UNIWERSYTET EKONOMICZNY W POZNANIU
Wydział Informatyki i Gospodarki Elektronicznej

Marek Wiśniewski

Uczenie ontologii z tekstu

Praca doktorska

Promotor: prof. dr hab. Witold Abramowicz

KATEDRA INFORMATYKI EKONOMICZNEJ
Poznań 2008

Spis treści

Spis rysunków

Spis tabel

Rozdział 1

Wprowadzenie

Polski Internet rozwija się. Wpływ na to mają dwa główne czynniki: rozwijający się rynek reklamy w Internecie oraz rosnący rynek handlu elektronicznego. Rynek handlu elektronicznego w Polsce dzieli się ze względu na siłę rynkową podmiotów na rynek sklepów internetowych oraz rynek platform (serwisów) aukcyjnych. Podstawowe wartości ekonomiczne podmiotów uczestniczących w rynku handlu elektronicznego przedstawione w sprawozdaniach finansowych za rok 2007 charakteryzują się silną tendencją wzrostową. Przychody polskich sklepów internetowych wzrosły średnio o 73%, łączne obroty platform aukcyjnych o ok. 53%, a cały rynek wzrósł o 62% i osiągnął wartość 8,1 mld zł (InternetStandard i Sklepy24.pl, 2008).

Cztery z pięciu największych polskich sklepów internetowych, do których należą: Agito.pl, Komputronik.pl, Merlin.pl, Oponeo.pl oraz Max24.pl zanotowaały w roku 2007 wzrost przychodu znacznie przekraczający 100% w porównaniu z rokiem 2006 (Tabela 1.1).

Badania przeprowadzone przez serwis Sklepy24.pl w dniach 16–20 grudnia 2007 r. wśród przedstawicieli polskich sklepów internetowych pokazały, że jednym z głównych czynników odpowiedzialnych za dynamiczny wzrost

Sklep	Przychód 2007 w zł	Wzrost przychodu
Agito.pl	120 mln	126%
Komputronik.pl	100 mln	110%
Merlin.pl	80 mln	27%
Oponeo.pl	63 mln	110%
Max24.pl	60 mln	140%

Tabela 1.1: Przychody największych polskich sklepów internetowych. Źródło: Raport e-commerce 2007, Internet Standard, Sklepy24.pl

rynku handlu elektronicznego w Polsce jest "upowszechnienie dostępu do Internetu", na co wskazało 88,3% respondentów.

1.1 Motywacja

Handel elektroniczny (ang. *e-commerce*) to jedna z pierwszych dziedzin informatyki ekonomicznej, która wykorzystała potencjał sieci Internet. Standardy z rodziny EDI umożliwiły prowadzenie biznesu na niespotykaną wcześniej skalę. Inicjatywy standaryzacyjne na poziomie technicznym umożliwiają sprawną i efektywną wymianę danych (Medjahed i in., 2003). Przeniesienie handlu w świat elektroniczny wymaga jednak transparentnych i bezpiecznych przepływów informacji i wiedzy. Konieczne jest zrozumienie, w jaki sposób warstwy pojęciowe, które stanowią przedmiot wymiany w elektronicznej gospodarce, winny być pozyskiwane, reprezentowane, współdzielone oraz przetwarzane zarówno poprzez ludzi, jak i inteligentne agenty.

Wizja Sieci Semantycznej (Berners-Lee i in., 2001) wyznacza fundamenty dla wymiany informacji pomiędzy współpracującymi na elektronicznym rynku podmiotami. Wymiany, w której ontologie dostarczają współdzieloną warstwę pojęciową precyzującą znaczenie danych, a inteligentne agenty w imieniu użytkowników pozyskują i wymieniają semantycznie wzbogacone informacje.

Osiągnięcia w dziedzinie ontologii umożliwiają jednoznaczną reprezentację pojęć będących przedmiotem wymiany na rynkach elektronicznych zarówno poprzez ludzi, jak i odpowiednio skonstruowane programy komputerowe (Hepp, 2008). Języki ontologii RDF[1] oraz OWL[2] są standardami organizacji W3C[3], która upowszechnia najważniejsze technologie związane ze współoperatywnością systemów informatycznych na poziomie technicznym i semantycznym[4].

Semantyczny handel elektroniczny to handel prowadzony przy pomocy środków elektronicznych, w którym wymiana informacji wzbogacona jest o semantykę. Semantyczny handel elektroniczny to podejście do zarządzania

[1] `http://www.w3.org/RDF/`

[2] `http://www.w3.org/2004/OWL/`

[3] `http://www.w3.org/`

[4] Współoperatywność to cecha jakościowa systemów informacyjnych polegająca na zdolności i otwartości na współpracę z innymi systemami. Współoperatwyność oznacza unikanie rozwiązań mogących negatywnie wpłynąć na współpracę z innymi systemami, pierwotnie nie przeznaczonymi do współpracy z nimi, a więc m.in. stosowanie zamkniętych lub prawnie zastrzeżonych standardów, wykorzystanie licencji ograniczających swobodne dysponowanie danymi, etc. Współoperatywność jest w trakcie niniejszej pracy traktowana jako synonim często używanego w praktyce terminu *interoperacyjność*.

wiedzą w procesach zachodzących na rynkach elektronicznych poprzez systematyczne zastosowanie technologii Sieci Semantycznej (Singh i in., 2005).

Handel elektroniczny składa się z segmentów, których akronimy odnoszą się do klasy podmiotów w nim uczestniczących. Przedmiot największego zainteresowania biznesu stanowią segmenty B2C (biznes-konsument) oraz B2B (biznes-biznes).

1.1.1 Rynek B2C

Segment B2C obejmuje cztery perspektywy:

- perspektywę konsumenta, którego celem jest wyszukiwanie informacji o produktach lub usługach. W 2007 r. na polskim rynku handlu elektronicznego konsumenci stanowią 41,9% internautów w Polsce[5] (źródło: Nettrack SMG/KRC),
- perspektywę dostawcy, który dostarcza informację o produktach lub usługach różnych producentów, np. Komputronik[6], Centrum Komputerowe Znak[7],
- perspektywę producenta, który wytwarza produkty lub usługi,
- perspektywę brokera, który pośredniczy pomiędzy producentami i konsumentami (np. serwisy aukcyjne). W 2007 r. polski rynek serwisów aukcyjnych został zdominowany przez serwis Allegro.pl, który posiadał ponad 80% udział w rynku(InternetStandard i Sklepy24.pl, 2008).

Wymienione klasy podmiotów (aktorzy) współuczestniczą na rynku handlu elektronicznego. Charakter ich współpracy oraz obecny stan rozwoju technologii stosowanych w segmencie B2C implikuje szereg problemów, tj.:

Wyszukiwanie produktów lub usług. Technologia stosowana na rynku handlu elektronicznego umożliwia szybszy i tańszy dostęp do produktów i usług. Bariery natury fizycznej i czasowej zostały zmniejszone w porównaniu z tradycyjnym dostępem. Odnalezienie konkretnego produktu czy usługi nadal jest jednak czasochłonne, zwłaszcza wtedy, gdy

[5] Wyróżnikiem konsumenta na rynku handlu elektronicznego jest zawarcie przynajmniej jednej transakcji przez Internet.

[6] `http://www.komputronik.pl/`

[7] `http://znak.pl/`

nie korzysta się z serwisów znanych dostawców lub poszukiwany produkt czy usługa posiada cechy o zaawansowanych warunkach brzegowych (np. konkretne wartości cech). Scenariusze związane z wyszukiwaniem dobrze ilustrują problem wieloznaczności terminów oraz niejednomianowości pojęć. Wieloznaczność terminów oznacza sytuację, w której termin składający sie z jednego lub kilku wyrazów, posiada wiele znaczeń (pojęć). Niejednomianowość oznacza zestaw różnych terminów do tego samego pojęcia. Na przykład, szukając konkretnego urządzenia naręcznego mechanizm wyszukiwawczy znaleźć powinien urządzenie nie tylko z klasy "urządzenie naręczne", lecz również: "PDA", "handheld" lub po prostu "telefon".

Filtrowanie. Mechanizmy budowania profilu użytkownika dedykowane są dla konkretnych dostawców. Profil użytkownika skonstruowany w jednym sklepie nie ma zastosowania w innym miejscu. Brakuje więc mechanizmów, które są w stanie ujednolicić proces budowania i utrzymywania profili użytkowników, a w konsekwencji umożliwić obniżenie kosztów budowania filtrów. Szerokie opracowanie poświęcone m.in. zastosowaniu filtrowania informacji na rynku handlu elektronicznego znaleźć można w Abramowicz (2008).

Przejrzystość rynku. Handel elektroniczny znacząco zwiększył przejrzystość rynku. Porównanie produktów nadal jednak jest trudne z powodu niejednomianowości pojęć. Odpowiednia konstrukcja strony dostawcy lub informacje dodatkowe np. w postaci obrazka mogą podsunąć znaczenie terminu. Są to jednak metody pomocnicze, nieskuteczne w przypadku przetwarzania maszynowego.

Dojrzałość usług. Większość dostawców umożliwia oprócz przeglądania listy dostępnych produktów, również zawieranie transakcji. Doświadczenia z dziedziny elektronicznej wymiany danych wskazują, że następnym etapem rozwoju jest automatyzacja procesów (Medjahed i in., 2003), np. zautomatyzowanie transakcji kupna produktów z systemem monitorującym stan zapasów. W celu uzyskania współoperatywności na poziomie semantycznym należy produkty będące przedmiotem wymiany ujednolicić pod względem wykorzystanej warstwy pojęciowej.

Negocjacje. Handel elektroniczny wymaga negocjacji warunków świadczenia usług (np. umów SLA[8]) lub parametrów produktów. W niektórych

[8]Service Level Agreement — umowa dotycząca parametrów świadczenia usług.

przypadkach negocjacje mogą przebiegać z wykorzystaniem odpowiedniego oprogramowania. Narzędzia takie muszą jednak posiadać dostęp do współdzielonego modelu pojęciowego.

Perspektywa konsumenta

Z punktu widzenia konsumenta rozwiązaniem niektórych ze wskazanych problemów mają być agenty, które w jego imieniu wyszukują relewantne produkty (tzw. *shopbots*) (Fasli, 2007; Palopoli i in., 2006; Garfinkel i in., 2006). W Polsce popularność zyskały serwisy, które korzystają z takich agentów na rzecz użytkownika np. Skąpiec[9] lub Ceneo[10]. Architektura takich narzędzi zależna jest od pozycji i możliwości producenta. Jeśli są one znaczące, np. poprzez duży udział w rynku, producent narzędzia może pozwolić sobie na wymuszenie modelu pojęciowego. Niestety nie jest on często dostosowany do specyfiki działalności wszystkich partnerów. W przeciwnym razie, dostawca jest zmuszony do implementacji kosztownych interfejsów. Obydwie strategie implikują znaczne bariery rozwoju.

Perspektywa dostawcy i producenta

Dostawcy informacji dotyczącej produktów lub usług (np. Komputronik, Centrum Komputerowe Znak) stają przed zadaniem prezentacji oferty. Najczęściej spotykany model prowadzi do dostarczenie opisu produktów przez producenców, najczęściej w postaci tekstu w języku naturalnym[11]. Do zadań dostawcy należy odpowiednia klasyfikacja produktów w schemacie stosowanym do prezentacji oferty. W zdecydowanej większości przypadków operacja ta wykonywana jest ręcznie. W szczególności na rynku produktów IT, ze względu na wysoką zmienność oferty, przeprowadzana jest dość często.

Wejście na nowy rynek, zmiana przedmiotu oferty lub jej rozszerzenie odnośnie nazw produktów oraz relacji pomiędzy nimi zachodzących wymaga znaczących nakładów. W fazie wstępnej stanowią one zatem barierę rozwoju.

Witryny internetowe dostawców prezentują oferty w sposób znacząco utrudniający pozyskanie prawdziwego znaczenia produktów. Stosowanie terminologii nie posiadającej odwzorowania do modelu formalnego (np. ontologii) oznacza, że informacje przedstawione na witrynie mogą być zrozumiałe wyłącznie dla człowieka. Zdarzają się przypadki, w których nawet człowiekowi trudno jest zinterpretować klasę produktu. Na przykład, produkt o na-

[9] `http://www.skapiec.pl/`

[10] `http://ceneo.pl/`

[11] Pominięte tym samym zostają przypadki szczególne, w których producent jest równocześnie dostawcą. Najsłynniejszym przypadkiem takiego modelu jest Dell.

zwie "A4-TECH Navigator Opto BW-5UP"[12] to urządzenie wskazujące, co nie dla wszystkich jest oczywiste. Programy komputerowe mają tym większy problem w prawidłowym rozpoznaniu takiego produktu. W konsekwencji witryna nie jest w stanie reagować na jakiekolwiek komunikaty wysyłane przez agenta, który reprezentować może np. profil użytkownika.

Perspektywa brokera

Segment B2C jest zdominowany przez pośredników, tzw. brokerów[13], którzy pośredniczą w transakcjach zawieranych pomiędzy producentami i konsumentami, umożliwiając negocjacje warunków kupna/sprzedaży.

Przed brokerem stoją problemy dotyczące zarówno konsumenta, jak i dostawcy. Połączenie potrzeb konsumenta z zakresu informacji o produktach i usługach z potrzebami dostawcy dodatkowo potęguje problemy niejednoznaczności i wielomianowości pojęć.

1.1.2 Rynek B2B

Rynek B2B to segment handlu elektronicznego, w którym interakcja następuje pomiędzy partnerami biznesowymi. Wyróżnia się trzy modele rynku B2B:

1:1 Dwa podmioty prowadzą handel przy pomocy środków elektronicznych. Wymagana jest zgoda w zakresie metody komunikacji oraz warstwy pojęciowej. Na poziomie technicznym rozwiązaniem są protokoły internetowe TCP/IP oraz SOAP, natomiast w warstwie syntaktycznej standardem staje się język XML.

1:N Jeden podmiot gospodarczy prowadzi handel z wieloma partnerami (przeważnie o słabszej pozycji rynkowej). W większości przypadków podmiot dominujący dyktuje warunki współpracy, wyznaczając jej narzędzia wraz z warstwą pojęciową. Taki scenariusz powoduje problemy związane z ograniczoną ekspresywnością modelu oraz niskim dopasowaniem do zmian.

N:M Wiele podmiotów współpracuje z wieloma partnerami. Często (ale nie zawsze) prowadzi to do tworzenia się elektronicznych rynków (ang. *e-marketplace*), które wymagają nie tylko standardów komunikacji, ale również wspólnej warstwy pojęciowej.

[12] Produkt z oferty sklepu komputerowego Komputronik.

[13] W klasycznym handlu pojęcie brokera ma nieco inne znaczenie, niż w przypadku handlu elektronicznego, gdzie pojęcie brokera jest pojemniejsze. Broker w handlu elektronicznym oznacza często hurtownika, dystrybutora lub integratora.

Standardy wymiany danych drogą elektroniczną (EDI) dla segmentu B2B umożliwiły przeniesienie wielu procesów w przestrzeń elektroniczną. Standardy EDIFACT[14] oraz architektury komponentowe CORBA[15] lub EJB[16] ustępują miejsca standardom komunikacji opartym na języku XML, który definiuje wyłącznie strukturę komunikatów. Konieczne jest zatem określenie znaczenia przesyłanych danych w postaci formalnej i współdzielonej warstwy pojęciowej zarówno na poziomie opisu zasobów podmiotu, jak i wymiany dokumentów (Guo, 2006).

1.1.3 Problem badawczy

Z przeprowadzonej analizy problemów biznesowych wynika, że kluczowym wyzwaniem dla narzędzi handlu elektronicznego jest stworzenie formalnej, współdzielonej warstwy pojęciowej, czyli ontologii. Tezę tę potwierdzają badania zarówno w dziedzinie nauk poznawczych (Kauffman i Walden, 2001), jak i nauk technicznych (Uschold i Gruninger, 2004; Fensel, 2003).

Ontologie stanowią aktywny przedmiot badań informatyki ekonomicznej (Hepp i in., 2008; Hepp, 2008). Chronologicznie pierwszą definicją ontologii w kontekście systemów informacyjnych uznaje się definicję Thomasa Grubera z roku 1993 (Gruber, 1993): "ontologia to formalna, współdzielona warstwa pojęciowa w danej dziedzinie". Rozwinięcie tej definicji podaje Gruber (2008).

Ontologie są zróżnicowane, co wynika ze źródeł badań nad ontologią wywodzących się z filozofii. Najczęściej wykorzystywana klasyfikacja ontologii obejmuje trzy poziomy szczegółowości (Navigli i Velardi, 2004):

1. Ontologie wyższego rzędu odzwierciedlające filozoficzne relacje pomiędzy bytami, które są możliwe do zastosowania w każdej ontologii, np. SUMO (Niles i Pease, 2001).

2. Podstawowe ontologie dziedzinowe, które odzwierciedlają byty i zależności w kluczowych dziedzinach, np. IT czy biologia. Ich wykorzystanie ograniczone jest tylko do zastosowań dziedzinowych, nie są jednak na tyle specyficzne, aby zaspokoić potrzeby konkretnych aplikacji.

3. Ontologie aplikacyjne, które definiują warstwę pojęciową dla specyficznej aplikacji w danej dziedzinie.

[14] `http://www.unece.org/trade/untdid/welcome.htm`
[15] `http://www.corba.org/`
[16] `http://java.sun.com/products/ejb/`

Istniejące repozytoria ontologii, np. SchemaWeb[17], biblioteka DAML[18] lub Swoogle[19] pozwalają na wyszukanie znacznej liczby gotowych ontologii.

Biblioteka SchemaWeb umożliwia pozyskanie 240 ontologii[20]. Niestety, znaczna ich część dotyczy standardów, odnaleźć można np. ontologie RDF, OWL-S, vcard, bibtex, ACL, Wordnet.

Biblioteka DAML udostępnia 282 ontologie[21]. Serwis umożliwia wygodną nawigację poprzez wybrane właściwości zgromadzonych ontologii, np. URI, słowa kluczowe, datę umieszczenia lub jednostkę odpowiedzialną za jej rozwój. Wśród ontologii ogólnych znaleźć można rodzinę ontologii CYC(Lenat, 1995). Biblioteka DAML zawiera większą liczbę ontologii dziedzinowych niż SchemaWeb.

Swoogle nie jest klasyczną biblioteką — to wyszukiwarka semantycznie zaanotowanych dokumentów, które wykorzystują pojęcia oraz relacje będące częścią ontologii. Powiązanie jest jawne. Pozyskanie semantycznie zaanotowanego dokumentu jest zatem dużym krokiem w celu pozyskania samych ontologii. Swoogle jest o tyle istotnym źródłem, że liczba semantycznie opisanych dokumentów wynosi 878 462, a składających się na tę liczbę wyrażeń jest 609 639 517[22].

Wydawałoby się zatem, że skoro istnieje potrzeba wykorzystania ontologii oraz dostępne są repozytoria, w których znaleźć można setki gotowych ontologii, to wykorzystanie właściwej ontologii nie jest problemem. Niestety, tak nie jest, co najmniej z następujących powodów:

1. W większości przypadków dziedzina aplikacji jest na tyle specyficzna, że gotowych ontologii nie ma.

2. Nawet jeśli istnieją ontologie dla dziedziny właściwej, to nie pokrywają się one z oczekiwaniami odnośnie zakresu. Na przykład ontologia eClassOWL ma na celu stworzenie relacji pomiędzy pojęciami z dziedziny produktów i usług IT (Hepp, 2006). Pomimo tego nie nadaje się do wykorzystania w świetle warunków brzegowych zdefiniowanych w scenariuszach dla rynków B2C i B2B.

3. Ontologie mogą nie odpowiadać oczekiwanemu stopniowi szczegółowości, tj. być zbyt szczegółowe lub zbyt ogólne.

[17] `http://www.schemaweb.info`
[18] `http://www.daml.org/ontologies/`
[19] `http://swoogle.umbc.edu/`
[20] Stan na dzień 4 kwietnia 2008 r.
[21] Stan na dzień 4 kwietnia 2008 r.
[22] Stan na dzień 4 kwietnia 2008 r.

Pojawia się zatem problem niedostępności lub nieadekwatności ontologii. W obu przypadkach skutkiem jest potrzeba tworzenia ontologii.

Proces tworzenia ontologii nazywa się inżynierią ontologii. Istnieją dwie klasy metod inżynierii ontologii: ręczne oraz automatyczne. Ręczne metody (Cristani i Cuel, 2005; Pinto i Martins, 2004) pozwalają na bardzo precyzyjną definicję ontologii, jednak są bardzo kosztowne. W celu szacowania kosztu ręcznego tworzenia ontologii opracowane zostały nawet specjalne modele (Simperl i in., 2006).

Proces inżynierii ontologii można w części lub w całości zautomatyzować. Półautomatyczne lub automatyczne tworzenie ontologii nazwane zostało uczeniem ontologii (Buitelaar i in., 2005b). Stopień automatyzmu zależy od udziału eksperta w procesie uczenia.

Proces uczenia ontologii składa się z 6. następujących po sobie faz (Cimiano, 2006)(por. z rysunkiem 2.1 na stronie 24):

1. Ekstrakcja terminologii[23].
2. Ekstrakcja synonimów.
3. Ekstrakcja pojęć.
4. Ekstrakcja relacji taksonomicznych.
5. Ekstrakcja relacji nietaksonomicznych.
6. Ekstrakcja reguł.

W kontekście zastosowań ontologii dla handlu elektronicznego najbardziej krytycznymi warstwami są terminologia oraz relacje, ponieważ:

1. W procesie uczenia ontologii z tekstu (opisy tekstowe produktów) terminologia stanowi podstawę wspólnej warstwy pojęciowej. Każdy inny element ontologii wywodzi się z terminologii.

[23]W literaturze polskojęzycznej istnieje rozbieżność w kwestii tłumaczenia angielskiego wyrazu *term*. W większości przypadków stosuje się kalkę językową, używając wyrazu *term* w polskim znaczeniu (np. ekstrakcja termów). W niniejszej pracy wykorzystuje się wyraz *termin*. Deklinacja rzeczownika *termin* w liczbie pojedynczej jest standardowa dla rodzaju męskiego, natomiast w liczbie mnogiej rodzi pewne problemy. W celu wskazania konkretnych terminów, ich danej grupy, stosuje się standardową deklinację dla liczby mnogiej i rodzaju męskiego (np. 2 terminy, dane terminy — tryb określony). Nazwanie ich grupy następuje natomiast w trybie nieokreślonym przy pomocy wyrazu *terminologia* i deklinacji rzeczownikowej dla rodzaju żeńskiego (stąd ekstrakcja terminologii). Podobne rozumienie stosuje się w normach organizacji ISO (ISO 1087-1:2000, 2000; ISO 704:2000, 2000) oraz Polskiego Komitetu Normalizacyjnego (PN-ISO 1087-1:2004, 2004).

2. Efektywność metod ekstrakcji terminologii jest propagowana na pozostałe zadania uczenia ontologii. Każdy błąd popełniony w tej fazie ma swoje konsekwencje w następnych fazach.

3. W wielu aplikacjach zakłada się równoznaczność pomiędzy terminami a pojęciami.

4. Standardowe metody ujednoznaczniania terminologii nie są problemem naukowym — ograniczają się do posiadania odpowiednio bogatych zasobów lingwistycznych. Przy pomocy standardowych zasobów lingwistycznych definiujących relację synonimiczności, np. Aspell[24] lub WordNet (Fellbaum, 1998), można w prosty sposób uzyskać dobre wyniki.

5. Relacje nietaksonomiczne są trudniejsze do wykrycia niż relacje taksonomiczne. Wymagają głębszej wiedzy dziedzinowej oraz poznania specyfiki ekstrahowanych relacji. Na przykład, przedmiot relacji *wyprodukowanyPrzez* ogranicza się do organizacji lub osoby. Właściwości ekstrahowanych relacji często muszą być podane na wstępie procesu ekstrakcji (por. np. Bunescu i Mooney (2007) na stronie 89). Ekstrakcja relacji nietaksonomicznych obejmuje większy zakres zagadnień, tj. identyfikację relacji, nazwanie oraz ekstrakcję właściwą.

6. Relacje taksonomiczne można traktować jako szczególne zagadnienie ekstrakcji relacji nietaksonomicznych (relacja nazwana "is-a").

Szczegółowa analiza relewancji poszczególnych etapów procesu uczenia ontologii z tekstu przeprowadzona jest w rozdziale poświęconym ogólnemu modelowi (rozdział 3 — por. tabelę 3.3 na stronie 114).

Ekstrakcja terminologii jest zadaniem trudnym, co wynika z problemów różnorodności postaci terminów, do których należą (Okazaki i Ananiadou, 2006):

- różnice w pisowni (np. ortograficzne),
- różnice morfologiczne (fleksja, derywacja, wieloczłonowość),
- różnice syntaktyczne (np. składniowe, szyk zdania),
- różnice semantyczne (znaczeniowe).

Większość metod ekstrakcji terminologii charakteryzujących się dobrą efektywnością niestety tworzona jest na potrzebny konkretnych aplikacji lub

[24] `http://aspell.net/`

dziedzin (Ananiadou i Mcnaught, 2006; Okazaki i Ananiadou, 2006). Z przeglądu obecnego stanu wiedzy (rozdział 2) wynika, że istnieje tylko jedna metoda ekstrakcji terminologii przeznaczona dla dziedziny handlu elektronicznego, tj. produktów IT (Khandelwal, 2007). Niestety, dostępna jest wyłącznie na zasadach komercyjnych oraz posiada znaczące ograniczenia w postaci zdefiniowanych źródeł ekstrakcji w postaci tabel. Klasyfikuje to metodę jako ekstrakcja terminologii ze źródeł ustrukturyzowanych. Standardowe metody ekstrakcji terminologii wykazują efektywność liczoną miarą F na poziomie 30–40% (wyniki eksperymentów szczegółowo przedstawiono w rozdziale 7, por. zwłaszcza rysunek 7.10 na stronie 174 oraz tabele 7.4 i 7.5 na stronach następnych).

Dodatkowo, ze względu na klasę problemów w ekstrakcji terminologii, większość metod jest zależna od języka naturalnego (jeśli źródłem jest tekst w języku naturalnym). Problemy wynikające z różnorodności morfologicznej i syntaktycznej tekstu są inne dla różnych języków. Większość metod ekstrakcji terminologii przeznaczona jest dla języka angielskiego. Niestety oznacza to, że ekstrakcja terminologii dla języka polskiego wymaga odrębnych metod.

Obecne metody ekstrakcji relacji charakteryzują się koniecznością znacznego nakładu pracy ekspertów w procesie ekstrakcji. Do eksperta należą takie czynności, jak definicja relewantnych dla dziedziny relacji, nazwanie ich, często również definicja lingwistycznych warunków brzegowych dla relacji.

1.2 Przedmiot pracy

Przedmiotem pracy objętej niniejszą rozprawą jest uczenie ontologii z tekstu w języku naturalnym[25]. Aplikacją opracowanych mechanizmów jest handel elektroniczny, lecz wyniki przeprowadzonych eksperymentów oraz dobór źródeł wskazują na pewną ogólność opracowanych metod.

Przedmiot pracy określa cele pracy wywodzące się z motywacji, określony przez cele zakres czasowy, przedmiotowy oraz przestrzenny pracy, wynikające tezy, a także miary osiągnięcia założonych celów pracy.

1.2.1 Cele pracy

Motywacja podjęcia badań określa trzy cele pracy:

1. Opracowanie uogólnionej metody uczenia ontologii z tekstu.

[25]W rzeczywistości chodzi o teksty (liczba mnoga). W pracy stosuje się jednak liczbę pojedynczą terminu *tekst*, ponieważ tak przyjęło się określać tę dyscyplinę w najbardziej liczących się publikacjach. Porównaj np. Buitelaar i Cimiano (2006) lub Buitelaar i in. (2005a).

2. Opracowanie metody ekstrakcji terminologii dla uczenia ontologii z tekstu.

3. Opracowanie metody ekstrakcji relewantnych relacji nietaksonomicznych dla uczenia ontologii z tekstu minimalizujących udział eksperta dziedzinowego.

Cel pierwszy jest celem ogólnym. Cele drugi i trzeci są celami szczegółowymi i stanowią jego rozwinięcie.

Ogólność metody uczenia ontologii z tekstu wyraża się możliwością jej zastosowania dla różnych tekstów, np. w różnym języku naturalnym, stylu lub dziedziny. Ze względu na rozważane problemy natury biznesowej, uwaga w szczególności zostanie poświęcona przetwarzaniu tekstu w języku polskim z dziedziny elektronicznego handlu. Dla tychże dokonana zostanie optymalizacji metod, tak aby wskazać związek opracowanych metod z wybraną dziedziną.

Cele szczegółowe pracy są rozwinięciem celu ogólnego w dwóch aspektach: ekstrakcji terminologii oraz relacji. Rozważana klasa problemów sytuuje cele szczegółowe jako cele niezależne od konkretnego języka naturalnego.

1.2.2 Zakres pracy

W ujęciu czasowym zakres pracy obejmuje najważniejsze osiągnięcia z ostatnich 15 lat z dziedziny uczenia ontologii (Hepp, 2008; Zavitsanos i in., 2006; Buitelaar i Cimiano, 2006; Buitelaar i in., 2005a; Cimiano, 2006) oraz pokrewnych, takich jak ekstrakcja informacji (Manning i Schutze, 1999), przetwarzanie tekstu naturalnego (Manning i Schutze, 1999; Jurafsky i Martin, 2000), uczenie maszynowe (Bishop, 2006) oraz reprezentacja wiedzy (Sowa, 2000a). Zakres czasowy pracy jest szczegłówo przedstawiony na rysunkach 2.5 oraz 2.6, gdzie chronologiczny układ pokazuje, że pierwsze ważne dla niniejszej pracy osiągnięcia miały miejsce na początku lat 90. ubiegłego wieku.

W ujęciu rzeczowym praca obejmuje uczenie ontologii z tekstu, w szczególności poszczególne fazy, tj. ekstrakcję terminologii, pojęć, relacji taksonomicznych oraz nietaksonomicznych. Praca zawiera uogólnioną metodę uczenia ontologii oraz modele szczegółowe poszczególnych faz. Dodatkowo metody zostaną dostosowane do konkretnych właściwości tekstów z dziedziny elektronicznego handlu.

W ujęciu przestrzennym rozprawa obejmuje ewaluację sprawdzanych tez dla dziedzin elektronicznego handlu (korpus e-commerce) oraz aktualności z działalności uniwersytetu (korpus KMi-News). Dobranie dwóch dziedzin miało na celu dowiedzenie pewnej ogólności opracowanych metod. Zakres

ewaluacji dotyczy tekstów w dwóch językach naturalnych, tj. języku polskim i angielskim.

1.2.3 Tezy pracy

Pierwszy cel pracy oraz próba syntezy różnych metod prowadzi do postawienia następującej tezy ogólnej pracy:

Teza 1 *Uogólniona metoda umożliwia przeprowadzenie procesu uczenia ontologii z tekstu w języku angielskim oraz polskim. Opisuje również w sposób abstrakcyjny elementy procesu, zależności pomiędzy elementami oraz wykorzystywane klasy zasobów.*

Ogólność metody oznacza zdolność do przetwarzania tekstu w różnych językach naturalnych, stylach i dziedzinach. Abstrakcyjny charakter metody oznacza, że musi ona do wspólnej postaci sprowadzać możliwie wiele reprezentatywnych podejść do uczenia ontologii z tekstu. Metoda musi zatem abstrahować od specyficznych cech poszczególnych metod, koncentrując się na wspólnych elementach i zależnościach pomiędzy nimi. Poszczególne metody uczenia ontologii z tekstu powinny być specjalizacjami postaci ogólnej metody.

Drugi cel pracy oraz zakres rzeczowy prowadzą do wyznaczenia szczegółowej tezy pracy, która ma na celu opracowanie nowej metody ekstrakcji terminologii:

Teza 2 *Nienadzorowana metoda ekstrakcji terminologii wykorzystująca dynamiczne okno kontekstowe jest bardziej efektywna niż klasyczne metody ekstrakcji terminologii dla uczenia ontologii z tekstu wykorzystujące podejścia lingwistyczno-statystyczne lub klasyczny model n-gram.*

Wykorzystując obecny stan wiedzy proponuje się wprowadzenie nowej metody, która cechować się będzie większą efektywnością niż obecnie stosowane metody.

Trzeci cel pracy oraz zakres rzeczowy pracy prowadzą do definicji szczegółowej tezy pracy, która ma na celu opracowanie nowej metody ekstrakcji relacji nietaksonomicznych:

Teza 3 *Wykorzystanie sprzężenia zwrotnego pomiędzy aksjomatami dziedzinowymi i informacją lingwistyczną prowadzi do zmniejszenia wymagań przedmiotowych oraz ilościowych w ekstrakcji relacji nietaksonomicznych dla uczenia ontologii z tekstu.*

Zbiory anotacji lingwistycznej konstruowane są poprzez ekspertów dziedzinowych. Wykorzystanie odpowiedniego źródła może odciążyć eksperta dziedzinowego, a w niektórych przypadkach umożliwić przeprowadzenie ekstrakcji relacji. Proponuje się zatem, aby informacja wskazująca na relewancję relacji w dziedzinie była ekstrahowana z tzw. aksjomatów dziedzinowych określających ogólne prawa istniejące w dziedzinie. Dodatkowo oba te źródła mogą być przesłanką dla siebie nawzajem, tj. informacja zawarta w aksjomatach dziedzinowych może świadczyć o postaci funkcji ekstrakcji relacji ze zbiorów informacji lingwistycznych i odwrotnie. Dlatego proponuje się zastosowanie mechanizmu sprzężenia zwrotnego w celu minimalizacji udziału eksperta dziedzinowego w procesie ekstrakcji relacji.

1.2.4 Miary osiągnięcia celów

Przeprowadzanie dowodu postawionych tez pracy zależne jest od specyfiki badanych problemów oraz metod prowadzonych badań. W niniejszej pracy badane są problemy, które można podzielić według następującego schematu:

1. Problem jest znany w danej dziedzinie nauki oraz istnieją metody jego rozwiązania. Istniejące metody nie charakteryzują się jednak oczekiwaną efektywnością lub powodują kolejne problemy naukowe. Podczas dowodu można porównać się do istniejących metod przy pomocy miary ilościowej (np. precyzja i zwrot).

2. Problem jest znany, istnieją również jego rozwiązania. Nie można jednak porównać się do istniejących metod, ponieważ istnieją obiektywne trudności w odtworzeniu materiałów pierwotnych, a rzetelność materiałów wtórnych jest wątpliwa. Nowa metoda rozwiązania problemu może również korzystać z innego zestawu środków, których zastosowanie w sposób nieznaczny zmienia definicję problemu.

3. Problem nie jest znany w danej dziedzinie lub zakresie pracy (np. przestrzennym) i nie istnieją znane metody jego rozwiązania.

Pierwsza teza pracy, tj. uogólniona metoda uczenia ontologii z tekstu dla języka polskiego, została sklasyfikowana jako metoda dotycząca trzeciego z wymienionych problemów. Przegląd literatury przedstawiony w rozdziale 2. pokazuje, że nie istnieją metody uczenia ontologii z tekstu dla języka polskiego, a przeniesienie obecnie stosowanych podejść nie jest możliwe (por. wnioski przedstawione w sekcji 7.5 na stronie 176).

Druga teza pracy dotycząca metody ekstrakcji terminologii dotyczy problemu pierwszego. Przegląd literatury przedstawiony w rozdziale 2. pokazuje,

że problem jest znany oraz istnieją metody jego rozwiązania. Należy zatem sprawdzić jakość opracowanej metody ekstrakcji terminologii w porównaniu z obecnie istniejącymi metodami.

Trzecia teza pracy dotycząca metody ekstrakcji relacji dotyczy drugiego problemu. Przegląd literatury przedstawiony w rozdziale 2. pokazuje, że problem jest znany oraz istnieją metody jego rozwiązania. Nakładają one jednak dodatkowe problemy w postaci nadmiernej pracy eksperta. Brak odniesienia do innych metod nie oznacza, że nie należy wyliczać miar jakości. Dlatego opracowana metoda ekstrakcji relacji zostanie poddana ewaluacji przy pomocy standardowych miar jakości.

W związku z tym, że tezy pracy wynikają bezpośrednio z celów pracy, miary osiągnięcia tez pracy są jednocześnie miarami osiągnięcia celów pracy.

Miarą osiągnięcia szczegółowych celów pracy są standardowe miary wyszukiwania informacji (Baeza-Yates i Ribeiro-Neto, 1999), które są odpowiedzią na następujące potrzeby:

- potrzeba posiadania tylko relewantnych pojęć i relacji z dziedziny (wysoka precyzja),
- potrzeba pozyskania wszystkich relewantnych pojęć i relacji z dziedziny (wysoki zwrot).

1.2.5 Wykorzystane materiały

W pracy wykorzystano materiały empiryczne pierwotne oraz wtórne. Zgromadzone podczas pracy materiały wtórne to:

- metody ekstrakcji przedstawione w rozdziale 2. wraz z takimi cechami jak efektywność (osiągane przez autorów wyniki),
- wykorzystane aplikacje, np. GATE (Cunningham i in., 2002), SProuT (Piskorski i in., 2005) czy WordNet (Fellbaum, 1998),
- zasoby lingwistyczne, w tym korpus anglojęzyczny KMi.

Niestety części materiałów wtórnych nie można poddać weryfikacji, np. środowisko dochodzenia do wyników danych metod ekstrakcji jest niedostępne, a przedstawione informacje nie dają takiej możliwości. W związku z tym, istnieją obiektywne trudności w porównaniu się z częścią zgromadzonych materiałów wtórnych (np. pod względem efektywności metod).

W niniejszej pracy część metod została odtworzona przy pomocy publikacji autorów oraz odpowiednich implementacji. W wyniku zmiany niektórych

cech takiego procesu (np. inny korpus) materiały te nabierają cech materiałów pierwotnych, tj. wynikają bezpośrednio z przeprowadzonych badań (implementacji). Do wykorzystanych materiałów pierwotnych zaliczyć zatem należy:

- reprezentatywny zbiór metod przedstawionych w rozdziale 2. wraz z ich efektywnością, czyli osiągane w środowisku testowym wyniki,
- aplikacje zaimplementowane w trakcie niniejszej pracy, z których bezpośrednio wynika efektywność zaproponowanych metod (przedstawione głównie w rozdziale 7.),
- wyniki przedstawionych metod (rozdział 2. oraz 7.),
- zgromadzone korpusy (rozdział 7.),
- zgromadzone zbiory zasobów lingwistycznych (rozdział 4. oraz 7.).

1.3 Metody badawcze

Praca ujęta niniejszą rozprawą przebiegała według określonego cyklu działania, w którym wyróżnić można fazę określania (diagnozy problemu), fazę poszukiwań oraz fazę realizacji.

Faza określania obejmowała rozpoznanie i sformułowanie problemu braku semantyki w rozwiązaniach elektronicznego handlu (ang. *e-commerce*) wraz ze identyfikowaniem problemów niedostępności i nieadekwatności ontologii. Na tym etapie określony został temat i cel główny pracy oraz wstępny plan pracy. Wyniki prac nad fazą określania przedstawione są w niniejszym rozdziale.

Faza poszukiwań obejmowała ustalenie możliwości realizacji celów pracy na podstawie dostępnej literatury oraz narzędzi. Przegląd dostępnych metod oraz narzędzi doprowadził do szczegółowej analizy problemów wraz ze zdefiniowaniem tez pracy oraz wstępnej wizji realizacji jej celów. Wyniki prac nad fazą poszukiwań przedstawione są w szczególności w części poświęconej analizie obecnego stanu wiedzy (rozdział 2.) oraz w części dotyczącej uogólnionej metody uczenia ontologii z tekstu (rozdział 3.).

W fazie realizacji prace koncentrowały się na implementacji prototypu przedstawionych metod oraz uszczegóławianiu metod ekstrakcji terminologii oraz relacji. Uzyskiwane wyniki były źródłem interpretacji oraz kolejnych kierunków badań. Dopiero podczas realizacji, zarówno tezy, jak i zakres pracy uzyskały ostateczny kształt. Wyniki prac nad fazą realizacji przedstawione

są głównie w rozdziałach poświęconych metodom ekstrakcji terminologii (rozdział 5.) oraz relacji (rozdział 6.), a także prezentacji uzyskanych wyników (rozdział 7.).

Podjęte prace w zależności od fazy miały różny charakter (typ pracy naukowej). W fazie określania zastosowanie miała głównie *praca koncepcyjna*, która charakteryzuje się właściwym opracowaniem zebranych materiałów oraz ułożeniem badanych problemów w postaci logicznej kolejności. W fazie poszukiwań, zwłaszcza w przypadku opracowywania metod ekstrakcji, zastosowano *pracę metodologiczną*, która miała na celu krytykę obecnie stosowanych podejść do ekstrakcji oraz opracowanie nowych metod.

Całość pracy posiada charakter *prac analitycznych oraz syntetycznych*. Prace analityczne, takie jak: przegląd literatury, analiza metod, opracowanie szczegółowych metod ekstrakcji, przeplatają się z pracami o charakterze syntetycznym, np. syntezą obecnie stosowanych metod, modelem ogólnym uczenia ontologii z tekstu, czy syntezą uzyskanych wyników.

Cykl działania oraz charakter pracy spowodował zróżnicowanie zastosowanych metod badawczych. W fazie przetwarzania materiałów pierwotnych i wtórnych, tj. analizy obecnego stanu wiedzy oraz dochodzenia do wyników, wykorzystano następujące metody badawcze:

1. Analiza — metoda polegająca na rozłożeniu badanego problemu na części składowe i badaniu każdej części osobno (Pytkowski, 1985). Zastosowano ją w następujących etapach pracy:

   - postawienie problemu, w tym określenie celów oraz przedmiotu pracy,
   - przegląd obecnego stanu wiedzy,
   - praca nad charakterystykami anotacji lingwistycznych,
   - opracowanie szczegółowych metod na podstawie ogólnego modelu uczenia ontologii z tekstu.

2. Synteza — metoda polegająca na składaniu, zestawianiu, ujmowaniu czegoś jako całości (Pytkowski, 1985). Syntezę zastosowano głównie podczas przedstawienia obecnego stanu wiedzy w postaci podsumowania i usystematyzowania oraz prezentacji modelu ogólnego uczenia ontologii z tekstu.

3. Wyodrębnianie cech (abstrahowanie) — metoda polegająca na oddzieleniu jednego lub wielu składników i poddaniu ich badaniu (abstrahowanie odosobniające) lub pomijania cech indywidualnych a wybieraniu cech wspólnych (abstrahowanie uogólniające) (Pytkowski, 1985). Abstrahowanie odosobniające zastosowano w następujących etapach pracy:

- badanie konkretnej metody ekstrakcji w oderwaniu od innych elementów modelu ogólnego,
- badanie wpływu poszczególnych parametrów na metody ekstrakcji,
- badanie efektywności poszczególnych metod ekstrakcji dla całości cyklu uczenia ontologii.

Abstrahowanie uogólniające zastosowano w następujących etapach pracy:

- systematyka obecnie stosowanych metod w uczeniu ontologii z tekstu,
- przygotowanie reprezentatywnego zbioru obecnych metod ekstrakcji terminologii oraz relacji w celu przeprowadzenia analizy porównawczej,
- badania nad uogólnioną metodą uczenia ontologii z tekstu,
- przygotowanie anotacji lingwistycznej (analiza oraz wykorzystanie reprezentatywnych cech różnych modeli).

4. Dedukcja i indukcja — zastosowane podczas przejścia z modelu ogólnego do modelów ekstrakcji (dedukcja) oraz wyprowadzanie wniosków z uzyskanych wyników (indukcja).

5. Analogia — zastosowana przy wykorzystaniu analogicznego formatu anotacji lingwistycznej. Analogia nie stanowi dowodu, ale daje przekonanie, np. o możliwości rzetelnej ewaluacji metod.

6. Ilościowe i jakościowe ujmowanie problemów — zaproponowanie modelu ogólnego w ujęciu ilościowym oraz zastosowanie podejścia jakościowego w opisie oraz interpretacji przeglądu obecnego stanu wiedzy i uzyskanych wyników.

Metody badawcze w fazie systematyzowania i opracowywania wyników pracy:

1. Interpretacja — wyjaśnienie znaczenia i możliwości obecnie stosowanych metod, analiza uzyskanych wyników oraz eksperymentów (np. wydajność narzędzi).

2. Wnioskowanie — w pracy zastosowanie ma głównie wnioskowanie indukcyjne, np. podczas analizy uzyskanych wyników. Wnioskowanie dedukcyjne wykorzystane jest pomocniczo w ramach metody ekstrakcji relacji (jednym ze źródeł są aksjomaty dziedzinowe).

3. Definicja — szerokie zastosowanie przy okazji omówienia obecnego stanu wiedzy, modelu ogólnego oraz poszczególnych modeli ekstrakcji (np. definicja okna kontekstowego).

4. Model — to podobizna rzeczywistości możliwej i od nas zależnej (Pytkowski, 1985). Zastosowano w opracowaniu modelu ogólnego oraz poszczególnych modeli ekstrakcji.

Wykorzystane metody badawcze są wzajemnie od siebie zależne. Na przykład abstrahowanie jest podstawą analizy, synteza jest metodą konstrukcji modeli. Dlatego prace na poszczególnymi fazami charakteryzowały się zastosowaniem kilku metod badawczych na raz, tj.:

- przegląd obecnego stanu wiedzy — analiza w formie jakościowego ujęcia problemów, następnie abstrahowanie i synteza podstawą interpretacji,
- uogólniona metoda — synteza podstawą modelu, dedukcja podstawą analizy i modelu; definiowanie oraz ilościowe ujmowanie problemów,
- uzyskane wyniki — wnioskowanie indukcyjne i interpretacja podstawą indukcji i syntezy.

1.4 Struktura pracy

Niniejsza rozprawa składa się z 8 rozdziałów. Pierwsze dwa rozdziały, tj. rozdział niniejszy oraz następny (przegląd obecnego stanu wiedzy) tworzą część nieoryginalną pracy. Przedstawione w tych rozdziałach treści są znane w dziedzinach objętych zakresem pracy. Począwszy od rozdziału trzeciego aż do rozdziału ostatniego rozprawa przedstawia nowe treści, które stanowią oryginalny wkład autora. Granica pomiędzy częścią nieoryginalną (odtwórczą) oraz oryginalną (wkładem do nauki) następuje zatem po rozdziale drugim.

Rozdział 1. (niniejszy) ma charakter wprowadzenia i obejmuje motywację podjętych badań, cele, zakres i tezy pracy, miary osiągnięcia celów oraz wykorzystane metody badawcze.

W rozdziale 2. następuje przegląd obecnego stanu wiedzy drogą analizy wszystkich dostępnych metod i narzędzi. Ze względu na szeroki zakres obszaru badawczego przegląd obejmuje prawie 50 metod dotyczących zagadnień objętych zakresem pracy. Dodatkowo analiza obejmuje kilka narzędzi (gotowych do wykorzystania aplikacji), których funkcjonalność zbieżna jest z obszarem pracy. Podsumowaniem przeglądu jest synteza obecnych metod.

Rozdział 3. przedstawia uogólnioną metodę (tzw. metamodel — model ogólny poszczególnych modeli ekstrakcji) uczenia ontologii z tekstu. Metamodel jest ilościowym ujęciem przeprowadzonej syntezy i stanowi odniesienie do pozostałych opracowanych i przedstawionych w rozprawie modeli. Rozdział poświęcony metamodelowi jest też zestawieniem niezbędnych dla całego procesu definicji.

Rozdział 4. pracy obejmuje, na zasadzie abstrahowania wyodrębniającego, przegląd zagadnień związanych z anotacją lingwistyczną. Przedstawiony przegląd procesów anotacyjnych, standardów oraz formatów ma na celu odpowiedni dobór cech dla zastosowanych tekstów, tj. źródeł dla procesów ekstrakcji w uczeniu ontologii.

Rozdział 5. oraz 6. stanowią odpowiedź na odpowiednio drugi oraz trzeci cel pracy, tj. konkretnych metod ekstrakcji terminologii oraz relacji. Oba rozdziały w sposób szczegółowy omawiają konstrukcję modeli, ich uruchomienie oraz wykorzystanie. Rozdział 6. powstał przy znacznym udziale dr Marii Vargas-Vera[26] z Knowledge Media Institute w Open University w Wielkiej Brytanii. Za ten wkład i przyjemność wspólnej pracy serdecznie dziękuję.

Rozdział 7. przedstawia wykonane eksperymenty z wykorzystaniem reprezentatywnych metod oraz metod opracowanych w ramach metamodelu. Szczegółowy opis warunków, w których dokonano badania poprzedza przedstawienie oraz dyskusję nad uzyskanymi wynikami.

Rozdział ostatni zawiera podsumowanie pracy i odniesienie się do problemów, celów oraz tez wskazanych we wprowadzeniu.

[26] `http://people.kmi.open.ac.uk/maria/`

Rozdział 2

Przegląd obecnego stanu wiedzy

Inżynieria ontologii wymaga znacznego udziału osób posiadających odpowiednią wiedzę z analizowanej dziedziny (zwanych ekspertami dziedzinowymi). Szacunki wskazują, że stworzenie prostej ontologii to 3 roboczo-tygodnie, natomiast stworzenie bardziej złożonej wymaga nawet kilku osobo-miesięcy (Sabou i in., 2005). W celu wspomagania procesu szacowania pracochłonności inżynierii ontologii opracowano dedykowane modele (Simperl i in., 2006, 2007; Simperl i Mochol, 2006). Proces ten jest więc kosztowny (Wroe i in., 2004). Ponadto, ontologie zgodnie z definicją, wymuszają zgodę w odniesieniu do podstawowych pojęć i mechanizmów przedstawianej dziedziny. Osiągnięcie takiego konsensusu, również w grupie eksperckiej, jest często niemożliwe. Przykładem źródła, w którym konstrukcja bazy wiedzy trwała latami i ciągle jest precyzowana jest CYC (Lenat, 1995).

Ontologie są formalizacją wiedzy, dlatego dużym wyzwaniem jest zarządzanie nimi. Jeśli bowiem ontologie przedstawiają daną dziedzinę, to już chwilę po ich zbudowaniu mogą być nieaktualne. Rzeczywistość nie jest monotoniczna, tj. nie spełnia warunku zamkniętości Świata (ang. *closed world assuptiom*), zatem same ontologie zmieniają się wraz z napływem nowych faktów (Sowa, 2000a). O ile jednorazowy nakład w postaci pracy eksperta, pomimo że kosztowny, jest realny, tak już stałą jej modyfikację uznać należy za wysoce niepraktyczną i nierealną.

Tworzenie ontologii z wykorzystaniem automatycznych metod jest więc korzystne i to nawet przy założeniu, że jakość ich działania (mierzona np. miara precyzji i zwrotu, por. sekcję 1.2.4 na stronie 14) jest z reguły niższa niż pracy eksperckiej. Automatyczne lub półautomatyczne tworzenie ontologii nazwane zostało *uczeniem ontologii*. Uczenie ontologii może odbywać się z nadzorem (półautomatyczne) oraz bez nadzoru (automatyczne).

Uczenie ontologii jest procesem. W odróżnieniu od ewolucji ontologii (Leenheer i Mens, 2008; Flouris, 2006; Haase i Sure, 2004) czy wersjonowania

(Voelkel, 2005), proces uczenia ontologii nie wykorzystuje ontologii początkowej, a przynajmniej nie jest to źródło podstawowe. Pominięta zostaje tym samym sytuacja, w której ontologia zewnętrzna wykorzystywana jest w procesie uczenia jako dodatkowe źródło zwiększające efektywność uczenia.

Uczenie ontologii odbywa się na podstawie danych wejściowych w postaci źródeł. Ze względu na ich charakter w dorobku dziedziny dostępne są metody uczenia ontologii z następujących źródeł:

- teksty w języku naturalnym, najczęściej w postaci kolekcji dokumentów (korpusu),
- słowniki, tezaurusy oraz źródła o stałej strukturze i zdefiniowanym interfejsie, np. WordNet (Fellbaum, 1998),
- bazy wiedzy,
- dane ze źródeł ustrukturyzowanych.

Każde z wymienionych źródeł stanowi podstawę odrębnej grupy mechanizmów. Uczenie ontologii z tekstu jest jednak najliczniej reprezentowanym z kierunków. Jest to spowodowane trudnością analizy języka naturalnego w celu konstrukcji ontologii oraz olbrzymią liczbą łatwo dostępnych dokumentów będących przedmiotem analizy. Ponadto przy pomocy języka naturalnego najłatwiej przekazywać informacje oraz wiedzę. Bazy wiedzy, tezaurusy, czy nawet ontologie, pomimo wysiłków, nadal są mało przystępne dla zwykłego pracownika organizacji opartej na wiedzy. Większość informacji możliwych do przetworzenia przez narzędzia lingwistyczne ciągle generowana jest w postaci języka naturalnego i dostępna jest przez standardowe mechanizmy wyszukiwawcze (abstrahujemy tym samym od tzw. Głębokiego Internetu, który wymaga innej klasy metod (Kaczmarek, 2007)). Do tej pory nie udało się stworzyć języka formalnego, który choćby w części był tak ekspresywny i zrozumiały, jak język naturalny (Sowa, 2000b). W rezultacie zdecydowanie największa liczba źródeł najlepiej predysponowanych do uczenia ontologii jest opisana w języku naturalnym. Proces uczenia ontologii na podstawie dokumentów tekstowych nazywany jest uczeniem ontologii z tekstu.

2.1 Proces uczenia ontologii z tekstu

Kompleksowa próba zdefiniowania procesu uczenia ontologii z tekstu dokonana została dwa razy. Zarówno proces zdefiniowany w Maedche (2002), jak

i ten przedstawiony w Cimiano (2006) dotyczą sekwencji kolejnych zadań, które należy wykonać, aby uzyskać ontologię.

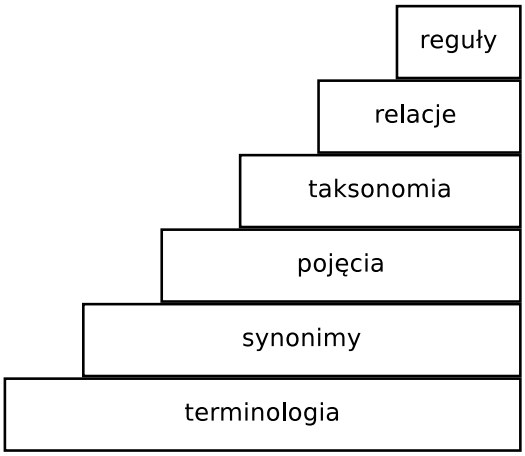
Proces zdefiniowany w Maedche (2002) jest bardziej ogólny, tj. składa się z fazy ekstrakcji terminologii, budowania ontologii oraz przycinania ontologii (ang. *ontology pruning*). Proces ten dokonuje klarownego, ale sztywnego, podziału pomiędzy fazę analizy lingwistycznej (ekstrakcja terminologii), fazę właściwego budowania ontologii oraz etap dostosowywania ontologii do rzeczywistych potrzeb. Charakterystyczna jest zwłaszcza trzecia faza, która kładzie nacisk na dostosowanie wynikowej ontologii do konkretnego zadania, aplikacji czy specyfiki danej dziedziny. W bardziej współczesnych podejściach do uczenia ontologii fazy te przenikają się nawzajem, na przykład filtrowanie terminologii następuje już w fazie analizy lingwistycznej (Missikoff i in., 2002) lub analiza lingwistyczna wykorzystywana jest także na etapie budowania ontologii (Maedche i Staab, 2000b).

Cimiano (2006) prezentuje podejście bardziej szczegółowo i lepiej klasyfikuje zidentyfikowane i badane problemy. W skład procesu wchodzą kolejno zadania ekstrakcji następujących obiektów:

- terminologii,
- synonimów,
- pojęć,
- relacji taksonomicznych,
- relacji nietaksonomicznych,
- reguł.

Ilustracją procesu uczenia ontologii jest rysunek 2.1, który przedstawia symboliczne zmniejszanie się liczby rozważanych elementów. W przypadku pierwszej warstwy analizy zbiór wyników w postaci terminów jest najliczniejszy, w przypadku warstwy ostatniej, elementów, czyli reguł, jest najmniej. Wiele podejść dotyczących uczenia ontologii z tekstu nie przeprowadza tych etapów w sposób sekwencyjny, lecz równoległy. Zdarzają się podejścia, które dokonują na przykład ekstrakcji pojęć równolegle z ekstrakcją relacji (Agirre i in., 2000). Analiza literatury dziedzinowej wykazuje również, że jest bardzo mało podejść obejmujących wszystkie wymienione fazy. Przeważnie rozwiązania skupiają się na poszczególnych warstwach.

Jedyne badania z dziedziny uczenia ontologii z tekstu dla języka polskiego pokazują wyłącznie wstępny zarys problemów i wyzwań (Wisniewski, 2006). Niniejsza praca stanowi rozwinięcie obranego kierunku badań.



Rysunek 2.1: Warstwowe przedstawienie procesu uczenia ontologii z tekstu. Źródło: (Cimiano, 2006) z modyfikacjami

2.1.1 Terminologia

Terminy są podstawowymi obiektami w procesie uczenia ontologii. Wyrażają one semantycznie jednorodne wyrażenie w tekście w postaci wyrazu lub grupy wyrazów.

Poniższe wyrażenie przedstawia fragment tekstu z dziedziny gospodarki. Zaznaczone zostały w nim terminy, które są wynikiem działania analizy tekstu i ekstrakcji terminologii.

> UniCredito Italiano zamierza przeprowadzić połączenie banków BPH i Pekao SA do końca 2006 roku
> wynika z komunikatu NBP.

Ekstrakcja terminologii odbywa się przy użyciu dwóch grup metod: lingwistycznych oraz statystycznych. Popularne są również podejścia hybrydowe, które łączą ze sobą analizę lingwistyczną oraz statystyczną, np. Daille (1996) (por. sekcję 2.2.2) oraz Frantzi i in. (2000) (por. sekcję 2.2.4). Analiza lingwistyczna jest wtedy stosowana do ekstrakcji potencjalnych terminów, natomiast analiza statystyczna do statystycznej oceny ich przydatności.

Najbardziej popularne są metody oparte wyłącznie na miarach statystycznych, ponieważ są one najłatwiejsze do opracowania oraz najbardziej znane, a także wymagają najskromniejszej anotacji lingwistycznej. Metody zaliczane do drugiej grupy są jednak bardziej skuteczne (Wermter i Hahn, 2006) i coraz częściej wykorzystuje się je łącznie z metodami statystycznymi.

Metody lingwistyczne

Metody lingwistyczne polegają na zastosowaniu analizy lingwistycznej (Manning i Schutze, 1999), a następnie wśród wygenerowanej informacji lingwistycznej, zastosowania wzorców ekstrakcji. Produktem analizy lingwistycznej jest informacja lingwistyczna w postaci zanotowanego tekstu. Informacja ta stanowi cenne źródło mechanizmów ekstrakcji, których podstawową jednostką jest reguła (tzw. *wzorzec ekstrakcji*). Najpopularniejszy wzorzec ekstrakcji terminologii klasyfikuje wszystkie znalezione rzeczowniki (zanotowane w informacji lingwistycznej) jako terminy.

W analizie lingwistycznej wykorzystuje się dwie grupy metod: ekstrakcję opartą na częściach mowy (tzw. *Part-of-Speech (POS) tagging*), nazywaną również *płytką analizą tekstu* oraz ekstrakcję opartą na zależnościach relacyjnych zdania zwaną *głęboką analizą tekstu*. Podział pomiędzy zakresem płytkiej i głębokiej analizy tekstu może być jednak różny w zależności od rodzaju oraz kontekstu poszczególnych prac.

Ekstrakcja oparta na częściach mowy wykorzystuje klasyfikację każdego wyrazu. Informacja lingwistyczna w wyrażeniu:

zamierza(VB) przeprowadzić(VB) połączenie(NN) banków(NN),

klasyfikuje pierwsze dwa wyrazy jako czasowniki (VB) oraz dwa ostatnie wyrazy jako rzeczowniki (NN). Tak przygotowana informacja lingwistyczna jest przedmiotem definicji wzorców powierzchni (ang. *surface patterns*). Wzorce te, oprócz klasyfikacji części mowy, używają również kolejności występowania wyrazów w zdaniu. Anotacja lingwistyczna wykorzystana dla wzorców powierzchni nie zawiera jednak żadnej informacji na temat logicznych zależności w zdaniu. Przykładowy wzorzec powierzchni definiuje wyrażenie rzeczownikowe jako występujące po sobie rzeczowniki (czyli NN, NN). Oczywiście na dowolnym etapie tego procesu można zastosować analizator morfologiczny, który m.in. sprowadza słowa do ich gramatycznej formy podstawowej.

Analiza oparta na anotacjach części mowy jest podstawową metodą popularnych systemów NLP (Cunningham i in., 2002; Piskorski i in., 2005; Hepple, 2000; Nadeau, 2005). Wzorce powierzchni można definiować w językach reguł, np. JAPE (Cunningham i in., 2000).

Ekstrakcja oparta na zależnościach relacyjnych wymaga znacznie bogatszej anotacji lingwistycznej. Oprócz informacji dotyczącej wyrazu, jego pozycji w zdaniu i klasyfikacji POS, wymaga dodatkowo informacji dotyczącej wyrażenia logicznego, czyli fragmentu logicznej struktury zdania, w której wyraz ten znajduje się. Podstawowymi elementami wyrażeń logicznych są asymetryczne binarne relacje pomiędzy wyrazem podstawowym (tzw. *głową wyrażenia*) oraz wyrazem modyfikującym. Na przykład omawiany fragment tekstu

	Token	POS	Głowa	Relacja
1	UniCredito Italiano	Rzeczownik	—	—
2	zamierza	Czasownik	1	rzeczownik-czasownik
3	przeprowadzić	Czasownik	2	czasownik-czasownik
4	połączenie	Rzeczownik	3	czasownik-rzeczownik
5	banków	Rzeczownik	4	rzeczownik-rzeczownik

Tabela 2.1: Przykładowa informacja lingwistyczna dla wyrażenia *UniCredito Italiano zamierza przeprowadzić połączenie banków* obejmująca informacje o częściach mowy oraz zależnościach relacyjnych

może prowadzić do zbioru informacji lingwistycznej przedstawionego w tabeli 2.1.

Ostatnia pozycja tabeli 2.1 powstaje poprzez relację głowy oraz słowa modyfikującego, czyli dwóch powiązanych elementów logicznej struktury zdania. Istnieją pewne klasy relacji, które nazywają się *wyrażeniami*. Dla przykładu, w języku angielskim wyrażenie rzeczownikowe przy użyciu wzorców syntaktycznych można uzyskać np. dla relacji rzeczownik-rzeczownik oraz rzeczownik-przymiotnik. Wyrażenia są podstawą tworzenia wzorców syntaktycznych dla analizy lingwistycznej opartej na zależnościach relacyjnych.

Ekstrakcja informacji przy pomocy definicji wzorców powierzchni i syntaktycznych jest dziedziną nauki rozwijającą się bardzo dynamicznie. Badania nad efektywnością wzorców prowadzone są dla wielu języków naturalnych. Najliczniej reprezentowany jest język angielski. Prace dla języka polskiego prowadzone są m.in. przez zespół z Katedry Informatyki Ekonomicznej Uniwersystetu Ekonomicznego w Poznaniu (Abramowicz i in., 2006).

Metody statystyczne

Metody statystyczne polegają bądź na analizie współwystępowania terminów w korpusie dokumentów, bądź na analizie porównawczej częstości występowania terminów w dziedzinie i korpusie ogólnym. Na przykład, terminy *komputer* i *stół* są równie popularne, aczkolwiek ich częstość występowania w dziedzinach IT i stolarskiej jest znacząco różna.

Do najczęściej wykorzystywanych miar statystycznych należą: miary Jaccarda, Dice'a i cosinusa przy analizie współwystępowalności, miara TFIDF oraz test χ^2. Zdecydowanie najpopularniejszą miarą metod statystycznych jest miara TFIDF (równanie 2.1).

$$tfidf(w) = tf(w) * log\frac{N}{df(w)}, \tag{2.1}$$

gdzie:
$tfidf(w)$ — względna ważność słowa w dokumencie,
$tf(w)$ — częstość terminu (liczba wystąpień terminu w dokumencie),
$df(w)$ — częstość dokumentu (liczba dokumentów z terminem),
N — liczba dokumentów w korpusie.

Miara TFIDF określa ważność terminu w dokumencie na podstawie jego częstości występowania w dokumencie oraz korpusie. Im większa częstość występowania w dokumencie oraz mniejsza w korpusie, tym dany termin będzie ważniejszy. TFIDF jest miarą popularnie wykorzystywaną w indeksowaniu dokumentów, gdzie wyjątkowość terminu gwarantuje wyróżnienie dokumentu wśród korpusu. Klasyczna postać TFIDF w ekstrakcji terminologii pełni przeważnie rolę drugorzędną, ponieważ fakt, iż termin nie jest wyjątkowy nie oznacza jego dyskwalifikacji.

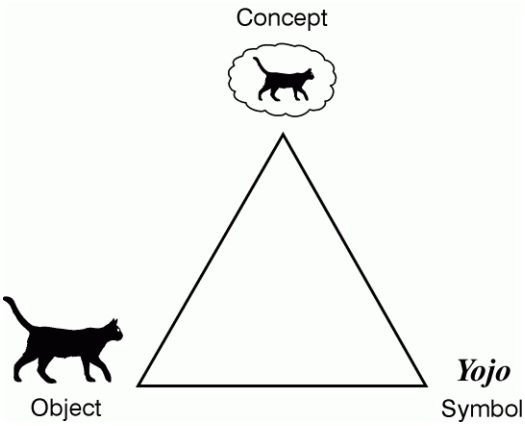
2.1.2 Synonimy

Drugim zadaniem w procesie uczenia ontologii z tekstu jest analiza terminów pod kątem ich wzajemnego podobieństwa. Podobieństwo jest cechą relatywną. Poziom podobieństwa jest zależny od kontekstu. Dla przykładu, terminy "urządzenie naręczne", "PDA", "smartphone" są bardzo podobne w kontekście dziedziny komórek macierzystych, natomiast z perspektywy dostawcy produktów na rynku handlu elektronicznego są to różne klasy produktów.

Synonimy o 100% wierności pojęciowej nie istnieją, są wyłącznie podobne do siebie wyrażenia będące są quasi-synonimami lub synonimami częściowymi przy spełnieniu określonych warunków brzegowych. Dlatego podobieństwa można mierzyć, a jego poziom jest specyficzny dla określonych warunków brzegowych (np. kontekstu dziedziny).

Podstawowymi metodami ekstrakcji synonimów są techniki klasyfikacji oraz analizy skupień (ang. *clustering*) terminów (Navigli i Velardi, 2005; Navigli, 2006b; Rinaldi i Yuste, 2005; Budanitsky i Hirst, 2006; Brody i in., 2006). Klasyfikacja polega na przyporządkowywaniu terminów do obecnych np. w WordNet (Fellbaum, 1998) klas. W analizie wykorzystywane są zbiory synonimów dla określonego terminu tzw. *synsets*. Analiza skupień stosowana jest do grupowania terminów pod względem dystrybucji ich cech, np. poprzez mierzenie ich współwystępowalności.

Faza ekstrakcji synonimów obejmuje również w przypadku analizy wielojęzycznej identyfikację synonimów w innych językach, czyli tłumaczeń (Grefenstette, 1998). Podobnie jak w przypadku synonimów, tak i w przypadku tłumaczeń, nie ma 100% wierności pojęciowej. Dla przykładu w Fisiak (2002)



Rysunek 2.2: Trójkąt znaczeniowy. Źródło: Ogden i Richards (1923); Sowa (2000b)

tłumaczenie polskiego wyrazu *mecz* na język angielski to *game*. Tłumaczenie odwrotne skutkuje natomiast polskim wyrazem *gra*.

W przypadku ekstrakcji tłumaczeń ważny jest fakt, że uczenie ontologii z tekstów wielojęzycznych jest złożeniem analiz z poszczególnych języków oraz zastosowanie filtrów, które zwiększają precyzję pozyskiwanych pojęć (Grefenstette, 1998). W konsekwencji nadal kluczowa jest miara efektywności ekstrakcji terminów z poszczególnych języków.

2.1.3 Pojęcia

W znanym trójkącie znaczeniowym (Ogden i Richards, 1923) pojęcie (ang. *concept*) jest według klasyfikacji wyższych poziomów ontologii formą abstrakcyjną (Sowa, 2000a). Pojęcie jest pewnym wyobrażeniem istniejących obiektów (ang. *object*), przedstawionym przy pomocy symboli (rysunek 2.2).

Interpretacja trójkąta znaczeniowego oparta jest na trzech poziomach abstrakcji. Konkretny czarny kot stanowi obiekt. Referowanie do obiektu następuje poprzez symbol. Symbol jest najczęściej wytworem języka reprezentacji bądź języka naturalnego. W przypadku rysunku 2.2 obiekt reprezentowany jest wyrazem *Yojo*. Reprezentacja symboliczna może być dowolna, ważne, że kojarzy obiekt. Klasa obiektów, tj. abstrakcyjna forma reprezentacji wszystkich obiektów o tych samych właściwościach, nazywana jest pojęciem. Z trójkątem znaczeniowym skojarzone są dwa problemy:

- problem niejednoznaczności oznaczający relację 1:n pomiędzy symbolem a pojęciem; jeden symbol może wskazywać na więcej niż jedno pojęcie (posiadać wiele znaczeń),

- problem niejednomianowości oznaczający relację n:1 pomiędzy symbolem a pojęciem; jedno pojęcie może być reprezentowane przez więcej niż jeden symbol (posiadać wiele tzw. *desygnatów pojęcia*).

Wieloznaczność jest wynikiem niedokładności powstałej w trakcie odwzorowania obiektów w pojęcia, a najczęstszą przyczyną jest niewłaściwy, niejednoznaczny symbol. W procesie uczenia ontologii z tekstu przetwarzane są symbole w postaci tekstu. Dobry mechanizm uczenia ontologii dąży do jak najdokładniejszego odwzorowania obiektów (przedstawionych przy pomocy symboli) w pojęcia.

Metody ekstrakcji pojęć z terminów wywodzą się z metod definicji form abstrakcyjnych przedstawionych w Sowa (2000a). Definiować abstrakcję można na dwa sposoby. Po pierwsze, pojęcie może być kombinacją dwóch lub więcej pojęć już istniejących. W każdym przypadku zastąpienie wyrazu *dziadek* wyrażeniem *ojciec ojca* jest semantycznie obojętne. Ten sposób definicji użyteczny jest wszędzie tam, gdzie wymagany jest szybszy, bardziej dokładny przekaz, a obecne pojęcia przekaz ten wydłużają. Sposób definicji pojęć poprzez ich składanie z pojęć już obecnych nazywany jest *definicją explicite*. Przeciwieństwem jest *definicja implicite*, która nie powstaje z konkretnego wyrażenia, lecz identyfikuje nowe, nienazwane pojęcie na podstawie pewnych ograniczeń lub reguł. Aby termin został potencjalnym kandydatem na pojęcie, musi spełniać warunki brzegowe. Popularnym narzędziem dla takiej klasyfikacji jest logika deskryptywna (Baader i in., 2003), przy pomocy której można klasyfikować nieznane byty na podstawie wartości ich własności. W konsekwencji reguły są tworzone automatycznie: w procesie uczenia, w momencie pojawienia się grupy podobnych do siebie instancji powstaje nowe pojęcie, które spełnia cechę podobieństwa tych instancji. Trzeci sposób na ekstrakcję pojęć wywodzi się już nie ze sposobów definicji form abstrakcyjnych, lecz z zastosowania analizy leksykalnej.

W fazie ekstrakcji pojęć w znakomitej większości podejść stosowane są źródła zewnętrzne w postaci tezaurusów, słowników lub ontologii. Zdecydowanie najpopularniejszym źródłem jest ontologia WordNet (Fellbaum, 1998) oraz jej narodowe odmiany (np. EuroWordnet (Vossen, 1998), GermaNet (Hamp i Feldweg, 1997)). Polska wersja ontologii WordNet jest (w trakcie pisania niniejszej pracy) opracowywana przez zespół dr Macieja Piaseckiego z Instytutu Informatyki Stosowanej na Politechnice Wrocławskiej[1].

[1] `http://plwordnet.pwr.wroc.pl/main/?cat=team`

2.1.4 Relacje taksonomiczne

Kolejną fazą cyklu uczenia ontologii z tekstu jest ekstrakcja relacji taksonomicznych, która dokonuje ekstrakcji relacji hiperonimicznych (bardziej ogólne byty w taksonomii) oraz hiponimicznych (bardziej szczegółowe byty w taksonomii). W obecnym dorobku nauki wyróżnia się szereg podejść do ekstrakcji tych dwóch kategorii:

- wzorce leksykalno-syntaktyczne,
- rozkład i analiza skupień,
- podejścia lingwistyczne,
- zawieranie się dokumentów,
- rozszerzenia taksonomii,
- podejścia łączone.

Wzorce leksykalno-syntaktyczne

Wzorce leksykalno-syntaktyczne jako metoda ekstrakcji relacji hiponimicznych zostały wprowadzone w Hearst (1992). Jego podejście opiera się na prostym założeniu, że uszczegółowione pojęcia występują w tekście często w podobnych wyrażeniach. Przykładem takich wyrażeń są:

> Produkty *takie jak* serwery, routery i światłowody.
> *Tak* dobre sklepy elektroniczne *jak* Komputronik i CK Znak.
> IBM, SAP, Microsoft *oraz inne* przedsiębiorstwa.
> Analiza taksonomiczna, a *w szczególności* analiza hiponimiczna.

Ogólną zasadą w tego typu analizie jest więc znalezienie instancji wzorców w tekście oraz ekstrakcja relacji typu *is-a*.

Od czasu publikacji pierwszych prac Hearsta powstało wiele podejść rozszerzających oryginalną pracę, m.in. (Alfonseca i Manandhar, 2002b; Kietz i in., 2000; Hearst, 1998). Podejścia te opierają się głównie na zwiększonej liczbie i precyzji reguł oraz zmianie przedmiotu analizy (np. ciekawa praca Sundblad (2003)).

Rozkład i analiza skupień

Drugie z wyróżnionych podejść opiera się na założeniu, że wyrazy są do siebie semantycznie podobne, jeśli występują w tym samym kontekście (Firth,

1957; Harris, 1986). W związku z tym przy ekstrakcji terminologii zbiera się również wyrazy sąsiadujące, które reprezentuje się jako wektor. Najprostsze z rozwiązań porównują powstałe wektory i mierzą miarę podobieństwa.

Podobieństwo wektorów można wyznaczyć przy pomocy trzech podejść do analizy skupień:

- opartej na analizie podobieństwa skupień,
- teorii zbiorów oraz prawdopodobieństwa,
- miękkiej analizy skupień.

Analiza z wykorzystaniem *podobieństwa skupień* w procesie ekstrakcji relacji taksonomicznych opiera się na trzech założeniach dotyczących sposobu mierzenia odległości, metody wiązania oraz algorytmu. *Odległość* pomiędzy skupieniami wyznaczyć można przy użyciu standardowych miar statystycznych, czyli: odległości euklidesowej, kwadratu odległości euklidesowej dla przypisania większej wagi obiektom od siebie najbardziej oddalonym lub odległości miejskiej (ang. *Manhattan distance*), która tłumi pojedyncze duże różnice. Wykorzystać również można inne miary, np. niezgodność procentową, która mierzy liczbę cech różnych w danych obiektach.

Posiadając wyliczone miary odległości pomiędzy obiektami należy połączyć je w skupienia. Aby połączyć skupienia o liczebności większej od 1, należy wybrać *metodę wiązania*. Skupienia można łączyć w zasadzie w sposób dowolny, jednak najbardziej rozpowszechnionymi strategiami łączenia są:

- metoda pojedynczego wiązania (najbliższego sąsiedztwa) — odległość między skupieniami określona jest na podstawie dwóch najbliżej położonych obiektów,
- metoda pełnego wiązania (najdalszego sąsiedztwa) — odległość skupień jest wynikiem odległości najdalej położonych obiektów obu analizowanych skupień,
- metoda średnich połączeń — odległość skupień to średnia odległości obiektów odpowiednich skupień,
- metoda średnich połączeń ważonych — odległość skupień wyznacza średnia ważona odległość obiektów w skupieniach, w których wagi są wyrazem liczebności danego skupienia,
- metoda środków ciężkości — odległość skupień to odległość od środków ciężkości skupień,

- metoda ważonych środków ciężkości — odległość skupień to ważona liczebnością skupienia odległość od środków ciężkości,
- metoda Warda — odległość skupień wyznaczana jest na podstawie minimalizacji sumy kwadratów odchyleń dowolnych dwóch skupień.

Oprócz miary odległości i metody łączenia w analizie skupień musi zostać odpowiednio dobrana strategia kierunku analizy. W klasycznej analizie statystycznej sytuacją wyjściową może być sytuacja, w której każdy obiekt stanowi osobne skupienie, bądź przeciwnie — na początku jest jedno duże skupienie zawierające wszystkie obiekty lub wszystkie sytuacje pośrednie, w którym zaczynamy analizę od pewnego stopnia klasyfikacji. Również sytuację wyjściową (docelową) można dowolnie ustalać, np. wychodząc od skupień jednoelementowych dążymy do jednego dużego skupienia (tzw. *metoda aglomeracji*), bądź odwrotnie. Można również zatrzymać analizę w dowolnym punkcie, celem doprowadzenia do dokładnie n-liczby skupień.

Analiza skupień w przypadku ekstrakcji relacji taksonomicznych najczęściej stosowana jest przy założeniu jednoelementowych skupień i dąży do uzyskania obiektów najbardziej podobnych przy zmiennej wartości progu. Wartość progu to wartość dobierana arbitralnie i oznacza wartość, przy której relacja może być sklasyfikowana jako taksonomiczna.

Wykrywanie relacji taksonomicznych metodą analizy skupień nastręcza jednak dużo problemów. Po pierwsze, powstające skupienia są nienazwane, co w przypadku celu ekstrakcji jest niedopuszczalne. Jedyne podejście do nazwania skupień (Caraballo, 1999, 2001) oparte jest na połączeniu analizy skupień ze wzorcami Hearsta. Po drugie, analiza skupień jest niestety podejściem statystycznym opartym na reprezentacji słownej. Oznacza to, że podobieństwo obiektów liczone przy pomocy tej metody jest często błędne. Istnieje w tej kwestii duża potrzeba uczestnictwa zewnętrznych mechanizmów, które zwiększą precyzję analizy skupień, co w zasadzie oznacza naprowadzenie strategii tworzenia skupień. Niedoskonałą metodę opisano w Cimiano i Staab (2005), która polega na znalezieniu przez eksperta relacji hiperonimicznej w stosunku do skupień. Skupienia zostają połączone tylko w przypadku, gdy mają taką samą relację hiperonimiczną.

Drugim sposobem na ekstrakcję relacji taksonomicznych z użyciem analizy skupień jest metoda znana pod nazwą FCA (ang. *Formal Concept Analysis*) (Ganter i Wille, 1999; Cimiano i in., 2005a). Metoda ta polega na analizie macierzy o n wierszach i kolumnach, gdzie n jest liczbą obiektów (tabela 2.2).

Analiza opiera się na stwierdzaniu przy każdej kombinacji obiektów począwszy od wierszy, czy zachodzi relacja hiponimiczna (*is-a*). Obiekt, który uzyska najmniejszą liczbę wystąpień relacji, traktowany jest jako pojęcie sto-

	Spółka	Spółka kapitałowa	Spółka akcyjna	Spółka osobowa
Spółka	X			
Spółka kapitałowa	X	X		
Spółka akcyjna	X	X	X	
Spółka osobowa	X			X

Tabela 2.2: Przykładowa analiza metodą FCA

jące w taksonomii najwyżej i zostaje usunięte z macierzy. Analiza powtarza się aż do wyczerpania wszystkich możliwości porównań.

Trzeci ze sposobów ekstrakcji relacji taksonomicznych z użyciem analizy skupień nazywa się *miękką analizą skupień*. Wykorzystuje ona analizę syntaktyczną. Niestety, terminy bardzo często mają niejedno znaczenie, co w tym przypadku oznacza przyporządkowanie obiektu do n-skupień. Na przykład, wyraz *mysz* zaklasyfikować można zarówno do skupienia *Sprzęt komputerowy*, jak i do *ssaki*. Wyzwaniem w tym przypadku jest rozpoznawanie wieloznaczności analizowanych pojęć (Yarowsky, 1992).

Podejścia lingwistyczne

Podejścia lingwistyczne do ekstrakcji relacji taksonomicznych wywodzą się z obserwacji dotyczących właściwości danego języka naturalnego. Zostało zauważone na przykład, że przymiotniki powiązane w zdaniu z rzeczownikami przeważnie zawężają zakres samego rzeczownika, tworząc w ten sposób relację taksonomiczną. Wyrażenia *spółka kapitałowa* i *spółka* to przykład klasycznej relacji taksonomicznej. Kolejnym popularnym podejściem jest leksykalne zawieranie się w sobie dwóch terminów (Buitelaar i in., 2004a; Velardi i in., 2001a) (por. sekcja 2.2.32 na stronie 73. Podejścia takie są stosowanie w narzędziach służących do analizy języka angielskiego. Rozbudowane narzędzia do analizy lingwistycznej dostępne są równiez dla tekstów w języku niemieckim (Piskorski, 2002; Xu i in., 2002; Piskorski i in., 2005) (por. np. sekcja 2.2.15 na stronie 52).

W związku z faktem, że analiza lingwistyczna jest specyficzna dla danego języka, analiza języka polskiego wymaga specyficznych podejść. Niestety, brakuje zarówno metod, jak i narzędzi do analizy języka naturalnego pod kątem ekstrakcji relacji taksonomicznych w procesie uczenia ontologii.

Zawieranie się dokumentów

Ekstrakcja relacji taksonomicznych odbywa się również zgodnie z założeniem, że jeżeli termin t_1 pojawia się we wszystkich dokumentach, w których występuje termin t_2, a także istnieją dokumenty, w których występuje wyłącznie t_1, to t_2 jest bardziej szczegółowym pojęciem (is-a(t_2, t_1)) (Sanderson i Croft, 1999). Pięć lat później podejście to zostało rozbudowane do postaci prawdopodobieństwa warunkowego (Fotzo i Gallinari, 2004). Kontynuując notację terminów:

$$P(t_2|t_1) = \frac{n(t_2, t_1)}{n(t_1)}. \tag{2.2}$$

Prawdopodobieństwo relacji hiperonimicznej t_2 i t_1 to stosunek liczby dokumentów, w których oba terminy współwystępują do liczby dokumentów, w których występuje wyłącznie termin t_1.

Rozszerzenia taksonomii

Wiele podejść do ekstrakcji relacji taksonomicznych opiera się na stosowaniu rozszerzeń do już istniejących metod (Widdows, 2003; Alfonseca i Manandhar, 2002a; Maedche, 2002; Witschel, 2005). Niestety, większość metod z tej grupy nie dokonuje ewaluacji własnego podejścia. Wynika to w dużej mierze ze złożoności problemu oraz z faktu, że jeżeli próbuje się kwantyfikować podejścia o różnych założeniach, to ich porównanie nie jest wiarygodne lub nawet możliwe. Wiąże się to z wielością i heterogenicznością korpusów, miar, a nawet samych ontologii użytych do ewaluacji.

Podejścia łączone

Różnorodność metod ekstrakcji relacji taksonomicznych generuje liczne próby łączenia różnych klas metod. W metodzie opisanej w Caraballo (1999) połączony został mechanizm nazywania skupień metodą aglomeracji z analizą leksykalno-syntaktyczną. Cederberg i Widdows (2003) wprowadzają mechanizm zwiększający dokładność oraz przedmiot analizy leksykalno-syntaktycznej. Cimiano i in. (2005b) proponują zastosowanie tzw. wyroczni eksperta związanej w relacjami hipernimicznymi do analizy strategii łączenia skupień.

Najprostszym w założeniu podejściem jest propozycja klasyfikacji podstawowowych metod (Cimiano i in., 2005b). Mechanizm dokonuje porównania rezultatów osiągniętych przy pomocy różnych metod i stosuje najbardziej pożądaną. Niestety, sam wybór metody musi zostać dokonany przez eksperta.

Metodę klasyfikacji taksonomii opartej na wielu pośrednich klasyfikatorach zastosowano w Snow i in. (2006). Autorzy skupiają się jednak na modelu

teoretycznym oraz rozszerzaniu ontologii WordNet, co powoduje, że metoda jest zbyt specyficzna do wykorzystania w innych warunkach.

2.1.5 Relacje nietaksonomiczne

Istniejące metody ekstrakcji relacji nietaksonomicznych wykorzystują analizę leksykalno-syntaktyczną oraz lingwistyczną. Najbardziej popularne metody opierają się na wykrywaniu *nazwanych relacji nietaksonomicznych*.

W ekstrakcji relacji nietaksonomicznych przeprowadzone badania odnoszą się głównie do czterech typów relacji nazwanych:

- relacja meronimiczna (ang. *part-of*), której zakres ekstrakcji jest pochodną definicji. Relacja ta może posiadać sens stricte fizyczny, np. stół i nogi, samochód i silnik. Pierwsza i zarazem najbardziej popularna metoda opiera się na wzorcu *X consists of Y* (Charniak i Berland, 1999),
- metody oparte na *strukturze qualia*, która definiuje cztery role wzorców: formalny, składowy, agentowy oraz celowy (Yamada i Baldwin, 2004; Cimiano i Wenderoth, 2005),
- metoda przyczynowa (ang. *causation*), polegająca na ekstrakcji wzorca *X leads to Y* (Girju i Moldovan, 2002),
- ekstrakcja atrybutów na zasadzie wzorca *the X of Y* (Poesio i Almuhareb, 2005).

Wykorzystanie struktury lingwistycznej tekstu umożliwia również ekstrakcję *nienazwanych relacji nietaksonomicznych*. Podejścia takie opierają się na przyporządkowywaniu występujących wzorców w tekście do odpowiednich struktur reprezentujących pojęcia w danych środowiskach. Na przykład, OntoLT (Buitelaar i Sintek, 2004) przyporządkowuje wzorzec lingwistyczny do elementów reprezentacji wiedzy opartej na ramkach (samej ramki, slotów, bądź zakresu slotu). Z kolei TextToOnto (Maedche i Staab, 2000a) pozwala na analizę zależności pomiędzy dwoma pojęciami, a następnie ich generalizację.

2.1.6 Reguły

Ostatnia warstwa cyklu uczenia ontologii polega na ekstrakcji informacji, która wskazuje na istnienie zależności aksjomatycznych w dziedzinie. Każda dziedzina posiada specyficzne aksjomaty, np. zatrudnienie na uczelni wyższej w Polsce powoduje przyporządkowanie do klasy pracowników dydaktycznych,

Ekstrakcja kolokacji

Referencje:	Smadja (1993)
Cel:	ekstrakcja kolokacji
Środek:	metody leksykalne i statystyczne
Wykorzystuje:	—
Rozszerzone w:	Xu i in. (2002)
Warstwa:	terminologia
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	bezpośrednio NIE
Ewaluacja:	ręczna, precyzja i zwrot

naukowych, technicznych lub administracyjnych. Prawidłowości te przedstawiać można w zasadzie w dowolnym logicznym języku reprezentacji, lecz najczęściej wykorzystywanym jest logika pierwszego rzędu, zwłaszcza logika Horna.

Lin i Pantel (2001) sugerują liczenie znaczących kolokacji na ścieżce zależności pojęć. Podejście opiera się na tezie, że jeżeli pomiędzy analizowanymi pojęciami w korpusie występuje duża liczba podobnych relacji, to być może istnieje relacja hiperonimiczna, która stanowi regułę dla tych pojęć. Z kolei Haase i Völker (2005) udowadniają skuteczność mechanizmu do ekstrakcji reguł rozłączności pomiędzy pojęciami na podstawie analizy statystycznej.

Niestety nie istnieje zbyt wiele metod ekstrakcji reguł dla uczenia ontologii z tekstu. Większość metod wykorzystuje już istniejące aksjomaty dziedzinowe do wzbogacenia procesów ekstrakcji innych warstw cyklu.

2.2 Metody uczenia ontologii z tekstu

Niniejsza sekcja obejmuje analizę dostępnych metod uczenia ontologii z tekstu. Przedstawienie metod zostało podzielone według warstwy, której metoda dotyczy, zgodnie z rysunkiem 2.1. Podobne zbiorcze zestawienia metod i narzędzi znaleźć można w Gómez-Pérez i Manzano-Macho (2004). Jest ono jednak nieaktualne i niepełne.

2.2.1 Ekstrakcja kolokacji

Metoda przedstawiona w Smadja (1993) była pierwszym tak obszernym podejściem do tematu ekstrakcji kolokacji, traktowanej przez niektórych autorów jako element ekstrakcji terminologii (Xu i in., 2002). Łączy ze sobą

podejścia do poszczególnych etapów ekstrakcji, czyli analizę leksykalną, składniową i statystyczną.

Kolokacja jest to grupa wyrazów występująca blisko siebie w tekście, posiadające znaczenie inne, niż by to wynikało ze znaczenia konkretnych wyrazów składających się na nią. Kolokacja w danym języku naturalnym lub dziedzinie jest wyjątkowo trudna do rozpoznania dla standardowych metod przetwarzania tekstów. Dla przykładu, *formacja niedźwiedzia* w dziedzinie rynków finansowych oznacza utrzymującą się tendencję spadkową. Oddzielnie oba te wyrazy mają zupełnie inne znaczenie. Idiomy występujące w językach naturalnych są klasycznym przykładem kolokacji trudno rozpoznawalnej zarówno dla standardowych metod przetwarzania tekstów, jak i dla cudzoziemców.

Metoda opisana w (Smadja, 1993) składa się z trzech etapów. Pierwszy etap dokonuje ekstrakcji powiązanych statystycznie terminów, np. poprzez zastosowanie analizy częstości współwystępowania wyrazów w tekście. Efektem fazy pierwszej są dwójki powiązanych terminów. Faza druga łączy *bigramy* w *n-gramy*, czyli n-powiązanych ze sobą terminów[2]. W trzeciej fazie dołączone zostają informacje lingwistyczne do potencjalnych kolokacji. Rezultatem są tylko te z nich, które spełniają warunek powtarzalności lingwistycznej, tj. jeżeli wszystkie zidentyfikowane przypadki *n-gramów* odpowiadają wzorcowi np. przymiotnik-rzeczownik. Przykładowo, efektem pierwszej fazy będzie wyrażenie "papierów wartościowych" (bez użycia analizatora morfologicznego), natomiast efektem drugiej fazy będzie "Warszawska Giełda Papierów Wartościowych". W związku z tym, że wyrazy wchodzące w skład tej kolokacji występują razem często przechodzą filtr w fazie trzeciej.

2.2.2 Połączenie analizy lingwistycznej i statystycznej

Metoda przedstawiona w Daille (1996) jest jedną z pierwszych prób połączenia analizy lingwistycznej z analizą statystyczną.

Zastosowana analiza lingwistyczna dotyczy języka francuskiego. Główny nacisk metody położony jest na ekstrakcję terminów wieloczłonowych. Terminy wieloczłonowe identyfikowane są poprzez wiele iteracji. Pojedyncza iteracja następuje zawsze poprzez łączenie dwóch członów. Zidentyfikowano trzy grupy połączeń dwóch członów dla ekstrakcji jednego terminu:

1. Kompozycja, która bądź to łączy dwa wyrazy bez zmiany jednego z nich, bądź zastępuje jeden termin nowym wyrażeniem.

[2]W tym miejscu pracy po raz pierwszy (nie licząc tez pracy) mowa jest o n-gramach. Termin ten w niniejszej pracy dotyczy całych wyrazów, w przeciwieństwie do np. klasycznej lingwistyki, w której przeważnie dotyczy ciągów liter.

Połączenie analizy lingwistycznej i statystycznej	
Referencje:	
Cel:	ekstrakcja terminologii
Środek:	połączenie analizy lingwistycznej i statystycznej
Wykorzystuje:	—
Rozszerzone w:	—
Warstwa:	terminologia
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	brak

2. Modyfikacja, która zmienia znaczenie terminu pod wpływem wyrażeń modyfikujących.

3. Koordynacja, która na podstawie dwóch terminów tworzy nowy termin będący złożeniem na podstawie wzorca lingwistycznego.

Każda z tych operacji jest specyficzna dla danego języka naturalnego. Następnie zastosowano filtr lingwistyczny, który klasyfikuje zidentyfikowane wyrażenia do dwóch grup: terminy i złe terminy. Złe terminy to takie terminy, które nie spełniają filtru statystycznego, a więc np. nie występują wystarczająco często. Na tym etapie dokonuje się identyfikacji najprostszych form kolokacji w celu ujednoznacznienia terminów.

Drugim etapem jest zastosowanie metod statystycznych do par wyrazów zidentyfikowanych w poprzednim etapie jako potencjalny termin. Zastosowano trzy miary częstości wystąpień:

1. Mutual Information, która zakłada, że wystąpienie jednego wyrazu zwiększa prawdopodobieństwa wystąpienia drugiego wyrazu:

$$I(x,y) = \frac{\log_2 P(x,y)}{P(x)P(y)}. \tag{2.3}$$

2. Log-Likelihood, która pokazuje, na ile prawdopodobne jest pojawienie się pary wyrazów x i y:

$$\begin{aligned} LogLike(x,y) = {} & a\log(a) + b\log(b) + c\log(c) + d\log(d) \\ & - (a+b)\log(a+b) - (a+c)\log(a+c) \\ & -(b+d)\log(b+d) - (c+d)\log(c+d) \\ & +(a+b+c+d)\log(a+b+c+d), \end{aligned} \tag{2.4}$$

Etykiety jakości	
Referencje:	Hahn i Schnattinger (1998)
Cel:	uczenie terminów
Środek:	redukcja przestrzeni hipotez
Wykorzystuje:	—
Rozszerzone w:	Hahn i Schulz (2000); Hahn i Markó (2001)
Warstwa:	terminologia
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	precyzja, zwrot, *parsimony*, *Learning accuracy*

gdzie a, b, c i d to elementy zbioru wszystkich kombinacji występowania i niewystępowania wyrazów x i y.

3. Łączna wydajność ϕ^2.

Wniosek z zastosowania wszystkich trzech miar wskazuje, że najlepsze wyniki uzyskano przy pomocy miary Log-Likelihood.

Dodatkowo zastosowaną miarą jest miara różnorodności (ang. *diversity*) wprowadzona przez Shannon (1948). Pokazuje ona rozkład formy podstawowej pojedynczego wyrazu w zbiorze wszystkich potencjalnych połączeń. Przy pomocy miary różnorodności można więc oszacować, czy dany wyraz wchodzi w relację wyłącznie z jednym wyrazem (wartość miary 0) czy takich powiązań istnieje wiele.

2.2.3 Etykiety jakości

Hahn i Schnattinger (1998) przedstawili jedną z pierwszych liczących się metod do ekstrakcji terminologii w procesie uczenia ontologii. Celem metody jest odkrywanie nowych terminów. W tym celu metoda stosuje analizę lingwistyczną oraz analizę obecnej struktury bazy wiedzy. Dla wszystkich dopuszczalnych możliwości klasyfikacji nieznanego pojęcia konstruuje się zbiór hipotez, które zostają w dalszej części weryfikowane poprzez definicję oraz estymację tzw. etykiet jakości (ang. *quality label*). Etykieta jakości lingwistycznej opiera się na poprawności lingwistycznej danej hipotezy, natomiast etykieta pojęciowa na poprawności strukturalnej (ontologicznej) hipotezy. Na podstawie oszacowanych wartości zbiór hipotez stopniowo się zmniejsza, aż do uzyskania jednej hipotezy, bądź przekroczenia zadanego progu jakościowego.

Ekstrakcja terminów wieloczłonowych miarą wartości C/NC	
Referencje:	Frantzi i in. (2000)
Cel:	uczenie terminów
Środek:	miara C-value / NC-value
Wykorzystuje:	—
Rozszerzone w:	Spasic i in. (2003); Nakagawa i Mori (2002, 2003); Ananiadou i Mcnaught (2006) Frantzi i Ananiadou (2007)
Warstwa:	terminologia
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	precyzja, zwrot

Metoda ta jest ciekawym podejściem teoretycznym. Niestety, mechanizmy lingwistyczne oraz terminologiczne są na nieweryfikowalnym poziomie z racji braku szczegółowego algorytmu oraz narzędzia reprezentującego metodę.

2.2.4 Metoda wartości C/NC

Metoda wartości C/NC (Frantzi i in., 2000) jest połączeniem analizy statystycznej z lingwistyczną. Została stworzona głównie w celu ekstrakcji terminów wieloczłonowych. Z powodzeniem jednak jest stosowana do ekstrakcji terminologii w ogóle. Metoda opiera się na dwóch miarach: wartości C oraz wartości NC. Wartość C oznacza statystyczną relewancję terminu, podczas gdy wartość NC zawiera w sobie kontekstowe właściwości terminu. Miara C/NC jest złożeniem obu wartości.

Metoda szacowania wartości C składa się z dwóch grup analiz: analizy lingwistycznej oraz analizy statystycznej. Analiza lingwistyczna zawiera trzy etapy:

Tagowanie polega na przyporządkowaniu znacznika części mowy do każdego wyrazu (standardowa anotacja POS).

Filtr lingwistyczny pozwala na dalszą analizę tylko pod warunkiem wystąpienia określonych właściwości lingwistycznych. Zastosowano filtr akceptujący sekwencje wyrazów będących przymiotnikiem, wyrażeniem lub rzeczownikiem i kończących się rzeczownikiem.

Lista wyrazów małoznaczących (tzw. stoplista) filtrująca zbiór analizowanych terminów. Listę dobrano eksperymentalnie nie podając szczegółów jej konstrukcji.

Główna część szacowania wartości C opiera się na analizie statystycznej, w której biorze się pod uwagę następujące cztery właściwości terminów złożonych:

1. Całkowitą częstość wystąpień kandydatów na terminy.
2. Częstość występowania kandydata na termin w innym, złożonym terminie.
3. Liczba zawierających termin terminów złożonych.
4. Długość kandydata na termin.

Na podstawie przedstawionych właściwości powstaje wartość C:

$$\text{C-value(a)} = \begin{cases} \log_2 |a| f(a) & \text{a niezagnieżdżony} \\ \log_2 |a| (f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)) & \text{a zagnieżdżony} \end{cases} \tag{2.5}$$

gdzie:
a to kandydat na termin,
$f(.)$ to liczba wystąpień w tekście,
T_a to zbiór terminów które zawierają a,
$P(T_a)$ to zbiór tych kandydatów na terminy.

Wartość NC bierze również pod uwagę wyrazy kontekstowe kandydata na termin. Analizowany jest wyraz bezpośrednio przed terminem i po terminie jeśli tylko spełniają kolejny filtr lingwistyczny. W tym przypadku analizowane są tylko i wyłącznie rzeczowniki, przymiotniki oraz czasowniki. Skład filtru dobrany został metodą eksperymentu. Następnie dla każdego wyrazu kontekstowego liczona jest waga w postaci:

$$waga(w) = \frac{t(w)}{n}, \tag{2.6}$$

gdzie:
w to wyraz kontekstowy,
$t(w)$ to liczba terminów, w których wystąpił wyraz kontekstowy w,
n to całkowita liczba terminów.

Na podstawie obliczonych wag dla wszystkich wyrazów kontekstowych można policzyć wartość NC według następującego wzoru:

$$\text{NC-value(a)} = 0.8\text{C-value(a)} + 0.2 \sum_{b \in C_a} f_a(b) waga(b), \tag{2.7}$$

Ekstrakcja terminów wieloczłonowych miarą FGM

Referencje:	Nakagawa i Mori (2003, 2002)
Cel:	uczenie terminów
Środek:	formowanie terminów wieloczłonowych
Wykorzystuje:	Nakagawa i Mori (1998)
Rozszerzone w:	—
Warstwa:	terminologia
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	precyzja, zwrot

gdzie:
a to kandydat na termin,
$C - value(a)$ jest wyliczone zgodnie z równaniem 2.5,
C_a to zbiór niepowtarzających się wyrazów kontekstowych dla a,
b to wyraz kontekstowy z C_a,
$f_a(b)$ to częstość b jako wyrazu kontekstowego a,
$waga(b)$ to waga wyrazu kontekstowego zgodnie z równaniem 2.6.

Miara oparta na wartości C/NC została porównana z klasycznymi metodami ekstrakcji terminologii opartych na częstości występowania wyrazów. W porównaniu z nimi już sama wartość C daje lepsze wyniki. Po zastosowaniu wartości NC wyniki są jeszcze lepsze i oscylują w granicach 70% precyzji. Niestety, nie podaje się miar zwrotu.

2.2.5 Miara FGM

Metoda C/NC (Frantzi i in., 2000) została w kilku pracach rozwinięta. Jedną z nich jest metoda przedstawiona w Nakagawa i Mori (2003, 2002), która zakłada, że większość terminów dziedzinowych to terminy wieloczłonowe.

Nakagawa i Mori (2003, 2002) proponują metodę, która najpierw oblicza częstość występowania danego jednowyrazowego terminu w innych wieloczłonowych terminach, a następnie wylicza wartość prawdopodobieństwa terminu. Funkcja obliczająca prawdopodobieństwo wystąpienia terminu jednowyrazowego jest sumą $n - tych$ potęg częstości jego występowania, przy czym parametr n jest poddany optymalizacji. Postać funkcji dla terminów wieloczłonowych jest średnią geometryczną funkcji dla terminów jednowyrazowych składających się na termin wieloczłonowy. Dodatkowo biorze się pod uwagę strukturę wystąpień terminu wieloczłonowego oraz wchodzących w jego skład terminów jednowyrazowych, a mianowicie to, czy terminy te występują w tekście razem (w tej samej grupie wyrazów) czy osobno (niezależnie

Wzorce komplementarności	
Referencje:	Spasic i in. (2003)
Cel:	uczenie pojęć
Środek:	wzorce komplementarności
Wykorzystuje:	Frantzi i in. (2000); Spasic i in. (2002)
Rozszerzone w:	Ananiadou i Mcnaught (2006); Frantzi i Ananiadou (2007); Nenadic i Ananiadou (2006)
Warstwa:	synonimy, pojęcia
Wykorzystywane ontologie:	dziedzinowa
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	precyzja, zwrot

od siebie). W wyniku tej obserwacji, średnia geometryczna jest pomnożona przez liczbę wystąpień niezależnych. W ten sposób powstaje miara FGM (ang. *Frequency Based Geometric Mean*).

Metoda oparta na mierze FGM porównana została z metodą wartości C w wersji z terminami jednoczłonowymi. Eksperymenty na testowym korpusie wykazały lekką przewagę miary FGM w przedziale do 1400 najwyżej sklasyfikowanych terminów.

Wskazuje się na duże podobieństwo metody wykorzystującej miarę FGM do metody opartej na wartości C. Analiza wyników pokazuje, że obie metody charakteryzują się podobnymi tendencjami, tj. relatywnie wysokim zwrotem przy wysokich poziomach precyzji.

2.2.6 Wzorce komplementarności

Wzorzec komplementarności to powiązanie pomiędzy czasownikiem dziedzinowym i zbiorem terminów. Czasownik dziedzinowy to czasownik charakterystyczny dla danej dziedziny, np. *świadczyć* dla dziedziny usług. Elementy te posiadają właściwość współwystępowania w tekstach dziedzinowych.

Metoda oparta na uczeniu wzorców komplementarności jest podstawą metody ekstrakcji synonimów i pojęć przedstawionej w Spasic i in. (2003). Metoda jest rozszerzeniem wcześniejszych prac tych samych autorów, głównie miary C/NC (Frantzi i in., 2000) oraz miary podobieństwa synonimicznego *CLS* (Spasic i in., 2002).

Metoda oparta na wzorcach komplementarności zakłada istnienie ontologii dziedzinowej. Ekstrakcja terminologii następuje przy użyciu klasycznej metody C/NC (Frantzi i in., 2000), której wyniki wyszukiwane są w ontologii. Następnie przy użyciu algorytmów genetycznych konstruowane są wzorce

Analiza odmian terminów

Referencje:	Nenadic i in. (2004)
Cel:	redukcja przestrzeni terminów
Środek:	normalizacja odmian terminów
Wykorzystuje:	Frantzi i in. (2000)
Rozszerzone w:	—
Warstwa:	terminologia
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	precyzja, zwrot

komplementarności, które wiążą znalezione pojęcia z ontologii z czasownikami z tekstu. W ten sposób uzyskane pary czasownik-pojęcia wraz z częstością ich występowania tworzą wzorce komplementarności. Jeżeli termin, który nie istnieje jeszcze w ontologii, współwystępuje z danym czasownikiem, zostaje częścią wzorca (uczenie wzorca) i w konsekwencji zostaje sklasyfikowany jako pojęcie w ontologii.

Ewaluacja metody została przeprowadzona z wykorzystaniem ontologii UMLS na korpusie biomedycznym. Średnie miary zwrotu, precyzji i miary F kształtowały się na poziomie odpowiednio: 12,2%, 63,83% oraz 20,48%.

Główną wadą rozwiązania jest nierealne wymaganie posiadania na wstępie względnie kompletnej ontologii. Wykorzystanie metody może dać dobre wyniki jedynie w rozszerzaniu lub ewolucji ontologii, co potwierdzają następne prace przedstawione przez Ananiadou i Mcnaught (2006); Frantzi i Ananiadou (2007); Nenadic i Ananiadou (2006).

2.2.7 Analiza odmian terminów

Różne postacie terminów są naturalną konsekwencją realizacji pojęć poprzez terminy. Terminy podlegają przemianom, takim jak fleksja, czy ortografia. W konsekwencji na jedno pojęcie wskazywać może wiele postaci terminów.

Nenadic i in. (2004) wskazują, że zastosowanie metod redukujących liczbę odmian terminów powoduje znaczący wzrost podstawowych miar ewaluacji. Na wykorzystanym korpusie biomedycznym udało się zwiększyć precyzję o 20-70%, a zwrot o 2-25% w porównaniu z wykorzystaniem form terminów w postaci różnych odmian.

W pracy Nenadic i in. (2004) wyróżnia się następujące typy odmian terminów:

- ortograficzne — np. wykorzystanie myślników, małych i wielkich liter lub używanie dopuszczalnie różnych form,

- morfologiczne — procesy fleksyjne, słowotwórcze oraz stosowanie wieloczłonowych nazw,
- leksykalne — np. synonimy,
- strukturalne — wykorzystanie łączników (np. *odmiana żeńska oraz męska*),
- akronimy i skróty — wykorzystanie różnych form skracania pełnej wersji terminu.

Wynikiem metody jest normalizacja postaci terminu. Normalizacja jest procesem łączenia wielu postaci terminu w tzw. *synterm*, czyli zbiór odmian tej samej gramatycznej formy podstawowej.

W celu uruchomienia metody wykorzystuje się ekstrakcję terminologii znaną z metody wartości C (Frantzi i in., 2000), czego wynikiem jest zbiór terminów w różnych odmianach. Następnie poprzez proces normalizacji powstają syntermy, dla których łącznie obliczana jest wartość C.

Proces normalizacji terminologii jest procesem nadzorowanym opartym na regułach. Oznacza to, że, dla każdego zaobserwowanego przypadku, opracowano reguły konwersji do formy podstawowej.

Metoda jest więc zależna od dziedziny oraz języka naturalnego, a jej skuteczność zależy od liczby i precyzji opracowanych reguł.

W znakomitej części innych metod w celu redukcji przestrzeni odmian terminów zastosowano bardziej praktyczne i prostsze rozwiązanie. W fazie ekstrakcji właściwej używa się form podstawowych, które często są elementem anotacji lingwistycznej.

2.2.8 Relewancja dziedzinowa

Metoda opisana w Velardi i in. (2001b) pozwala na wydzielenie z dokumentów terminów specyficznych dla danej dziedziny. Metoda wykorzystuje system NLP Ariosto + Chaos (Basili i in., 1996) do analizy lingwistycznej, która umożliwia rozpoznawanie bytów nazwanych oraz terminologii.

Przedstawiona analiza lingwistyczna opiera się na dwóch krokach. Pierwszym z nich jest narzędzie do rozpoznawania bytów nazwanych. Wykorzystanie dostępnych narzędzi umożliwia dość dobrą efektywność procesu. Dokonana weryfikacja rozpoznawalności bytów nazwanych wykazała efektywność na poziomie 89% miary F (Velardi i in., 2001b). Drugim krokiem jest analiza potencjalnych terminów niebędących bytami nazwanymi. Podczas tej fazy posłużono się podejściami opartymi na miarach *Mutual Information*, *czynnika Dice* oraz na miarach dystrybucji (*TFIDF*).

Relewancja dziedzinowa

Referencje:	Velardi i in. (2001b)
Cel:	ekstrakcja terminologii przy pomocy relewancji dziedzinowej
Środek:	miara relewancji dziedzinowej terminów
Wykorzystuje:	Basili i in. (1996)
Rozszerzone w:	Missikoff i in. (2002); Navigli i Velardi (2004)
Warstwa:	terminologia
Wykorzystywane ontologie:	—
Powiązane narzędzia:	OntoLearn
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna, miara F

Ekstrakcja terminologii dziedzinowych — KFIDF

Referencje:	Xu i in. (2002)
Cel:	ekstrakcja terminologii dziedzinowych
Środek:	miara KFIDF
Wykorzystuje:	Piskorski i Neumann (2000)
Rozszerzone w:	—
Warstwa:	terminologia
Wykorzystywane ontologie:	GermaNet (Hamp i Feldweg, 1997)
Powiązane narzędzia:	SPPC (Piskorski i Neumann, 2000)
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja, zwrot

Dla poprawy efektywności ekstrakcji terminologii wprowadza się miarę relewancji dziedzinowej terminu t:

$$DR(t, D_i) = \frac{P(t|D_i)}{\sum_{i=1...n} P(t|D_i)}, \tag{2.8}$$

gdzie:
n oznacza liczbę analizowanych dziedzin,
D_i to dziedzina ze zbioru $D_1, D_2, \ldots, D_n$,
$P(t/D_i)$ to prawdopodobieństwo warunkowe wystąpienia terminu t w dziedzinie D_i, liczone jako stosunek częstości wystąpienia terminu t w dziedzinie D_i do sumy wszystkich wystąpień terminu t.

Zmienność form terminów wieloczłonowych

Referencje:	Wermter i Hahn (2005b,a)
Cel:	ekstrakcja terminów wieloczłonowych
Środek:	miara P-Mod
Wykorzystuje:	—
Rozszerzone w:	—
Warstwa:	terminologia
Wykorzystywane ontologie:	—
Powiązane narzędzia:	(Tsuruoka i in., 2005); (Kudo i Matsumoto, 2003)
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja, zwrot

2.2.9 KFIDF

Metoda ekstrakcji terminologii relewantnej dla dziedziny przedstawiona w Xu i in. (2002) oparta jest na systemie przetwarzania języka niemieckiego SPPC (Shallow Processing Production Center) (Piskorski i Neumann, 2000) oraz ontologii GermaNet (Hamp i Feldweg, 1997). Metoda jest częścią systemu obejmującego fazy ekstrakcji terminologii, synonimów, pojęć oraz relacji taksonomicznych (Xu i in., 2002).

Xu i in. (2002) zaproponowali, zaimplementowali oraz zweryfikowali metodę ekstrakcji terminologii opartą na specyficznej odmianie miary TFIDF przeznaczoną do sklasyfikowanych dokumentów — KFIDF (Salton, 1991):

$$KFIDF(w, cat) = docs(w, cat) * \log(\frac{n * |cats|}{cats(w)} + 1), \tag{2.9}$$

gdzie:
$docs(w, cat)$ — liczba dokumentów w kategorii cat zawierająca wyraz w,
n — czynnik wygładzający,
$cats(w)$ — liczba kategorii zawierających wyraz w.

Miara KFIDF jest użyteczna w celu wyróżnienia wyrazów z kategorii w przypadku posiadania danych z co najmniej dwóch różnych kategorii. Metoda została zweryfikowana w trzech dziedzinach: zarządzania, rynków papierów wartościowych oraz przestępstw narkotykowych. Wyniki potwierdziły tezę, że miara KFIDF dobrze identyfikuje terminy dziedzinowe.

2.2.10 Zmienność form terminów wieloczłonowych

Metoda opisana w Wermter i Hahn (2005b) wskazuje, że terminy wieloczłonowe wykazują znacząco mniejszą tendencję do występowania w odmienionej

formie (tj. innej niż podstawowa forma). Prowadzi to do hipotezy, że sekwencje wyrazów często występujących w niezmiennej formie są dobrymi terminami.

Metoda należy do grupy metod statystycznych, wykorzystuje jednak filtr lingwistyczny, który filtruje z korpusu dokumentów tylko i wyłącznie wyrażenia rzeczownikowe oraz występujące w nich n-gramy o stopniach 2., 3. i 4. W wyniku tego procesu metoda dokonuje obliczeń spośród przygotowanych w ten sposób n-gramów. Dla celów anotacji POS wykorzystuje Tsuruoka i in. (2005), natomiast w celu anotacji wyrażeń Kudo i Matsumoto (2003).

W celu zmierzenia zmienności wykorzystuje się miarę *P-Mod*, która operuje na zmiennej $0 \leqslant k \leqslant n$, gdzie n to długość n-gramu, oznaczającej liczbę wyrazów modyfikowanych w n-gramie. Na przykład, dla n-gramu *long terminal repeat* i $k = 1$ wyróżnia się następujące wzorce:

k_1 terminal repeat,
long k_2 repeat,
long terminal k_3.

Zmienność n-gramu o zmiennej k wyrażona jest przy pomocy następującej zależności:

$$mod_k(ngram) = \prod_{i=1}^{s} \frac{f(ngram)}{f(sel_i, ngram)}, \tag{2.10}$$

gdzie:
s oznacza liczbę wzorców utworzonych przez k,
$f(ngram)$ oznacza liczbę n-gramów bez odmian,
$f(sel_i, ngram)$ oznacza stosunek $f(ngram)$ do liczby k-wzorca.

Miara zmienności form wynosi:

$$P - Mod(ngram) = \prod_{k=1}^{n} mod_k(ngram). \tag{2.11}$$

Na korpusie medycznym wykazano znaczącą poprawę skuteczności ekstrakcji terminów wieloczłonowych w porównaniu z miarami t-test oraz wartości C (Frantzi i in., 2000).

2.2.11 Metoda Hwanga

Metoda opisana w Hwang (1999) jest prostym mechanizmem ekstrakcji terminologii, relacji taksonomicznych oraz nietaksonomicznych na podstawie wskazań eksperta oraz narzędzi NLP.

Metoda Hwanga	
Referencje:	Hwang (1999)
Cel:	opracowanie prostego, ale całościowego systemu ekstrakcji
Środek:	wzorce powierzchni, analiza wyrażeń rzeczownikowych
Wykorzystuje:	—
Rozszerzone w:	—
Warstwa:	ekstrakcja terminologii, relacji taksonomicznych oraz nietaksonomicznych
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	brak

Faza ekstrakcji terminologii rozpoczyna się od wskazania przez eksperta przykładowych terminów występujących w dziedzinie. Na podstawie tych terminów oraz zidentyfikowanych w tekście wyrażeń rzeczownikowych identyfikowane są kolejne terminy. Hwang (1999) użył informacji lingwistycznej opartej na zależnościach relacyjnych zdania. Dla zidentyfikowanych wyrażeń rzeczownikowych zawierających wskazane wcześniej terminy dokonywana jest ekstrakcja nowych terminów. Na tym samym etapie metoda klasyfikuje relacje pomiędzy wyrazami wchodzącymi w skład wyrażenia rzeczownikowego. Wynikiem klasyfikacji jest relacja *is-a* (taksonomiczna) lub *assoc-with* (nietaksonomiczna, niezidentyfikowana), w zależności od struktury wyrażenia rzeczownikowego.

Metoda jest iteracyjna. Zidentyfikowane terminy w danej fazie są terminami początkowymi dla iteracji następnej. Liczbę faz może dowolnie definiować ekspert.

Ekstrakcja relacji nietaksonomicznych (ang. *assoc-with*) wymaga podjęcia decyzji przez eksperta na temat poprawności, kształtu oraz nazwy relacji.

2.2.12 Swiss Life

Metoda jest częścią środowiska On-To-Knowledge (Fensel i in., 2000) oraz TextToOnto (Maedche i Staab, 2000c,b), których przedmiot i zakres działania jest tyle obszerny, co nieprecyzyjny. Metoda opiera się na tekście naturalnym znajdującym się w słowniku firmy ubezpieczeniowej Swiss Life. Do każdego pojęcia w nim się znajdującym istnieje syntetyczny opis w języku naturalnym.

Swiss Life	
Referencje:	Kietz i in. (2000)
Cel:	ekstrakcja pojęć z wykorzystaniem GermaNet
Środek:	GermaNet, słownik organizacji Swiss Life
Wykorzystuje:	Maedche i Staab (2000b,c)
Rozszerzone w:	Cimiano i Völker (2005)
Warstwa:	ekstrakcja terminologii, synonimów, pojęć
Wykorzystywane ontologie:	GermaNet (Hamp i Feldweg, 1997) słownik organizacji Swiss Life
Powiązane narzędzia:	On-To-Knowledge, TextToOnto (Fensel i in., 2000; Maedche i Staab, 2000c,b)
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna, zwrot oraz precyzja

Metoda składa się z ekstrakcji terminologii z opisów pojęć w słowniku Swiss Life, porównania ich z zasobami pojęciowymi tezaurusa GermaNet, w tym synsetami oraz ekstrakcji pojęć na podstawie tych dwóch źródeł. Proces sterowany jest przez zbiór reguł, które określają jak konkretne pojęcia językowe (np. akronimy) rozwiązywane są do pojęć.

Opracowana metoda prezentuje ekstrakcję pojęć z sieci lokalnej, tj. ze zdefiniowanego słownika organizacji. Pod pojęciem intranetu ukrywa się największa wada rozwiązania — służy ono wyłącznie do ekstrakcji pojęć ze zdefiniowanych źródeł. Niestety, nie pokuszono się o ewaluację metody bez bardzo ułatwiających zadanie źródeł pomocniczych, czyli w tym przypadku GermaNet oraz słownika Swiss Life.

Należy jednak odnotować próbę i wkład do stanu wiedzy na etapie ekstrakcji synonimów oraz pojęć z użyciem niemieckiej odmiany słownika WordNet.

2.2.13 Ekstrakcja produktów IT

Metoda przedstawiona przez Holzinger i in. (2006) jest specyficzną metodą ekstrakcji terminologii oraz ich wartości dla dziedziny sprzętu komputerowego. Celem jest dostarczenie informacji na temat produktów IT, ich cech oraz wartości cech na podstawie półustrukturyzowanej informacji w postaci tabel na stronach HTML.

Architektura podejścia opiera się na na algorytmie ekstrakcji tabel, który na podstawie ontologii tabel (wiersze, kolumny, komórki i dopuszczalne wartości) pozyskuje ze stron HTML sformalizowaną postać tabel. Następnie istniejąca ontologia treści (ontologia produktów) wykorzystywana jest do po-

Ekstrakcja produktów IT

Referencje:	Holzinger i in. (2006)
Cel:	uczenie terminów
Środek:	ekstrakcja tabel HTML z produktami IT
Wykorzystuje:	—
Rozszerzone w:	Khandelwal (2007)
Warstwa:	terminologia
Wykorzystywane ontologie:	dziedzinowa
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	precyzja, zwrot

Ekstrakcja akronimów

Referencje:	Okazaki i Ananiadou (2006), podobne: Schwartz i Hearst (2003); Adar (2004); Torii i in. (2006); Chang i Schutze (2006)
Cel:	ekstrakcja par akronim-definicja
Środek:	C/NC, miara współwystępowalności
Wykorzystuje:	Schwartz i Hearst (2003); Frantzi i in. (2000)
Rozszerzone w:	—
Warstwa:	terminologia
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	precyzja, zwrot

brania terminów odpowiadających przewidzianej strukturze. Pozyskane terminy są anotowane wystąpieniami pojęć z ontologii treści. Następnie terminy zostają dodawane do istniejącej ontologii wraz z wypełnieniem odpowiednich właściwości i ich wartości.

Ewaluacja metody została przeprowadzona na tekstach opisujących aparaty cyfrowe. Miara F kształtowała się na poziomach: dla terminów 64,27%, dla wartości 48,90% oraz dla właściwości produktów 79,72%. Metoda została komercyjnie wykorzystana przez Khandelwal (2007).

Przedstawiona metoda jest interesująca ze względu na dziedzinę. Niestety, zastosowane podejście klasyfikuje ją bardziej jako metodę uczenia ze źródeł półustrukturyzowanych.

2.2.14 Ekstrakcja akronimów

Stosowane podejścia do ekstrakcji akronimów polegają na konstrukcji wzorców syntaktycznych, które umożliwiają ekstrakcję akronimu wraz z jego rozwinięciem stanowiącym jednocześnie jego definicję (Schwartz i Hearst, 2003; Adar, 2004; Torii i in., 2006; Chang i Schutze, 2006). W związku z powszechnością stosowania akronimów w dziedzinie biomedycznej, większość metod jest dostosowana do tej dziedziny.

Metoda ekstrakcji akronimów przedstawiona przez Okazaki i Ananiadou (2006) jest wyjątkowa ze względu na fakt, że w celu ekstrakcji akronimów wykorzystano bardziej ogólną metodę ekstrakcji terminologii.

Podejście składa się z trzech etapów:

Drążenie akronimów. Pierwszy etap polega na ekstrakcji potencjalnych akronimów, tzw. *krótkich form*. W tym celu wykorzystuje się metodę opisaną przez Schwartz i Hearst (2003), która polega na ekstrakcji przy pomocy zdefiniowanych reguł postaci: *długa forma (krótka forma)*.

Drążenie definicji. Drugi etap polega na analizie kontekstu krótkiej formy przy pomocy zmodyfikowanej metody wartości C (Frantzi i in., 2000). Następnie liczona jest miara współwystępowalności dla terminów pozyskanych metodą wartości C i wybrane zostają te, które charakteryzują się największymi wartościami miary współwystępowalności. W ten sposób dla każdej krótkiej formy przyporządkowany jest zbiór tzw. *długich form*.

Walidacja definicji. Ostatnim etapem jest analiza długich form dla każdej krótkiej formy. Polega ona na zastosowaniu dodatkowych ograniczeń syntaktycznych, np. konieczności wystąpienia w długiej formie wszystkich liter krótkiej formy.

Ewaluacja metody została przeprowadzaona na korpusie składającym się z abstraktów biblioteki MEDLINE. Uzyskane wyniki precyzji i zwrotu (78% i 85%) pozwalają stwierdzić, że standardowe metody ekstrakcji terminologii mogą być wykorzystane w celu ekstrakcji akronimów.

2.2.15 Kolokacja w językach o swobodnym szyku zdania

Zgodnie z Smadja (1993) kolokacja to często używane zestawienie wyrazów, w którym znaczenie całości wynika ze znaczeń poszczególnych wyrazów.

Metoda kolokacji terminów przedstawiona w Xu i in. (2002) oparta jest na systemie przetwarzania języka niemieckiego SPPC (Shallow Processing Production Center) (Piskorski i Neumann, 2000) oraz ontologii GermaNet

Kolokacja terminów w językach o swobodnym szyku zdania

Referencje:	Xu i in. (2002)
Cel:	kolokacja terminów w celu wykrycia wieloczłonowych terminów oraz relacji taksonomicznych i nietaksonomicznych dla języków o dowolnym szyku zdania
Środek:	wykorzystanie miar Mutual Information, Log-Likelihood oraz t-test studenta
Wykorzystuje:	Smadja (1993); Finkelstein-Landau i Morin (1999); Piskorski i Neumann (2000)
Rozszerzone w:	—
Warstwa:	ekstrakcja terminologii, ekstrakcja relacji taksonomicznych i nietaksonomicznych
Wykorzystywane ontologie:	GermaNet (Hamp i Feldweg, 1997)
Powiązane narzędzia:	SPPC (Piskorski i Neumann, 2000)
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja, zwrot

(Hamp i Feldweg, 1997). Metoda jest częścią systemu obejmującego fazy ekstrakcji terminologii, synonimów, pojęć oraz relacji taksonomicznych (Xu i in., 2002).

Celem metody jest identyfikacja wieloczłonowych terminów (składających się z więcej niż jednego wyrazu) oraz ekstrakcja wzorców syntaktycznych zawierających terminy powiązane relacjami semantycznymi. Podejścia statystyczne polegające na analizie współwystępowalności oparte są na analizie par, trójek, itd. terminów połączonych w zdefiniowany przez gramatykę danego języka sposób. Niestety istnieje pewna grupa języków, w których dopuszcza się stosowanie dość dużej swobody w porządku wyrazów w zdaniu. Należą do nich na przykład język niemiecki oraz język polski. Dlatego dla celów analizy konstruuje się zbiór wszystkich kombinacji terminów w danym wyrażeniu. Na podstawie tych kombinacji obliczają miary statystyczne:

1. Mutual Information, która zakłada, że wystąpienie jednego wyrazu zwiększa prawdopodobieństwa wystąpienia drugiego wyrazu (równanie 2.3).

2. Log-Likelihood, która pokazuje na ile prawdopodobne jest pojawienie się pary wyrazów x i y (równanie 2.4).

Ekstrakcja pojęć z wykorzystaniem podpisu tematu

Referencje:	Agirre i in. (2000)
Cel:	rozszerzenie listy pojęć powiązanych
Środek:	*Topic signatures*
Wykorzystuje:	Lin i Hovy (2000); Hovy i Lin (1999)
Rozszerzone w:	—
Warstwa:	ekstrakcja synonimów, ekstrakcja pojęć
Wykorzystywane ontologie:	WordNet (Fellbaum, 1998)
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna, precyzja

3. T-test studenta, który pokazuje na ile prawdopodobne jest pojawienie się w tekście określonego wzorca:

$$T = \frac{x - \mu}{\sqrt{\frac{s^2}{N}}}, \tag{2.12}$$

gdzie:
x oznacza średnią próby,
μ to średnia całości korpusu,
s^2 to wariancja próby,
N to rozmiar próby.

Obserwacje potwierdzają tezę, że miara LogLike daje najlepszą precyzję dla danych o małej częstości występowania. Ponadto miary sprawdzają się różnie w zależności od wzorca powierzchni, tj. miara oparta na statystyce *t-test studenta* daje najlepsze wyniki dla wzorca przyimek-rzeczownik-czasownik, z kolei *LogLike* dla wzorca przymiotnik-rzeczownik. Dla małych korpusów *LogLike* sprawdza się najlepiej.

Przy znacznej liczbie powtarzających się kolokacji charakteryzujących się jednakowym wzorcem można przypuszczać, że jest to wzorzec ogólny. W ten sposób bez udziału eksperta powstają wzorce powierzchni.

Duża miara kolokacji terminów wskazuje również możliwość istnienia relacji semantycznej pomiędzy terminami.

2.2.16 Podpis tematu

Metoda Agirre i in. (2000) ma na celu rozszerzanie listy pojęć terminami występującymi w tekście razem z pojęciem. Geneza podejścia wywodzi się z niedoskonałości ontologii WordNet, czyli braku powiązań nietaksonomicznych pomiędzy synsetami oraz mnożenie znaczeń pojęć (ang. *word senses*).

Oryginalna metoda podpisu tematu opiera się tylko i wyłącznie na przesłance zniesienia wymienionych niedogodności ontologii WordNet. Metoda wykorzystuje algorytmy podpisu tematów (ang. *topic signatures*) (Lin i Hovy, 2000) zastosowane z powodzeniem do automatycznego podsumowywania tekstów (Hovy i Lin, 1999).

Działanie metody podpisu tematu rozpoczyna się od pobrania analizowanego terminu z ontologii WordNet. Następnie pobierane są wszystkie znaczenia skojarzone z danym synsetem. Dla przykładu, dla terminu *business* wyszukane zostanie 9 znaczeń. Dla zwięzłości wywodu omówione zostaną wybrane trzy:

- business, concern, business concern, business organization, business organisation (a commercial or industrial enterprise and the people who constitute it) "he bought his brother's business"; "a small mom-and-pop business"; "a racially integrated business concern",
- business, business sector (business concerns collectively) "Government and business could not agree",
- occupation, business, job, line of work (the principal activity in your life that you do to earn money) "he's not in my line of business".

Każde z tych znaczeń jest *de facto* innym pojęciem, pomimo że odnoszą się do tego samego terminu. Dlatego metoda ta jest wykorzystywana w fazie ekstrakcji pojęć dla naniesienia do ontologii pojęć z zadanej listy terminów.

Dla każdego obiektu z przykładowej listy znaczeń wraz z ich opisem (ang. *sense + information*) konstruowane jest zapytanie, które jest przekazywane do wyszukiwarki w celu odnalezienia dokumentów zawierających analizowane terminy. Następnie na podstawie kolekcji dokumentów dla każdego terminu oraz algorytmu podpisu tematu konstruowana jest lista rozszerzająca.

Zgodnie z Hovy i Lin (1999); Lin i Hovy (2000) podpis tematu to rodzina powiązanych terminów:

$$t, < (w_1, s_1) \ldots (w_i, s_i) >, \tag{2.13}$$

gdzie t to temat (analizowany termin), a w_i to powiązany termin z siłą powiązania s_i.

Powiązane terminy to wszystkie terminy występujące w kolekcji dokumentów uzyskanej przy pomocy wyszukiwania. Siła powiązania mierzona jest miarą premiującą liczną występowalność terminu w jednym dokumencie w stosunku do całej kolekcji. Miara ta zdefiniowana została w Lin (1997) i przypomina znaną miarę TFIDF (równanie 2.1).

Wieloznaczność pojęciowa	
Referencje:	Yarowsky (1992)
Cel:	analiza niejednoznaczności pojęciowej
Środek:	analiza kontekstu tekstu, tezaurus Roget's thesaurus
Wykorzystuje:	—
Rozszerzone w:	Faatz i Steinmetz (2002)
Warstwa:	ekstrakcja pojęć
Wykorzystywane ontologie:	*Roget's Thesaurus*
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna, dokładność odwzorowania dla znaczenia i ogólna

2.2.17 Wieloznaczność pojęciowa

W wielu metodach ekstrakcji pojęć (np. Faatz i Steinmetz (2002)) w celu usunięcia niejednoznaczności pojęciowej używa się metody przedstawionej w Yarowsky (1992). Metoda ta pomaga z dużą precyzją dokonać klasyfikacji wyrazu do zbioru jego znaczeń. W przypadku omawianej metody użyty został tezaurus *Roget's Thesaurus*, który był niezwykle popularnym i kompleksowym źródłem relacji synonimicznych do czasu pojawienia się ontologii WordNet. Obecnie znaczenie i użycie *Roget's Thesaurus* znacznie spadło (Ozsu i Snodgrass, 2001).

Metoda polega na trzech następujących po sobie krokach:

1. Wyznaczenie kontekstu dla analizowanego wyrazu.

2. Identyfikacja najistotniejszych wyrazów z kontekstu oraz obliczenie wag dla każdego z nich.

3. Użycie obliczonych wag do rozpoznania odpowiedniego znaczenia dla wieloznacznych terminów występujących w tekście.

Metoda była jedną z pierwszych podejść do identyfikacji niejednoznaczności na podstawie klasyfikacji, dlatego wszystkie trzy kroki opierają się na niezwykle prostych algorytmach. Kontekstem danego wyrazu nazywana jest określona z góry liczba wyrazów poprzedzających oraz następujących po analizowanym wyrazie w tekście. Waga wyrazu z kroku 2. wyliczania jest przez podzielenie prawdopodobieństwa wystąpienia wyrazu w korpusie tezaurusa przez prawdopodobieństwo wystąpienia w korpusie. W kroku 3. analizowany jest kontekst badanego wyrazu, a następnie znaczenie o największej wartości podobieństwa jest wybierane jako najbardziej odpowiednie.

Wykorzystanie grafów pojęciowych

Referencje:	Roux i in. (2000)
Cel:	ekstrakcja nowych pojęć
Środek:	reprezentacja przy pomocy grafów pojęciowych
Wykorzystuje:	A'it-Mokhtar i Chanod (1997); Sowa (1984)
Rozszerzone w:	—
Warstwa:	ekstrakcja pojęć
Wykorzystywane ontologie:	ontologia początkowa
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	brak

2.2.18 Wykorzystanie grafów pojęciowych

System ekstrakcji informacji przedstawiony w Roux i in. (2000) wprowadza połączenie analizy lingwistycznej z reprezentacją modelu przy pomocy *grafów pojęciowych* (ang. *conceptual graphs*) (Sowa, 1984).

Punktem wyjścia metody jest ontologia opisana przy pomocy grafu pojęciowego. System na podstawie tekstu potrafi dokonać ekstrakcji pojęć dotyczących relacji już w ontologii się znajdujących. Metoda korzysta z systemu NLP przedstawionego w A'it-Mokhtar i Chanod (1997). W przypadku ekstrakcji nowego pojęcia w ontologii sprawdzana jest relacja, która go dotyczy. Jeżeli nie ma bezpośredniego związku z pojęciem stanowiącym obiekt relacji, to nowe pojęcie jest dodawane jako następny obiekt danej relacji. W przypadku, gdy występuje relacja pomiędzy starym a nowym pojęciem (np. wynikająca ze struktury relacja taksonomiczna), stary lub nowy obiekt jest zastępowany.

Przedstawiona metoda jest niezwykle prosta. Posiada również szereg ograniczeń. Na przykład można dokonać ekstrakcji tylko pojęć, które są połączone relacją występującą już w bazie wiedzy. System nie może dokonać ekstrakcji relacji. Charakteryzuje się więc zapewne (praca nie przytacza ewaluacji) wysoką precyzją i niskim zwrotem. Do poprawnego funkcjonowania potrzebuje na starcie dość rozbudowanej ontologii. W zasadzie więc służy wyłącznie do operacji dodawania małych porcji nowych pojęć.

Nowością metody jest zastosowanie jej w dziedzinie genetyki oraz przedstawienie wzorców oraz nowych pojęć przy pomocy grafów pojęciowych.

Ekstrakcja nawiązań	
Referencje:	Poesio i in. (2002)
Cel:	automatyczne tworzenie leksykonu
Środek:	łączenie nawiązań w tekście
Wykorzystuje:	Vieira i Poesio (2000); Fellbaum (1998); Mikheev i in. (1999)
Rozszerzone w:	—
Warstwa:	ekstrakcja pojęć
Wykorzystywane ontologie:	WordNet
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja i zwrot

2.2.19 Ekstrakcja nawiązań

Nawiązanie (ang. *anaphora*) to pojawiające się w tekście wyrażenie, które odnosi się do wcześniej wprowadzonych terminów. Ekstrakcja nawiązań (ang. *anaphora resolution*) to identyfikacja terminu oraz informacji, która go dotyczy, a znajduje się niekoniecznie w jego bezpośrednim sąsiedztwie. W praktyce oznacza, że nie jest powiązana informacją lingwistyczną, czyli nie należy np. do jednego wyrażenia rzeczownikowego. Odległość nawiązania może być różna. To samo zdanie, dwa sąsiadujące zdania, a czasem nawet różne akapity tekstu. Ograniczeniem odległości jest tzw. okno, w terminologii lingwistycznej oznaczające analizowany na raz fragment tekstu. W znakomitej jednak większości analizowane nawiązania ograniczają się do przypadku najczęściej występującego, czyli nawiązania, które występuje na początku następnego zdania.

Metoda opisana w Poesio i in. (2002) jest zbiorem badań prowadzonych m.in. w Vieira i Poesio (2000) w celu automatycznej konstrukcji leksykonów. Motywacją są znane błędy ontologii WordNet, przede wszystkim jej nadmierne rozróżnianie znaczeń oraz brak ukontekstowienia. Ekstrakcja nawiązań jest jedną z metod automatycznej konstrukcji leksykonów.

Ekstrakcja nawiązań metodą przedstawioną rozpoczyna się od ekstrakcji opisów bezwzględnych (Vieira i Poesio, 2000). Na ich podstawie zostaje opracowany zestaw 204 wzorców łączących informację powiązaną. Zaproponowane algorytmy łączą informację pozyskaną z ontologii WordNet (Fellbaum, 1998) oraz opracowane heurystyki (Vieira i Poesio, 2000).

Niestety, po zastosowaniu metody okazało się, że uzyskane wyniki nie są satysfakcjonujące dla żadnej aplikacji (precyzja i zwrot oscylowały w przedziale 50-70%). Rozszerzeniem metody było zastosowanie łączenia nawiązań dla bytów nazwanych, w których nawiązanie występowało w określonym kon-

Wzbogacanie metodą kolokatorów	
Referencje:	Faatz i Steinmetz (2002)
Cel:	ekstrakcja pojęć
Środek:	analiza statystyczna
Wykorzystuje:	Yarowsky (1992)
Rozszerzone w:	—
Warstwa:	ekstrakcja pojęć
Wykorzystywane ontologie:	ontologia medyczna powstała w ramach projektu k-med
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	ręczna

tekście zdania, tj. na początku następnego zdania. W celu ekstrakcji bytów nazwanych wykorzystano mechanizm przedstawiony w Mikheev i in. (1999). Wyniki dla takiego zadania były już dużo lepsze (precyzja 95%, zwrot 66%).

Dużym wkładem metody jest przeprowadzenie analizy wrażliwości dla parametrów lingwistycznych takich jak: rozmiar okna, ważność odległości analizowanej informacji lingwistycznej, typ analizy korpusu (parametry lematyzacji, anotowania, etc.).

2.2.20 Wzbogacanie metodą kolokatorów

Metoda opisana w Faatz i Steinmetz (2002) ma na celu wzbogacenie już istniejącej ontologii pojęciami wydobytymi z korpusów dokumentów. Na podstawie kontekstu występowania w dokumentach pojęcia oraz reguł syntaktycznych identyfikowane są tzw. *kolokatory* oraz mierzona zostaje odległość, która reprezentuje podobieństwo sugerowanego kandydata do ontologii. Potencjalnie interesujące nowe pojęcia prezentowane są ekspertowi do ręcznej analizy.

Proces wzbogacania ontologii składa się z trzech kroków:

1. Definicja korpusu dokumentów, który w ramach eksperymentów składał się bądź to z ogólnego korpusu, bądź z dziedzinowego pozyskanego w wyniku zapytania skierowanego do systemu wyszukiwawczego.

2. Analiza oraz identyfikacja pojęć-kandydatów. Głównym zadaniem jest odkrycie reguł, które nie występują w ontologii, a regulują wzorce występowalności pojęć ze sobą powiązanych. Przy tym, respektuje się wyłącznie relacje taksonomiczne w istniejących już ontologiach. Na podstawie analizy statystycznej opartej na kontekście pojęcia w tekście

Semantyczna interpretacja złożonych terminów	
Referencje:	Missikoff i in. (2002); Navigli i Velardi (2005)
Cel:	ekstrakcja pojęć z terminów wieloczłonowych
Środek:	analiza relacji pomiędzy znaczeniami ontologii WordNet
Wykorzystuje:	Basili i in. (1996); Velardi i in. (2001a,b)
Rozszerzone w:	Navigli i Velardi (2004); Navigli (2006b); Brody i in. (2006)
Warstwa:	ekstrakcja pojęć
Wykorzystywane ontologie:	WordNet, ontologia wstępna
Powiązane narzędzia:	OntoLearn
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja i zwrot

mierzona jest odległość do pojęcia. Kontekst pojęcia może być zdefiniowany np. poprzez liczbę wyrazów lub kontekst zdania.

3. Wybór relewantnych pojęć-kandydatów przez eksperta.

2.2.21 Semantyczna interpretacja złożonych terminów

Jedną z części procesu uczenia ontologii przedstawionego w Missikoff i in. (2002) jest metoda identyfikacji pojęć na podstawie semantycznej interpretacji złożonych terminów.

Metoda opiera się na złożonych terminach (czyli wieloczłonowych), które pozyskiwane są z dwóch źródeł: korpusu specyficznego dla dziedziny turystycznej oraz korpusie ogólnym. Korpus specyficzny służy do ekstrakcji kandydatów na terminy, które następnie porównywane są z korpusem ogólnym, dobieranym tak, aby wyróżnić te, które są specyficzne dla dziedziny. Porównania z korpusem ogólnym nazywane jest filtrowaniem dziedzinowym dokumentów. W obu typach ekstrakcji użyty został system do przetwarzania języka naturalnego Ariosto+Chaos (Basili i in., 1996).

W kolejnej fazie wyselekcjonowane terminy są interpretowane semantycznie. Semantyczna interpretacja terminów polega na przypisaniu do danego terminu odpowiedniej nazwy pojęcia. Proces przypisywania rozszerza każdy złożony termin zgodnie z przypisanym zbiorem synsetów ontologii WordNet. Następnie tworzone są wszystkie możliwe kombinacje znaczeń pojedynczych członów złożonego terminu. Dla przykładu znaczenie terminu *transport company* rozwiązywane jest do postaci:

S("transport company") = transportation, shipping, transport, company

zgodnie z faktem, że WordNet zawiera trzy znaczenia wyrazu *transport* i jedno wyrazu *company*. Dla każdej pary liczona jest miara na podstawie opracowanych wzorców. Wzorce wykorzystują dostępne elementy ontologii WordNet i mogą opierać się na:

- tzw. *glosses*, czyli wpisach słownikowych zawierających syntetyczną definicję wyrazu,
- relacji taksonomicznych, czyli hiperonimicznych oraz hiponimicznych,
- relacjach meronimicznych oraz holonimicznych, czyli relacjach odwrotnych oznaczających "jest częścią",
- miarach podobieństwa, czyli relacjach synonimicznej (dokładne znaczenie), podobieństwa oraz korelacji (oznaczających mniejszy stopień podobieństwa), a także relacji odwrotnej, tj. antonimicznej. Ponadto wykorzystywana jest relacja pertonimiczna, która łączy wystąpienia wyrazu w postaci rzeczownikowej i przymiotnikowej (np. matka i matczyny),
- relacji powiązania semantycznego, która oznacza powiązania terminów występujących w ontologii WordNet.

Przykładowy wzorzec oparty na wpisie słownikowym definiuje wystąpienie jednego wyrazu w rozszerzeniu drugiego jako znaczący element podwyższający miarę powiązania obu terminów. Szczegółowej definicji wzorców nie ujawnia się.

Wadą analizowanej metody jest fakt, że ma ona jakikolwiek sens zastosowania tylko w przypadku terminów złożonych. Pomimo, że może być zastosowana również dla jednoczłonowych terminów, interpretacja semantyczna traci w takim przypadku sens, ponieważ miara podobieństwa nie może zostać wyliczona.

2.2.22 Wzorce Hearsta

Metoda Hearsta wywodzi się z niedoskonałości ontologii WordNet i próbie rozszerzenia liczby relacji w nim istniejących (Hearst, 1998). Metoda dotyczy ekstrakcji relacji hiponimicznych, która zdefiniowana jest w WordNet jako:

Wzorce Hearsta	
Referencje:	Hearst (1992, 1998)
Cel:	ekstrakcja wzorców i relacji hiperonimicznych
Środek:	wzorce ekstrakcji
Wykorzystuje:	—
Rozszerzone w:	Kietz i in. (2000); Alfonseca i Manandhar (2002b); Charniak i Berland (1999); Finkelstein-Landau i Morin (1999); Caraballo (1999, 2001); Iwanska i in. (2000); Xu i in. (2002); Poesio i in. (2002); Markert i in. (2003); Sundblad (2003); Cimiano i Staab (2005); Cimiano i in. (2005b); Cimiano (2006)
Warstwa:	relacje taksonomiczne
Wykorzystywane ontologie:	WordNet
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	wersja ogólna TAK, wersja szczególna NIE
Ewaluacja:	ręczna

Definicja 1 (Relacja hiponimiczna) *Pojęcie, które jest reprezentowane przez wpis leksykalny L_0, jest hiponimem pojęcia reprezentowanego przez wpis leksykalny L_1, jeśli* native speaker *danego języka akceptuje jako prawdziwe zdania skonstruowane ze wzorca L_0 jest (rodzajem) L_1.*

Postać ogólna metody Hearsta polega na tym, że WordNet zostaje przeszukany w celu odnalezienia pojęć, które są połączone relacją hiperonimiczną. Następnie, poszukuje się w tekście wystąpień obu pojęć i szuka się wzorców leksykalno-syntaktycznych, przy pomocy których pojęcia te w tekście zostały połączone. Wzorce takie stanowią podstawę do wyszukiwania innych pojęć, które w bazie WordNet nie występują, a na zasadzie podobieństwa wzorca, bardzo prawdopodobne, że połączone są relacją hiponimiczną.

Metoda Hearsta dąży więc do ekstrakcji wzorców takich jak:

$$NP_0 \; such \; as \; NP_1\{, NP_2 \ldots, \; (and|or) \; NP_i\}, \; gdzie \; i \geqslant 1, \tag{2.14}$$

które implikują, że:

$$\forall_{i \geqslant 0} NP_i \; hyponym(NP_i, NP_0).$$

Niestety metoda ta nie wykazała wysokiej skuteczności ekstrakcji relacji, dlatego wykorzystana została tylko jej szczególna wersja. Polegała ona

Klasyfikacja wzorców leksykalno-syntaktycznych

Referencje:	Finkelstein-Landau i Morin (1999)
Cel:	ekstrakcja wzorców leksykalno-syntaktycznych
Środek:	klasyfikacja wzorców
Wykorzystuje:	Hearst (1992, 1998); Smadja (1993); Daille (1996), system NLP dostępny w narzędziu Promethee
Rozszerzone w:	Xu i in. (2002)
Warstwa:	relacje taksonomiczne i nietaksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	Promethee
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	ręczna: precyzja i zwrot

na ręcznym odkryciu najczęściej występujących w tekście reguł, a dopiero następnie wyszukaniu tych reguł w tekście. W ten sposób powstało sześć reguł wykorzystywanych do ekstrakcji relacji hiperonimicznych. Pierwsza z nich została przytoczona wyżej (2.14) jako przykład, pozostałe wyglądają w sposób następujący:

$$such\ NP\ as\ \{NP,\}*\{(or|and)\}\ NP \tag{2.15}$$

$$NP\ \{,NP\}*\{,\}\ or\ other\ NP \tag{2.16}$$

$$NP\ \{,NP\}*\{,\}\ and\ other\ NP \tag{2.17}$$

$$NP\ \{,\}\ including\ \{NP,\}*\{or|and\}\ NP \tag{2.18}$$

$$NP\ \{,\}\ especially\ \{NP,\}*\{or|and\}\ NP \tag{2.19}$$

2.2.23 Klasyfikacja wzorców leksykalno-syntaktycznych

Metoda przedstawiona w Finkelstein-Landau i Morin (1999) wprowadza mechanizm analizy skupień do procesu ekstrakcji wzorców leksykalno-syntaktycznych.

Klasyczne wzorce opisane w Hearst (1992) są nadzorowane. Ich definicja wymaga aktywnego udziału eksperta poprzez analizę dostępnego korpusu. Rzadko kiedy istnieje pewność, że zdefiniowane wzorce dadzą dobre rezultaty oraz że wszystkie istotne wzorce z korpusu udało się zidentyfikować. Lecz przede wszystkim, koszt uważnej analizy korpusu jest bardzo wysoki. Problemem bezpośrednio poruszanym w tej pracy jest fakt, że rzadko kiedy

Połączenie metod ekstrakcji relacji z nadzorem i bez nadzoru

Referencje:	Finkelstein-Landau i Morin (1999)
Cel:	ekstrakcja relacji połączonymi metodami bez nadzoru i z nadzorem leksykalno-syntaktycznych
Środek:	połączenie metod ekstrakcji relacji bez nadzoru (lingwistyczne i statystyczne) i z nadzorem (wzorce Hearsta)
Wykorzystuje:	Hearst (1992, 1998); Smadja (1993); Daille (1996); Charniak i Berland (1999)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne i nietaksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	Promethee
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	ręczna: precyzja i zwrot

udaje się w pierwszym podejściu uzyskać wzorzec ogólny. Przeważnie zidentyfikowany wzorzec jest doprecyzowywany na podstawie nowych dokumentów i nowych kontekstów jego użycia.

Krokiem w celu doprowadzenia do ekstrakcji wzorców ogólnych, jest metoda klasyfikacji wzorców. Dostępny zbiór zidentyfikowanych wzorców (niekoniecznie kompletnych i uogólnionych) poddawany jest analizie. Polega ona na sprawdzeniu dla każdej dowolnej pary wzorców, czy nie posiadają syntaktycznej części wspólnej. Dla przykładu na podstawie wzorców:

NP find in NP such as LIST,
NP such as LIST continue to plague NP

powstaje wzorzec ogólny:

NP such as LIST.

Klasyfikacja przebiega bez nadzoru, dlatego przy większej liczbie analizowanych wzorców doprowadza do znaczącego wzrostu efektywności metod.

2.2.24 Połączenie metod ekstrakcji relacji z nadzorem i bez nadzoru

Druga z zaprezentowanych metod w Finkelstein-Landau i Morin (1999) sprowadza się do rozróżnienia, a następnie złożenia wyników dwóch metod ekstrakcji relacji: metody z nadzorem oraz metod bez nadzoru.

Analiza skupień Caraballo

Referencje:	Caraballo (1999, 2001)
Cel:	uporządkowanie terminów w skupienia
Środek:	analiza skupień, wzorce Hearsta
Wykorzystuje:	Pereira i in. (1993); Riloff i Shepherd (1997); Roark i Charniak (1998); Hearst (1992)
Rozszerzone w:	Cimiano i Staab (2005)
Warstwa:	relacje taksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	ręczna: precyzja

Metody bez nadzoru to głównie podejścia statystyczne oparte na informacji lingwistycznej rozszerzone o identyfikację wieloczłonowych terminów (Daille, 1996; Smadja, 1993). Metody z nadzorem to głównie prace Hearst (1992, 1998) oraz jego następców (np. Charniak i Berland (1999)).

Finkelstein-Landau i Morin (1999) stawiają hipotezę, że obie te grupy metod są komplementarne. Weryfikacji dokonują na podstawie obserwacji wyników obu metod przeprowadzonych w dziedzinie fuzji oraz w dziedzinie przedmiotów produkcji.

Metody z nadzorem, takie jak metoda Hearsta, cechują się dużą precyzją ze względu na specyfikę wzorców, które tworzone są na podstawie empirycznych obserwacji. Niestety z tego samego powodu cechują się niskim zwrotem — wiele potencjalnie relewantnych informacji nie jest odzyskana. Metody bez nadzoru zachowują się dokładnie odwrotnie. Charakteryzują się wysokim zwrotem odpowiedzi. Niestety cierpi na tym precyzja. Ważną cechą metod bez nadzoru jest również możliwość oszacowania siły relacji wyrażonej na przykład przez stosunek występowalności terminu w związku i bez związku z innym terminem.

Na podstawie dwóch analizowanych dziedzin dowiedziono hipotezy, że metody bez nadzoru oraz z nadzorem są komplementarne. Ich jednoczesne użycie zwiększa zarówno precyzję jak i zwrot uzyskanych wyników.

2.2.25 Analiza skupień w ekstrakcji relacji taksonomicznych

Metoda przedstawiona w Caraballo (1999, 2001) służy do automatycznego klasyfikowania terminów w nazwane zbiory. Klasyfikacja przebiega metodą

analizy skupień, natomiast nazywanie skupień — metodą opartą na wzorcach leksykalno-syntaktycznych Hearst (1992).

Zanim metoda rozpocznie swoje działanie należy w miarę prostym i skutecznym sposobem zebrać kolekcję terminów. W tym celu skorzystano z prac Riloff i Shepherd (1997); Roark i Charniak (1998). Z korpusu dokonano ekstrakcji terminologii wchodzących w skład dwóch relacji:

NP, NP, ...and NP (np. serwery, drukarki i światłowody),
NP, NP (np. serwer IBM Blade, rozwiązanie klasy enterprise).

Dodatkową motywacją dla wykorzystania właśnie tych wzorców jest fakt, że uzyskane w ten sposób terminy są ze sobą powiązane (Riloff i Shepherd, 1997; Roark i Charniak, 1998). W ten sposób można uzyskać kolekcję terminów, która jest bardziej podatna na strukturyzację w taksonomię.

Pierwszym etapem metody jest pogrupowanie zbioru wszystkich zidentyfikowanych na podstawie powyższych wzorców terminów w nienazwane skupienia. W tym celu, wykorzystuje się metodę *bottom-up*, tj. rozpoczyna analizę od liczby skupień równej liczbie terminów, wszystkich jednoelementowych. Pereira i in. (1993) stosują dokładnie odwrotne podejście, tj. *top-down*). Dla każdego terminu konstruowany jest wektor składający się z liczb określających, ile razy wystąpił on z każdym innym terminem w powyższych wzorcach. Dla tak skonstruowanych wektorów można wyliczyć miarę podobieństwa opartą na cosinusie kąta pomiędzy dwoma wektorami zgodnie z równaniem:

$$\cos(v, w) = \frac{v * w}{|v| * |w|}. \tag{2.20}$$

Dla mierzenia odległości pomiędzy skupieniami wykorzystano prostą metodę średniej miary cosinusa pomiędzy wszystkimi kombinacjami elementów obydwu skupień, tj:

$$sim(a, b) = \frac{\sum_{v,w} \cos(v, w)}{size(A) * size(b)}. \tag{2.21}$$

W przypadku złączenia dwóch skupień A i B w jedno skupienie C, miarę podobieństwa skupienia C z dowolnym innym skupieniem i wyliczyć można opierając się na:

$$sim(C, i) = \frac{sim(A, i) * size(A) + sim(B, i) * size(B)}{size(A) + size(B)}. \tag{2.22}$$

Analiza skupień kończy się w momencie, gdy powstanie jedno duże skupienie.

Drugi etap metody polega na nazwaniu nienazwanych skupień. W tym celu na podstawie wzorców leksykalno-syntaktycznych opisanych w Hearst

Wzorce Kietz i in. (2000)

Referencje:	Kietz i in. (2000)
Cel:	ekstrakcja relacji hiperonimicznych
Środek:	zmodyfikowany wzorzec Hearsta
Wykorzystuje:	Hearst (1998)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne
Wykorzystywane ontologie:	GermaNet (Hamp i Feldweg, 1997), słownik Swiss Life
Powiązane narzędzia:	On-To-Knowledge (Fensel i in., 2000)
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja i zwrot

(1992), w szczególności wzorca 3. (równanie 2.16) i 4. (równanie 2.17) (*NP, NP, and other NPs*), dokonywana jest konstrukcja wektora dla każdego jednoelementowego skupienia. Wektor ten składa się z elementu 1, jeśli dany termin został zgodnie ze wzorcem sklasyfikowany jako hipernim lub 0 jeśli nie. Następnie dla każdej gałęzi drzewa konstruowany jest wektor będący sumą wektorów jego dzieci. Dla każdego skupienia można w ten sposób wyznaczyć termin (lub terminy gdy istnieje jednakowa liczba przy różnych elementach wektora), który najczęściej występuje w relacji hiperonimicznej. Nazwa tak wyznaczonego terminu staje się nazwą skupienia.

Niestety, przedstawiona metoda powoduje w praktyce wiele problemów. Metoda tworzy drzewo binarne — bezpośrednio połączyć można tylko dwa skupienia. W bardzo wielu przypadkach powoduje to niepotrzebne wydłużenie drzewa. Na przykład, gdy analizowana jest lista wszystkich krajów, można sobie wyobrazić strukturę ze skupieniami: kraje europejskie, kraje azjatyckie, etc. W praktyce jednak najlepsza jest w tym przypadku struktura dwupoziomowa z pojęciem hipernimicznym *kraj*.

W celu zlikwidowania niektórych problemów związanych z nadmiarowością gałęzi pośrednich, zastosowano kompresję drzewa. Jeśli istnieją problemy w rozwiązaniu hiperonimia dla wewnętrznej gałęzi drzewa, to jest ona usuwana, natomiast wszystkie jej dzieci są dołączane do rodzica usuwanego skupienia.

Przedstawiona metoda jest znaczącym osiągnięciem. Jako pierwsza metoda niewykorzystująca zewnętrznych leksykonów pozwoliła uzyskać wyniki lepsze niż uzyskiwane przy wykorzystaniu ontologii WordNet.

	Miary CSM
Referencje:	Yamamoto i in. (2005)
Cel:	ekstrakcja relacji taksonomicznych
Środek:	miary CSM
Wykorzystuje:	Hearst (1998); Sawaki i in. (1997); Kanzaki i in. (2004)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	zgodność poziomu hierarchii ze słownikiem

2.2.26 Wzorce Swiss Life

Wzorce opisane w Kietz i in. (2000) są częścią środowiska On-To-Knowledge (Fensel i in., 2000). Oparte są na wzorcach Hearsta (Hearst, 1992).

Dla konkretnego słownika firmy ubezpieczeniowej Swiss Life opracowany został wzorzec postaci:

lexicon entry : $(NP_1, NP_2, NP_i, and/or NP_n)$
$forall \quad NP_i, 1 \leqslant i \leqslant n \qquad hypernym(NP_i, lexiconentry)$
Result: hypernym("electronic service", "A.D.T.")

Wzorzec ten umożliwia większą efektywność ekstrakcji relacji hiperonimicznych ze słownika organizacji Swiss Life. Oczywiście metoda ta jest mocno ograniczona z tytułu dostosowania jej do konkretnego źródła.

2.2.27 Miary CSM

W celu ekstrakcji relacji taksonomicznych Yamamoto i in. (2005) zastosowali miary CSM (ang. *complementary similarity measures*) (Yamamoto i in., 2005), znane z rozpoznawania obrazów (Sawaki i in., 1997), wcześniej wykorzystane do ekstrakcji relacji nazwanych (Kanzaki i in., 2004).

Miara CSM jest kombinacją wektorów cech kontekstu występowania analizowanych wyrazów i istnieją jej dwa warianty: CSM stworzona dla rozpoznawania binarnych obrazów (CSM_b) oraz dla obrazów w odcieniach szarości (CSM_g).

Po zastosowaniu prostego filtru lingwistycznego dla każdej pary terminów liczona jest miara podobieństwa CSM. Następnie, dla posortowanych wyników, tworzone jest drzewo hierarchii według zależności pomiędzy terminami (wartość miary CSM, aktualny stan hierarchi oraz występowalność terminów w drzewie).

Ekstrakcja relacji meronimicznych Charniak

Referencje:	Charniak i Berland (1999)
Cel:	ekstrakcja relacji meronimicznych
Środek:	wzorce lingwistyczne (powierzchni)
Wykorzystuje:	Hearst (1992)
Rozszerzone w:	—
Warstwa:	relacje nietaksonomiczne
Wykorzystywane ontologie:	WordNet
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna – grupa ekspertów, oparta na precyzji

Uzyskane wyniki pokazują, że miara CSM_b lepiej ekstrahuje drzewa o niższym poziomie głębokości, natomiast miara CSM_g lepiej sprawdza się przy wyższych poziomach głębokości drzewa.

Metoda została poddana ewaluacji poprzez porównanie ze słownikiem EDR.

2.2.28 Relacje meronimiczne Charniak

Hearst w swoich badaniach nad wzorcami ekstrakcji relacji taksonomicznych (Hearst, 1992) próbował zastosować wzorce ogólne do innych typów relacji nietaksonomicznych. Niestety, jak sam przyznał, nie udało mu się osiągnąć sensownych wyników.

Jedną z udanych prób wykorzystującą pomysły Hearsta jest metoda opisana w Charniak i Berland (1999). Zaprezentowane podejście umożliwia ekstrakcję relacji meronimicznych (*całość-część, part-of*). Na podstawie obserwacji zdań języka angielskiego opracowali oni zbiór reguł z powodzeniem zastosowany do ekstrakcji relacji meronimicznych dla języka angielskiego. Zbiór ten składa się z pięciu wzorców.

A. whole NN[-PL]'s POS part NN[-PL]
... building's basement...
B. part NN[-PL] of PREP {the/a} DET mods [JJ/NN]* whole NN
... basement of a building...
C. part NN in PREP {the/a} DET modes [JJ/NN]* whole NN
... basement in a building...
D. parts NN-PL of PREP wholes NN-PL
... basements of buildings...
E. parts NN-PL in PREP wholes NN-PL
... basements in buildings...

Analiza skupień z wykorzystaniem podpisu tematu

Referencje:	Agirre i in. (2000)
Cel:	analiza skupień znaczeń terminów
Środek:	analiza skupień na podstawie topic signatures
Wykorzystuje:	Lin i Hovy (2000); Hovy i Lin (1999)
Rozszerzone w:	—
Warstwa:	relacje nietaksonomiczne
Wykorzystywane ontologie:	WordNet
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja i zwrot

Najlepsze wyniki uzyskano przy zastosowaniu pierwszych trzech wzorców.

Metoda ekstrakcji wzorców meronimicznych ma jednak dwie podstawowe wady. Po pierwsze, niestety nie daje sobie rady w nadmiarowością dokumentów w korpusie. Jeśli jeden dokument pojawia się dwa razy, to wyniki są zdeformowane. Po drugie, relacja meronimiczna jest relacją bardzo subiektywną. Na przykład odpowiedź na pytanie, czy wiedza jest częścią przedsiębiorstwa czy odwrotnie, zależeć będzie od kontekstu i doświadczeń eksperta. Dlatego autorzy spotkali się z problemami w ewaluacji mechanizmu.

2.2.29 Analiza skupień z wykorzystaniem podpisu tematu

Metoda opisywana w Agirre i in. (2000) oraz w sekcji 2.2.15 ma na celu rozszerzanie listy pojęć terminami występującymi w tekście razem z pojęciem. Metoda została wykorzystana do ekstrakcji pojęć, ale nie tylko. Pojęcia wraz z rozszerzoną listą terminów zostały również wykorzystane na drodze analizy skupień do ekstrakcji relacji nietaksonomicznych.

Analizę skupień można oczywiście zastosować do synsetów w ontologii WordNet. Autorzy jednak zdecydowali się na zmierzenie podobieństwa znaczeń wyrazów do danego synsetu. Zastosowanie binarnego algorytmu hierarchicznego pozwoliło na stwierdzenie, które ze znaczeń analizowanego synsetu można pominąć dla uproszczenia np. procesu wyszukiwania informacji. Niejako na marginesie pokazali jednak, w jaki sposób wykrywać relacje nietaksonomiczne analizowanych znaczeń wyrazów, czyli różnych pojęć.

Wzorce UNO	
Referencje:	Iwanska i in. (2000)
Cel:	ekstrakcja relacji
Środek:	zastosowanie płytkich metod (Hearst)
Wykorzystuje:	Iwanska (2000); Hearst (1998)
Rozszerzone w:	—
Warstwa:	relacje nietaksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	ręczna: precyzja

2.2.30 Wzorce UNO

Iwanska i in. (2000) prezentują metodę, która zawiera połączenie płytkiej oraz głębokiej metody ekstrakcji informacji z tekstu. Płytka metoda polega na zastosowaniu wzorców Hearsta (Hearst, 1998) do analizy niskoprzetworzonego tekstu (bez rozbudowanej informacji lingwistycznej, np. w postaci zależności relacyjnych). Głęboka metoda to reprezentacja powstałych pojęć i relacji w postaci rozbudowanego modelu lingwistycznego UNO (Iwanska, 2000).

Na podstawie metody Hearsta zidentyfikowano zestaw ośmiu powtarzających się w korpusie wzorców:

1. *X and not Y* charakteryzujący powiązane typy.
2. *X if not Y* charakteryzujący stopniowalność typów.
3. *not only X but Y* charakteryzujący stopniowalność typów.
4. *X indeed Y* charakteryzujący stopniowalność typów.
5. *X, which BE Y* charakteryzujący definicję.
6. *X, or Y,* charakteryzujący definicję.
7. *X such as Y,. . . , Z* charakteryzujący definicję.
8. *X rather than Y* charakteryzujący powiązane typy.

Następnie z korpusu składającego się z dokumentów z magazynu *Time* oszacowano liczbę wystąpień każdego ze wzorców oraz liczbę wystąpień zgodnie z klasyfikacją powyżej. W rezultacie za najbardziej obiecujące uznano dwa ostatnie wzorce.

Odkrywanie relacji pojęciowych w TextToOnto	
Referencje:	Maedche i Staab (2000b,a)
Cel:	ekstrakcja relacji nietaksonomicznych
Środek:	heurystyki lingwistyczne
Wykorzystuje:	Srikant i Agrawal (1995); Neumann i in. (1997); Maedche i Staab (2000c)
Rozszerzone w:	—
Warstwa:	relacje nietaksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	On-To-Knowledge, TextToOnto, (Fensel i in., 2000); (Maedche i Staab, 2000c)
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja i zwrot

Dla wzorców 7. i 8. przeprowadzono dalsze analizy. Polegały one na przetworzeniu tekstu przez parser własnej produkcji, oszacowaniu listy wyrazów, które są charakterystyczne dla każdego wzorca i definiują zakres pojęć. Na przykład pojęcie X ze wzorca 7. przeważnie jest oddzielone od poprzedniego tekstu znakiem interpunkcyjnym.

Ewaluacja obu wzorców opierała się na precyzji. Dla wzorca 7. wynosiła 64% natomiast dla wzorca 8. była dużo gorsza. Ten ostatni występuje w tekście dużo częściej i nie zawsze opisuje relację powiązania.

Rzekoma unikatowość metody polega na tym, że model UNO zyskuje dodatkową metodę gromadzenia danych.

2.2.31 Odkrywanie relacji pojęciowych w TextToOnto

Podejście opisane w Maedche i Staab (2000b,a) opiera się na dwóch zasadniczych filarach.

Po pierwsze jest to system ekstrakcji informacji z tekstów w języku niemieckim — SMES (Saarbrucken Message Extraction System) (Neumann i in., 1997). Wykorzystanie SMES umożliwia zastosowania standardowych mechanizmów przetwarzania tekstu naturalnego, czyli: tokenizera, analizy leksykalnej oraz mechanizmów łączących współwystępowalne terminy. Analiza leksykalna obejmuje w tym przypadku analizę morfologiczną, rozpoznanie bytów nazwanych, ekstrakcję informacji specyficznych dla dziedziny oraz analizę części mowy. SMES zawiera również mechanizmy pozwalające na wskazanie par terminów, które podczas analizy językowej wykazują zależności. SMES zawiera zbiór reguł oraz algorytmy lingwistyczne (nazywane heurystykami),

które umożliwiają już na etapie analizy lingwistycznej rozpoznanie terminów, które ze sobą współwystępują. Niestety, sama analiza lingwistyczna wykazuje słabe wyniki w ekstrakcji relacji zależności.

Metoda proponuje zastosowanie dodatkowych heurystyk, mianowicie:

- Heurystyka NP-PP, która łączy wszystkie wyrażenia przyimkowe z odpowiednimi rzeczownikami.
- Heurystyka zdania, która łączy wszystkie terminy w danych zdaniu ze sobą, jeśli żadna inna reguła nie zadziała.
- Heurystyka tytułu, która łączy z tytułem dokumentu wszystkie w nim występujące terminy.

Heurystyka NP-PP sprawdza się w znacznej liczbie przypadków, zwłaszcza dla testowanych korpusów z dziedziny turystyki. Heurystyka tytułu wynika z układu dokumentów, którymi posługiwano się i jest wysoce specyficzna dla aplikacji.

Maedche i Staab (2000b,a) uzyskali poprawę rezultatów, ale badania ich opierały się na jednej dziedzinie oraz na dokumentach półustrukturyzowanych, co dyskwalifikuje metodę w szerszej aplikacji oraz dokumentach nieustrukturyzowanych.

Na bazie procesów skojarzonych z systemem SMES działa drugi filar opisywanej metody — algorytm uczący. Algorytm oparty został na mechanizmie odkrywania uogólnionych reguł asocjacyjnych (Srikant i Agrawal, 1995). Mechanizm posługuje się miarami wsparcia oraz ufności do wyznaczania relewantnych relacji. Relacja klasyfikowana jest jako relewantna jeśli obie miary przekraczają próg zdefiniowany przez eksperta.

Metoda była ewaluowana na korpusie 2234 dokumentów HTML z użyciem miar precyzji i zwrotu.

2.2.32 Odkrywanie relacji pionowej

Odkrywanie relacji pionowej (wertykalnej) jest bardzo prostą metodą opartą na założeniu, że jeżeli dane są dwa pojęcia i nazwa jednego zawiera się w nazwie drugiego, to między nimi zachodzi relacja taksonomiczna. Na przykład *spółka* i *spółka kapitałowa* to relacja hiponimiczna.

Metoda jest wykorzystywana praktycznie we wszystkich systemach umożliwiających ekstrakcję relacji taksonomicznych, najczęściej jako metoda pomocnicza.

Odkrywanie relacji pionowej

Referencje:	Velardi i in. (2001a)
Cel:	ekstrakcja relacji taksonomicznych wyrazów podobnych syntaktycznie
Środek:	zawieranie się nazw, analiza syntaktyczna
Wykorzystuje:	Morin i Jacquemin (1999); Vossen (2001)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne
Wykorzystywane ontologie:	WordNet
Powiązane narzędzia:	OntoLearn
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: miara F

Klasyfikacja pojęć w ontologii WordNet

Referencje:	Alfonseca i Manandhar (2002c,b)
Cel:	klasyfikacja w oparciu o podejście top-down oraz podobieństwo semantyczne
Środek:	zawieranie się nazw, analiza syntaktyczna
Wykorzystuje:	Fellbaum (1998); Rajman i Bonnet (1992)
Rozszerzone w:	Alfonseca i Manandhar (2002b)
Warstwa:	relacje taksonomiczne
Wykorzystywane ontologie:	WordNet (Fellbaum, 1998)
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja i zwrot

First decision: entity		
synset	**synset Id**	*total*
being, organism	n00002908	0.3207
causal agency	n00004753	0.3121
location	n00018241	0.1383
body of water	n07411542	0.1112
thing (anything)	n03781420	0.0457
thing (object)	n00002254	0.0442
cell	n00004081	0.0087
(15 more)	...	...
Second decision: being		
synset	**synset Id**	*total*
human	n00005145	0.6161
animal	n00010787	0.2790
host	n01015823	0.0243
parasite	n01015154	0.0192
flora	n00011740	0.0169
(34 more)	...	...

Rysunek 2.3: Przykład dwuetapowego algorytmu klasyfikacji pojęć według Alfonseca i Manandhar (2002c)

2.2.33 Klasyfikacja pojęć w ontologii WordNet

Celem metody opisanej w Alfonseca i Manandhar (2002c) jest poprawna klasyfikacja nieznanych pojęć w ramach relacji taksonomicznych występujących w ontologii WordNet.

Metoda stosuje podejście odgórne (ang. *top-down*), czyli analizę rozpoczyna od najwyżej położonego pojęcia w ontologii WordNet, a następnie rekurencyjnie analizuje niższe poziomy drzewa. W każdym punkcie decyzyjnym metoda wylicza podobieństwo semantyczne analizowanego pojęcia w ontologii WordNet oraz nieznanego pojęcia będącego przedmiotem klasyfikacji.

Istotą algorytmu jest liczenie odległości semantycznej. Opiera się ona na prostej zasadzie, że pojęcia występujące w podobnym kontekście, cechują się wysoką korelacją pojęciową (Rajman i Bonnet, 1992).

Dla przykładu pojęcie *orc*, jako nieznane w ontologii WordNet, może zostać sklasyfikowane, zgodnie z dwoma krokami algorytmu przedstawionego na rysunku 2.3, jako hiponim *human*.

W pierwszym kroku największą wartość podobieństwa semantycznego wykazało pojęcie *being*, w przypadku drugiego kroku wybrany został synset *human*.

Klasyfikacja pojęć oraz metoda Hearsta	
Referencje:	Alfonseca i Manandhar (2002b)
Cel:	ekstrakcja relacji taksonomicznych
Środek:	połączenie klasyfikacji pojęć oraz wzorców Hearsta
Wykorzystuje:	Hearst (1998); Alfonseca i Manandhar (2002c)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne
Wykorzystywane ontologie:	WordNet (Fellbaum, 1998)
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna, miary: dokładność odwzorowania nieznanych pojęć w WordNet, procent dobrych wyborów, średnia pozycja, którą zajmował poprawny, synset na każdym poziomie analizy, dokładność uczenia.

2.2.34 Połączenie klasyfikacji pojęć oraz metody Hearsta

Metoda połączenia klasyfikacji pojęć oraz metody Hearsta (Alfonseca i Manandhar, 2002b) wykorzystuje połączenie dwóch metod:

- metody wzorców Hearsta (Hearst, 1998) (por. sekcję 2.2.22 na stronie 61) oraz
- metody klasyfikacji pojęć opracowanej przez tych samych autorów (Alfonseca i Manandhar, 2002c) (por. sekcję 2.2.33 na stronie 75).

Kluczową rolę w przypadku obu metod pełni ontologia WordNet. W metodzie wzorców WordNet jest źródłem par hiperonimów, których wzorców następnie wyszukuje się w tekście. W metodzie klasyfikacji natomiast, WordNet służy do analizy pozycji nieznanego pojęcia w taksonomii.

Połączenie obu metod polega na zastosowaniu klasyfikacji pojęć do momentu, w którym wyznaczane są miary podobieństwa semantycznego kandydata oraz pojęcia i jego hiponimów. Jeżeli pomiędzy pojęciem a którymkolwiek z jego hiponimów zachodzi relacja określona wykrytym wzorcem Hearsta, miara podobieństwa modyfikowana jest o z góry ustaloną wartość. Wartość ta w testach ustalana była na 10 i malała o połowę z każdym kolejnym poziomem drzewa taksonomicznego.

Tworzenie drzewa pojęciowego dziedziny

Referencje:	Missikoff i in. (2002)
Cel:	tworzenie taksonomii pojęć
Środek:	analiza relacji pomiędzy znaczeniami ontologii WordNet
Wykorzystuje:	Velardi i in. (2001a,b)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne
Wykorzystywane ontologie:	WordNet (Fellbaum, 1998), ontologia wstępna
Powiązane narzędzia:	OntoLearn
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja i zwrot

Alfonseca i Manandhar (2002b) uzyskują w ten sposób znacznie lepsze wyniki, ponieważ do klasyfikacji pojęć, która nie zawsze daje dobre wyniki, dołączają analizę wzorców. Natomiast sama analiza wzorców cechuje się wysoką precyzją, ale niższym zwrotem.

2.2.35 Tworzenie drzewa pojęciowego dziedziny

Metoda tworzenia drzewa pojęciowego dziedziny opisana w Missikoff i in. (2002) opiera się na tych samych założeniach, o którym mowa była w sekcji 2.2.21.

W odróżnieniu jednak do referowanej metody, tworzenie drzewa pojęciowego polega na łączeniu pojęć w drzewa taksonomiczne, czyli metoda dotyczy fazy ekstrakcji relacji taksonomicznych. Ekstrakcja polega na wykorzystaniu opracowanych wzorców, do tworzenia których wykorzystywane są:

- relacje hiperonimiczne oraz hiponimiczne,
- relacje pertonimiczne,
- relacje podobieństwa oraz synonimiczne.

Na podstawie wykrytych zależności pomiędzy pojęciami tworzone są niepełne drzewa pojęć, w których krawędzie oznaczają relacje hipertoniczne i hiponimiczne.

Uczenie taksonomii przy użyciu szczegółowości i podobieństwa terminów	
Referencje:	Ryu i Choi (2006)
Cel:	uczenie taksonomii
Środek:	miary szczegółowości i podobieństwa terminów
Wykorzystuje:	Ryu i Choi (2005)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	precyzja, zwrot

2.2.36 Szczegółowość oraz podobieństwo terminów

Metoda przedstawiona przez Ryu i Choi (2006) służy do konstrukcji taksonomii z wykorzystaniem miar szczegółowości oraz podobieństwa terminów (ang. *terms specifity and similarity*).

Autorzy wyszli z założenia, że mając dwa terminy $t_1 \neq t_2$ prawdopodobieństwo tego, że termin t_1 jest generalizacją terminu t_2, jest tym większe, im terminy te są do siebie bardziej podobne (miara podobieństwa) oraz t_2 jest bardziej szczegółowe od t_1 (miara szczegółowości lub specyficzności). Zastosowana miara szczegółowości oparta jest na stopniu zawierania współwystępujących terminów, natomiast miara podobieństwa na statystycznym wyróżnieniu dziedzinowym.

Ewaluacja metody została przeprowadzona dla różnych części mowy. Miary precyzji i zwrotu kształtowały się w przedziałach odpowiednio 66-79% oraz 60-90%.

2.2.37 Wzorce syntaktyczne dla języka niemieckiego

Metoda ekstrakcji relacji z języka niemieckiego (Xu i in., 2002) oparta jest na systemie przetwarzania języka niemieckiego SPPC (Shallow Processing Production Center) (Piskorski i Neumann, 2000) oraz ontologii GermaNet (Hamp i Feldweg, 1997). Metoda jest częścią systemu obejmującego fazy ekstrakcji terminologii, synonimów, pojęć oraz relacji taksonomicznych (Xu i in., 2002).

W odróżnieniu od pracy Hearsta (Hearst, 1998), która wymaga na wstępie podania wzorców syntaktycznych, przedstawiona metoda nie wymaga ich definicji *a priori*. Zamiast tego, korzysta z semantycznych relacji obecnych w ontologii GermaNet. Należą do nich relacje synonimiczne, hiponimiczne

Wzorce syntaktyczne dla j. niemieckiego

Referencje:	Xu i in. (2002)
Cel:	ekstrakcja wzorców oraz relacji taksonomicznych i nietaksonomicznych
Środek:	konstrukcja i zastosowanie wzorców syntaktycznych
Wykorzystuje:	Piskorski i Neumann (2000); Hearst (1998); Finkelstein-Landau i Morin (1999)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne i nietaksonomiczne
Wykorzystywane ontologie:	GermaNet (Hamp i Feldweg, 1997)
Powiązane narzędzia:	SPPC (Piskorski i Neumann, 2000)
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja i zwrot

Ekstrakcja relacji meronimicznych Poesio

Referencje:	Poesio i in. (2002)
Cel:	automatyczne tworzenie leksykonu
Środek:	wzorce syntaktyczne
Wykorzystuje:	Vieira i Poesio (2000); Mikheev i in. (1999); Hearst (1998)
Rozszerzone w:	—
Warstwa:	relacje nietaksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	ręczna: precyzja i zwrot

oraz meronimiczne. Pojęcia powiązane tymi relacjami odnajdywane są w tekście, a następnie identyfikowane są wzorce, w których pojęcia te występują. Po zastosowaniu algorytmu klasyfikacyjnego opisanego w Finkelstein-Landau i Morin (1999) wzorce dzielone są na specyficzne dla dziedziny oraz ogólne.

2.2.38 Relacje meronimiczne Poesio

Metoda opisana w Poesio i in. (2002) jest zbiorem badań prowadzonych m.in. w Vieira i Poesio (2000) w celu automatycznej konstrukcji leksykonów. Motywacją są znane wady ontologii WordNet, przede wszystkim jej nadmierne rozróżnianie znaczeń oraz brak ukontekstowienia.

Rozwiązywanie nawiązań przy użyciu źródeł zewnętrznych	
Referencje:	Markert i in. (2003)
Cel:	rozwiązywanie nawiązań
Środek:	identyfikacja siły wyrażeń nawiązań przy pomocy zapytań do sieci Internet
Wykorzystuje:	Hearst (1998); Cunningham i in. (2002)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne i nietaksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	GATE
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	ręczna: precyzja i zwrot

Ekstrakcja relacji meronimicznych powstała na podstawie prac Hearsta (Hearst, 1998), parsera CASS[3] oraz korpusu BNC (British National Corpus). Na podstawie obserwacji Poesio i in. (2002) zdefiniowali zestaw następujących wzorców do ekstrakcji relacji meronimicznych:

- the NP of NP,
- NP of NP,
- NP's NP,
- NP N.

Metoda nie wykorzystuje ontologii WordNet lub pochodnych. W zamian, podczas ewaluacji wykorzystuje miarę *Mutual Information* dla poprawnej identyfikacji najbardziej powiązanych semantycznie terminów.

2.2.39 Rozwiązywanie nawiązań przy użyciu źródeł zewnętrznych

Większość rozwiązań do ekstrakcji relacji wykorzystuje ręcznie zdefiniowane zasoby leksykalne, najczęściej w postaci ontologii WordNet. Pomijając sam fakt, że WordNet ma swoje wady, nie zawsze możliwe jest jego wykorzystanie. Dla znakomitej większości języków naturalnych zasoby takie jak WordNet nie są dostępne. Język polski jest tego doskonałym przykładem.

Metoda przedstawiona w Markert i in. (2003) wykorzystuje w ekstrakcji relacji rozwiązywanie nawiązań (ang. *anaphora resolution*). Nowatorskim rozwiązaniem jest wykorzystanie dokumentów z sieci Internet dla mierzenia

[3] `http://complingone.georgetown.edu/~linguist/parser.html`

prawdopodobieństwa potencjalnych kandydatów na terminy powiązane relacjami.

W analizowanych przypadkach obie części wyrażenia nawiązania (samo nawiązanie, jak i głowa wyrażenia) nie są *explicite* połączone w analizowanym tekście. Nie można więc przy pomocy prostej informacji lingwistycznej (w postaci anotacji lub zależności relacyjnych) dokonać ekstrakcji relacji. Istnieje natomiast silna semantyczna relacja pomiędzy terminami w rozpatrywanym modelu dziedziny. Oznacza to, że relacja ta jest obecna *explicite* w innych dokumentach z tej samej dziedziny.

Proces ekstrakcji relacji przebiega w czterech fazach:

1. Gromadzenie i przygotowanie danych. Dokumenty z korpusu są kolejno analizowane pod kątem ekstrakcji nawiązań. Identyfikowane są potencjalne wyrażenia nawiązania (nawiązanie oraz poprzednik). Identyfikowane są byty nazwane. Dokonywana jest lematyzacja. Dla celów eksperymentów wykorzystano system GATE (Cunningham i in., 2002).

2. Wybór relacji leksykalnej. Na podstawie typu nawiązania wybierana jest relacja leksykalna, która najczęściej występuje z danym typem.

3. Wybór wzorca leksykalno-syntaktycznego, który odpowiada analizowanej relacji leksykalnej.

4. Ekstrakcja potencjalnych wyrażeń nawiązania. Dla każdego rozpoznanego nawiązania identyfikowani są potencjalni poprzednicy. Wynikiem tej fazy jest zbiór potencjalnych wyrażeń nawiązania.

5. Mierzenie relewancji kandydatów z wykorzystaniem sieci Internet. Dla każdego elementu zidentyfikowanego zbioru wyrażeń odniesienia konstruowane jest zapytanie z wykorzystaniem Google API. Ze zbioru kandydatów odnoszących się do jednego nawiązania wybierany jest ten, który posiada najwięcej trafień.

2.2.40 Ekstrakcja relacji z pytań

Prosta, lecz zadziwiniająco efektywna metoda przedstawiona została w Sundblad (2003). Autor podszedł do problemu małej precyzji uzyskiwanych rezultatów poprzez analizę źródła – korpusu, z którego proces uczenia ontologii następuje. Zamiast dowolnego korpusu dokumentów, który przeważnie charakteryzuje się niskim stopniem strukturyzacji, zastosowano korpus składający się wyłącznie z pytań. Pytania charakteryzują się dość wysokim stopniem strukturalizacji.

Ekstrakcja relacji z pytań

Referencje:	Sundblad (2003)
Cel:	ekstrakcja relacji z użyciem korpusu pytań
Środek:	wzorce syntaktyczno-leksykalne dla pytań
Wykorzystuje:	Hearst (1998)
Rozszerzone w:	—
Warstwa:	relacje taksonomiczne i nietaksonomiczne
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	ręczna: precyzja i zwrot

Dla korpusu zostały wyznaczone wzorce leksykalno-syntaktyczne, zgodnie z metodą Hearsta (Hearst, 1998). W odróżnieniu jednak od pracy Hearsta, wzorce przedstawione przez Sundblad (2003) nie są ogólnymi wzorcami, tj. na ich podstawie nie można dokonać ekstrakcji dowolnego terminu spełniającego daną relację. Opracowano zestaw wzorców, które można zastosować do ekstrakcji relacji taksonomicznych następujących klas terminów:

- Osób:

  Kim jest/był X?
  Z czego X jest najbardziej sławny?

- Lokacji:

  Gdzie X jest położony?
  Gdzie znajduje się X?

- Skrótów i akronimów:

  Co znaczy X?
  Od czego skrótem/akronimem jest X?

Jedynym ogólnym zidentyfikowanym wzorcem jest:

Jakim typem/rodzajem Y jest/był X?

Metoda ekstrakcji relacji hiponimicznych daje bardzo dobre wyniki. Precyzja jest równa lub bardzo bliska 100%.

Metoda została przetestowana również do jednego typu relacji nietaksonomicznej, tj. relacji meronimicznej. Opracowane zostały następujące wzorce:

Tworzenie modelu pojęciowego z dokumentacji

Referencje:	Aussenac-Gilles i in. (2000a,b)
Cel:	definicja ram i potrzebnych zasobów w ramach uczenia ontologii z tekstu
Środek:	doświadczenia z modelowania dziedziny
Wykorzystuje:	—
Rozszerzone w:	—
Warstwa:	wszystkie
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	brak

What is the X of Y? (Jakie jest Y X?)
What is the X for Y? (Jakie jest X dla Y?)
What is X's Y (Jakie jest X Y?)
How many X are in/on Y? (Ile jest X na/w Y?)

Niestety, ekstrakcja relacji meronimicznych nie dała już tak doskonałych wyników jak ekstrakcja relacji hiponimicznych. Nie mniej jednak, zarówno precyzja jak i zwrot wynosiły 83,2%.

2.2.41 Tworzenie modelu pojęciowego z dokumentacji

Metoda przedstawiona w Aussenac-Gilles i in. (2000a) definiuje proces konstrukcji modeli pojęciowych wykorzystywanych w modelowaniu dziedziny dla potrzeb projektowania systemów informatycznych.

W myśl autorów, dla celów stworzenia modelu pojęciowego (w teorii projektowania systemów — modelu dziedzinowego) niezbędne jest zagwarantowanie następujących zasobów:

1. Zestawu wymagań dla modelowanej aplikacji.
2. Dokumentacji technicznej.
3. Części istniejących modeli, które mogą być wykorzystane.
4. Wiedzy eksperckiej.
5. Narzędzi przetwarzania tekstu naturalnego.

Każdy z zasobów jest wykorzystywany inaczej, w zależności od cech obiektywnych dziedziny, jak i subiektywnych analityka. Wynikiem procesu jest model pojęciowy. Sam proces rozwiązywania zasobów do modelu pojęciowego podzielony jest na cztery główne fazy:

Ekstrakcja nazwanych relacji binarnych Snowball

Referencje:	Agichtein i Gravano (2000); Agichtein i in. (2001)
Cel:	ekstrakcja nazwanych, binarnych relacji
Środek:	analiza płytkiej struktury tekstu
Wykorzystuje:	Brin (1999)
Rozszerzone w:	—
Warstwa:	relacje
Wykorzystywane ontologie:	—
Powiązane narzędzia:	Snowball (Agichtein i in., 2001)
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	precyzja, zwrot

1. Tworzenie korpusu. Ekspert musi dokonać analizy dostępnych dokumentów i wybrać te, które w najlepszy sposób opisują dziedzinę. Główną rolę pełnią w tej fazie dokumenty oraz wiedza ekspercka.

2. Analiza lingwistyczna. Ekspert dokonuje wyboru najbardziej odpowiednich narzędzi oraz metod przetwarzania tekstu. Następnie dokumenty są przetwarzane w celu ekstrakcji podstawowych pojęć z dziedziny.

3. Normalizacja. Po prostej analizie lingwistycznej następuje faza eksploracji korpusu w celu dodefiniowania pojęć, nazwania ich oraz ekstrakcji powiązań pomiędzy nimi. Na tym etapie następuje ekstrakcja relacji z tekstu. Faza analizy lingwistycznej oraz normalizacji mogą cyklicznie następować po sobie, aż do satysfakcjonującego wyniku.

4. Formalizacja. Etap odpowiedzialny za budowę ontologii w postaci sformalizowanej, np. w postaci pliku. Faza obejmuje również walidację wyniku.

Zaprezentowana metoda nie podaje szczegółowego opisu mechanizmów w poszczególnych fazach. System został przetestowany na dokumentach w języku francuskim.

2.2.42 Snowball

Snowball to nadzorowana metoda i system przeznaczony do ekstrakcji binarnych relacji nietaksonomicznych. System potrzebuje próby zdefiniowanych przez eksperta wzorców leksykalnych, na podstawie których dokonuje ekstrakcji. Proces ekstrakcji wspierany jest przez ciągłą ewaluację reguł przy użyciu miar wsparcia.

Ekstrakcja nazwanych relacji binarnych LEILA

Referencje:	Suchanek i in. (2006b)
Cel:	ekstrakcja nazwanych, binarnych relacji
Środek:	analiza głębokiej struktury tekstu
Wykorzystuje:	Suchanek i in. (2006a)
Rozszerzone w:	Kasneci i in. (2007); Suchanek i in. (2007)
Warstwa:	relacje
Wykorzystywane ontologie:	—
Powiązane narzędzia:	LEILA
Wykorzystanie w j. polskim:	NIE
Ewaluacja:	precyzja, zwrot

Celem autorów (Agichtein i Gravano, 2000) była klasyczna informacja ekstrakcji tj. uzupełnienie tabel, w których znajduje się dokładna informacja, jakiej klasy bytów należy szukać.

Metoda jest rozwinięciem podejść opartych na wzorcach leksykalnych Hearsta (Hearst, 1998). Wartością dodaną jest proces ciągłej ewaluacji reguł.

Ewaluacja metody przeprowadzona została na relacji pomiędzy organizacją i miejscem siedziby (ang. *Organization-Location*). Wersja metody ze znakami interpunkcyjnymi na testowanej próbie osiąga rezultaty lepsze niż metoda bazowa.

2.2.43 LEILA

LEILA to system służący do nadzorowanej ekstrakcji nazwanych binarnych relacji nietaksonomicznych. System potrafi dokonać ekstrakcji danej relacji, którą należy na wstępie zdefiniować i scharakteryzować, np. *instanceOf* lub *birthdate*. Ekstrakcja jest binarna tzn. pozyskiwany jest zarówno jej podmiot, jak i przedmiot, np. dla relacji *birthdate* będzie to *osoba* oraz wartość daty urodzenia.

System wykorzystuje metodę opartą na głębokiej analizie lingwistycznej, która bada drzewo zależności w zdaniu. Algorytm metody składa się z następujących etapów:

1. W fazie odkrywania analizowane są drzewa zależności w zdaniach w korpusie. Wyszukane zostają wcześniej przygotowane pary, które są tzw. *pozytywnymi przykładami*. W miejscu drzewa, w którym zostaną wyszukane, wstawiane są specjalne znaczniki tworzące wzorce. Są one następnie wykorzystane do wyszukania kolejnych przykładów, niekoniecznie zgodnych z poprawnymi wskazaniami (tzw. *negatywne przykłady*).

Ekstrakcja pamięciowa tzw. *rote extractors*

Referencje:	Alfonseca i in. (2006b,a)
Cel:	ekstrakcja nazwanych relacji
Środek:	ekstrakcja pamięciowa tzw. *rote extractors*
Wykorzystuje:	Mann i Yarowsky (2005); Brin (1999); Ravichandran i Hovy (2001)
Rozszerzone w:	Ruiz-Casado i in. (2007)
Warstwa:	relacje
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	precyzja

2. W fazie trenowania obliczane są przy pomocy miar statystycznych klasyfikatory dla relacji. W przykładowej implementacji zastosowano metodę klasyfikatora k-NN (k-tego najbliższego sąsiada) oraz SVM (Suchanek i in., 2006a).

3. W fazie testowej analizowane są po raz kolejny wszystkie drzewa w korpusie. Dla każdego znacznika produkuje wszystkie możliwe pary podmiotu i przedmiotu relacji. Jeżeli klasyfikator daje wynik pozytywny, nowa para jest dołączana do wyników metody.

Ewaluacja metody została przeprowadzona na 4 różnych korpusach, tj. Wikicomposers, Wikigeography, Wikigeneral, Googlecomposers oraz 3 zdefiniowanych relacjach, tj. *birthdate*, *synonymy* i *instanceOf*. Uzyskano wyniki miar precyzji oraz zwrotu w granicach odpowiednio 26-80% oraz 15-70%.

Metoda została zastosowana w wyszukiwarce opartej na semantyce słów kluczowych — NAGA (Kasneci i in., 2007). NAGA umożliwia wyszukiwanie oparte na pojęciach i relacjach pomiędzy nimi.

Wyniki przeprowadzonych eksperymentów z wykorzystaniem systemu doprowadziły do powstania pełnowartościowej ontologii YAGO charakteryzującej się dużą precyzją zawartych w niej informacji (Suchanek i in., 2007).

2.2.44 Ekstrakcja pamięciowa

Ekstrakcja pamięciowa (tzw. *rote extractors*) szacuje prawdopodobieństwo relacji $r(p, q)$ przy danym kontekście zdaniowym $A_1pA_2qA_3$ (Alfonseca i in., 2006b,a; Mann i Yarowsky, 2005; Brin, 1999; Ravichandran i Hovy, 2001). Wartość szacowana jest na podstawie korpusu testowego jako częstość wystąpienia dwóch elementów $r(x, y)$ w kontekście $A_1xA_2yA_3$ podzielone przez

częstość wystąpienia x w kontekście każdego innego wyrazu. Wtedy x jest nazywane *hakiem*, a y *celem* (zgodnie z Ravichandran i Hovy (2001)).

W porównaniu do Mann i Yarowsky (2005); Ravichandran i Hovy (2001) metoda przedstawiona w Alfonseca i in. (2006b,a) dodaje następujące udoskonalenia:

- Wyszukiwanie następuje w kolekcji składającej się zarówno z korpusu utworzonego na podstawie haku, jaki i na podstawie celu.
- Testowanie uzyskanych wzorców następuje niezależnie od konkretnej reguły, tzn. testowanie danej reguły przebiega na regułach odnoszących się również do innych relacji.
- Zastosowanie dodatkowego mechanizmu weryfikacji poprawności uzyskanej pary przy pomocy zapytania ogólnego.

Uruchomienie metody wymaga znaczącego wysiłku polegającego na opracowaniu zestawu startowego reguł. Składa się on m.in. z: nazwy relacji, listy pozytywnych przykładów, kwantyfikacji relacji.

Algorytm metody polega na:

1. Pozyskaniu korpusu dla haka.
2. Pozyskaniu korpusu dla celu.
3. Dla każdej relacji, dla każdego wzorca uzyskanego podczas trenowania modelu:
   (a) sprawdzenie, czy znajduje się na liście przykładów,
   (b) sprawdzenie, czy znajduje się na liście przykładów innej relacji,
   (c) sprawdzenie, czy zgadzają się pozostałe cechy w zestawie startowym,
   (d) wygenerowanie zapytania do Google API składającego się z pozyskanych informacji,
   (e) obliczenie miary prawdopodobieństwa.

Ewaluacja metody została przeprowadzona na wewnętrznym korpusie i narzędziach anotacyjnych. Przedstawione wyniki precyzji były bardzo zróżnicowane i kształtowały się w przedziale 3-100%. Autorzy nie podali wartości miar zwrotu.

Przedstawiona metoda została wykorzystana do wzbogacenia *Simple English Wikipedia* oraz *WordNet 1.7* (Ruiz-Casado i in., 2007). Uzyskane wyniki pozwoliły na dodanie nowych relacji z precyzją 60-70%.

Metody jądra	
Referencje:	Zhao i Grishman (2005)
Cel:	ekstrakcja relacji
Środek:	metody jądra
Wykorzystuje:	GuoDong i in. (2005); Zhou i Zhang (2007)
Rozszerzone w:	Zhang i in. (2006a,b); Yang i in. (2006)
Warstwa:	relacje
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	wykrywanie relacji ACE (Doddington i in., 2004), precyzja i zwrot

2.2.45 Metody jądra

Wiele metod ekstrakcji opartych na uczeniu maszynowym opiera się na liczeniu klasycznego iloczynu kartezjańskiego pomiędzy wektorami, które reprezentują indywidualne obiekty ekstrakcji. Metody oparte na wykorzystaniu tzw. jądra (ang. *kernel methods*) polegają na zastosowaniu zamiast iloczynu kartezjańskiego specyficznej funkcji, która jest matematycznie bardziej złożona niż iloczyn każdego z elementów zbioru.

Metoda przedstawiona w Zhao i Grishman (2005) jest nadzorowaną metodą ekstrakcji relacji z wykorzystaniem metod jądra.

Punktem wyjścia jest zastosowanie źródeł informacji lingwistycznej na 3. różnych poziomach, tj.:

1. Płytka informacja lingwistyczna, w tym informacja morfosyntaktyczna tokenów.

2. Analiza wyrażeń.

3. Głęboka informacja lingwistyczna, w tym funkcje gramatyczne wyrażeń oraz zależności pomiędzy nimi.

Dobór źródeł przeprowadzono na podstawie analizy dotyczącej przydatności poszczególnych poziomów informacji lingwistycznej do zadania (GuoDong i in., 2005; Zhou i Zhang, 2007). Dla zdefiniowanych źródeł informacji opracowano pięć podstawowych metod jądra, tj.:

- jądro argumentów dla ważenia zależności pomiędzy bytami,

- jądro bigramów dla ważenia zależności pomiędzy dwoma tokenami,

Minimalny nadzór

Referencje:	Bunescu i Mooney (2007)
Cel:	ekstrakcja relacji
Środek:	minimalny nadzór
Wykorzystuje:	Bunescu i Mooney (2005)
Rozszerzone w:	—
Warstwa:	relacje
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	precyzja i zwrot

- jądro powiązania pomiędzy zdaniami,
- jądro zależności pomiędzy wyrażeniami,
- jądro zależności lokalnej dla ważenia bezpośredniego kontekstu poszukiwanego wyrazu.

Podstawowe metody jądra są następnie ważone z wykorzystaniem złożonych metod jądra.

Główną wadą metod jądra jest fakt, że ich kształt jest w pewnym stopniu pochodną definicji zadania. Decydując się na ewaluację metody zgodnie z zadaniem wykrywania relacji ACE (Doddington i in., 2004), kształt funkcji jądra uzależniono od znajdujących się tam bytów. Na przykład postać funkcji jądra argumentów posiada cechy relacji pomiędzy osobą a organizacją (PER-ORG), która jest jedną z kluczowych relacji w zdefiniowanym zadaniu. Ewaluacja w innym zadaniu lub w przypadku ogólnym może charakteryzować się odmiennymi wynikami. W zadaniu ACE uzyskano wyniki na poziomie 60-70%.

2.2.46 Minimalny nadzór

Nadzorowana metoda ekstrakcji nazwanych, binarnych relacji nietaksonomicznych przedstawiona w Bunescu i Mooney (2007) wyróżnia się niewielką liczbą wymaganych wskazań eksperta. Zgodnie z przedstawionymi eksperymentami wystarczy zaledwie kilka wskazań pozytywnych oraz negatywnych dla wybranej relacji, aby system mógł dokonać ekstrakcji z precyzją i zwrotem na poziomie 70-80%.

Dla danej relacji nietaksonomicznej należy zdefiniować pozytywne oraz negatywne przykłady. Na ich podstawie konstruowane są modele prawdopodobieństwa oparte na wektorach zawierających cechy lingwistyczne kontekstu

Integracja relacji

Referencje:	Huang i in. (2007)
Cel:	ekstrakcja relacji
Środek:	miary popularności i unikatowości
Wykorzystuje:	—
Rozszerzone w:	—
Warstwa:	relacje
Wykorzystywane ontologie:	—
Powiązane narzędzia:	—
Wykorzystanie w j. polskim:	TAK
Ewaluacja:	—

wystąpienia wskazanych w przykładach wyrazów. Cechy lingwistyczne odnoszą się wyłącznie do płytkiej analizy tekstu, głównie występowalności form podstawowych wyrazów.

Modele tworzone są z wykorzystaniem funkcji jądra zdefiniowanych w celu klasyfikacji relacji. Dotyczą one głównie takich cech jak występowanie oraz kolejność wyrazów w analizowanym tekście.

Na podstawie stworzonych modeli system potrafi dokonać klasyfikacji, czy wskazane dwa byty powiązane są zdefiniowaną relacją czy relacja taka nie zachodzi.

Podstawową wadą metody jest konieczność definicji relacji nazwanej oraz oderwanie od ekstrakcji pojęć. Na wstępie należy wskazać, pomiędzy którymi pojęciami model ma dokonać klasyfikacji. Niewątpliwą zaletą metody jest minimalny nadzór.

2.2.47 Integracja relacji

W zasadzie wszystkie podejścia do ekstrakcji relacji nietaksonomicznych wymagają od eksperta wskazania, które relacje mają stanowić przedmiot zadania ekstrakcji. Identyfikacja właściwych relacji na potrzeby inżynierii i uczenia ontologii jest dużym problemem nawet dla ekspertów.

Metoda integracji relacji (Huang i in., 2007) wskazuje na problemy w poprawnej identyfikacji relacji jako przedmiotu ekstrakcji. W tym celu proponuje się proste operacje syntaktyczne oraz ujednoznacznianie relacji z użyciem ontologii WordNet. Oba źródła stanowią podstawę dwóch zaproponowanych miar, tj. popularności oraz unikatowości relacji. Popularność odnosi się do liczby wystąpień relacji w dziedzinie, natomiast unikatowość bierze pod uwagę, czy w dziedzinie (tekście) nie znajdują się inne relacji o podobnych znaczeniu. Na podstawie obu miar metoda identyfikuje, które relacje są ważne dla dziedziny.

Prace przedstawione w Huang i in. (2007) należy uznać za początkujące — autorzy nie podają żadnych szczegółów metody, ani ewaluacji.

2.3 Narzędzia

Przedstawione metody w większości przypadków nie są rozpowszechniane z narzędziami umożliwiającymi ewaluację przeprowadzonych eksperymentów. Problem ten jest po części rozwiązany przez niektóre ośrodki naukowe, które tworzą i udostępniają środowiska zawierające określoną grupę omawianych metod.

Niniejsza sekcja przedstawia najbardziej popularne narzędzia wykorzystywane do uczenia ontologii, tj. takie, które są powszechnie dostępne, rozwijane, posiadające dokumentację i wykorzystywane przez autorów metod. Wszystkie z wymienionych narzędzi są programami udostępnianymi dla środowiska naukowego nieodpłatnie i na zasadzie otwartego kodu źródłowego.

2.3.1 OntoLT

OntoLT jest wtyczką do najpopularniejszego środowiska inżynierii ontologii — Protege (Knublauch i in., 2004). Procesy ekstrakcji zostały oparte na formacie anotacji lingwistycznej o strukturze składającej się z trzech elementów (Vintar i in., 2001):

- tekst — tj. zbiór tokenów wraz z informacją morfosyntaktyczną (lemat, część mowy, etc.),
- wyrażenie — zbiór sekwencji tokenów pełniących określoną funkcję w zdaniu,
- człon zdania — funkcje gramatyczne wyrażeń.

Procesy anotacyjne pozostają poza zakresem funkcjonalności narzędzia i są przeprowadzane przy pomocy narzędzia SCHUG (Declerck, 2002).

OntoLT składa się z trzech kluczowych elementów:

Reguły opisują i definiują zastosowane mechanizmy. Stanowią one podstawowy element procesu ekstrakcji i składają się z wyrażeń w języku warunków wstępnych oraz następujących operatorów. Narzędzie zawiera domyślnie dwie reguły, np. pierwsza z nich dokonuje ekstrakcji głowy wyrażenia rzeczownikowego do klasy oraz głowy wyrażenia plus wyrażenie zmieniające do podklasy. Część drzewa dla tej reguły przedstawiono na rysunku 2.4.

OntoLT	
Referencje:	Sintek i in. (2004); Buitelaar i in. (2004b,a); Buitelaar i Sintek (2004)
Wykorzystuje:	Knublauch i in. (2004)
Rozszerzone w:	—
Warstwa:	terminologia, pojęcia, relacje
Powiązane narzędzia:	SCHUG (Declerck, 2002)

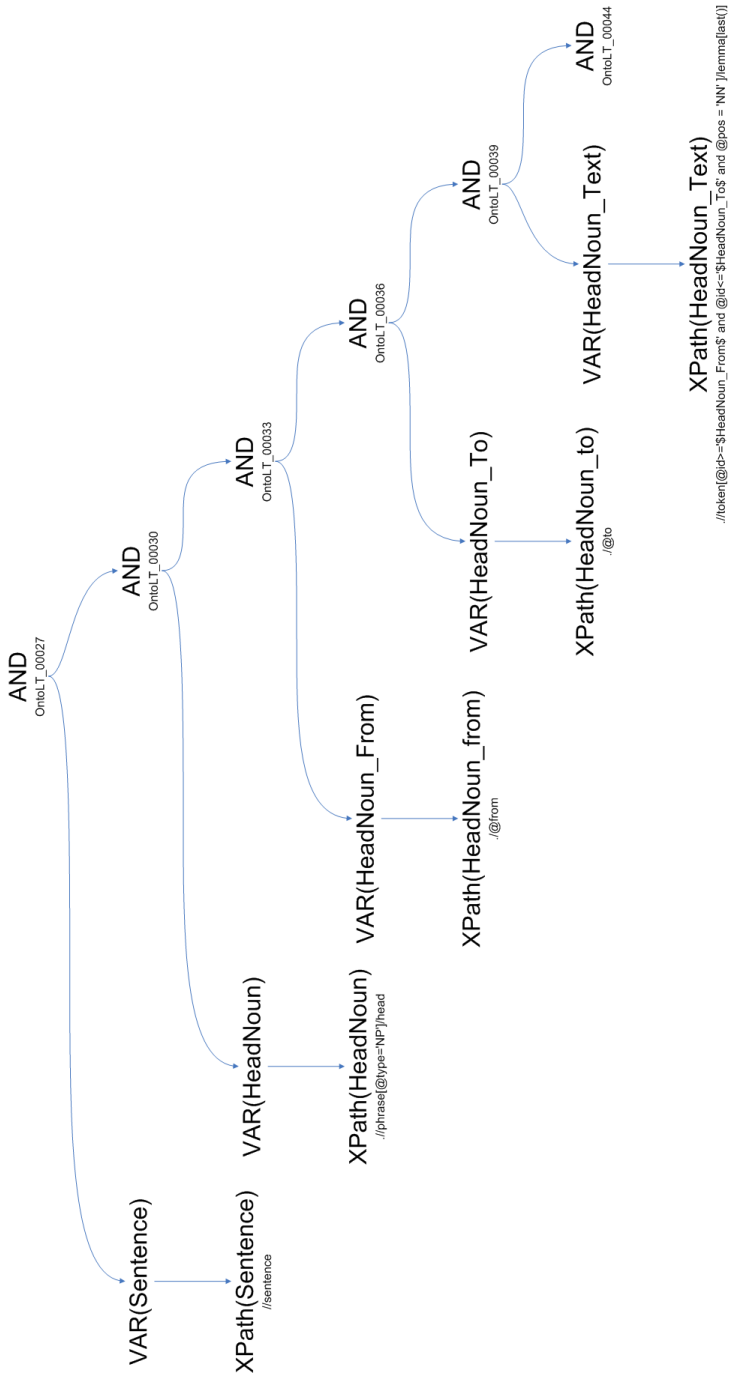
Język warunków wstępnych tzw. *precondition language* jest językiem definiowania reguł i umożliwia jednoznaczne wskazanie elementów lingwistycznych. Wykorzystuje język XPath, co przy założeniu reprezentacji anotacji lingwistycznej w XML umożliwia wygodne i efektywne wyszukiwanie elementów drzewa. Dostępne rozszerzenia elementów języka warunków wstępnych to predykaty oraz funkcje. Do najważniejszych predykatów należą: *containsPath*, *hasValue*, *pred*, *subject* oraz *object*. Predykaty umożliwiają odwołanie się do konkretnych elementów lingwistycznych, np. wyrażenie *containsPath(sentence)* umożliwia odwołanie do zdania.

Operatory umożliwiają dodawanie elementów ontologii. Dostępne operatory to: *CreateCls* (utworzenie nowej klasy), *CreateInstance* (utworzenie instancji), AddSlot (utworzenie właściwości), FillSlot (ustawienie nowej wartości dla właściwości).

OntoLT powstał w roku 2004 w wersji 1.0. W roku 2006 pojawiła się wersja 2.0. Oprócz mechanizmów umożliwiających regułowe, nadzorowane uczenie umożliwia zastosowanie prostych miar statystycznych opartych na porównaniu częstości wystąpienia w używanym korpusie do częstości wystąpienia w korpusie ogólnym.

Narzędzie osiągnęło niepodważalny sukces mierzony liczbą odniesień w innych artykułach. Nie jest ono jednak wolne od wad, do których zaliczyć należy:

- nadzorowaną, opartą na regułach ekstrakcję,
- model reprezentacji reguł, który jest skomplikowany. Do konstrukcji nawet prostej reguły należy opracować obszerne drzewo,
- konstrukcja operatorów — dopuszczalne operatory charakteryzują się małą ekspresywnością w stosunku do powszechnie stosowanych bibliotek inżynierii ontologii, np. Jena (McBride, 2002).



Rysunek 2.4: Część przykładowej reguły (16 z 36 węzłów) dla wyrażenia rzeczownikowego zgodnie z modelem reguł dla OntoLT. Podział drzewa zastosowano w węźle *OntoLT*_00044. Podtytuły węzłów zawierają względne ścieżki zdefiniowane w języku XPath, które pozyskują kolejne elementy wyrażenia

OntoLearn	
Referencje:	Missikoff i in. (2002); Navigli i Velardi (2004); Navigli i in. (2004); Navigli i Velardi (2005); Navigli (2006c,a)
Wykorzystuje:	Velardi i in. (2001b,a)
Rozszerzone w:	—
Warstwa:	terminologia, synonimy, pojęcia, relacje taksonomiczne
Powiązane narzędzia:	—

2.3.2 OntoLearn

OntoLearn to zbiór narzędzi udostępnianych przez *Linguistic Computing Laboratory* przy Uniwersytecie w Rzymie[4]. Obecnie udostępniane narzędzia to[5]:

- *Structural Semantic Interconnections* opisane w sekcji 2.2.21 (Missikoff i in., 2002; Navigli i Velardi, 2005),
- *Terms Extractor* (Velardi i in., 2001b),
- *Taxonomy Validator* — narzędzie do walidacji taksonomii,
- *Valido* — narzędzie do interaktywnego wspomagania ekstrakcji w fazie analizy poprawnego znaczenia terminów (problem niejednoznaczności) (Navigli, 2006c,a).

Narzędzia te udostępniają niektóre metody przedstawione w przeglądzie metod — zwłaszcza opisaną w sekcji 2.2.21 semantyczną interpretację złożonych terminów oraz niektóre proste miary statystyczne. Relewantne dla pracy są dwa pierwsze narzędzia. *Terms Extractor* w celu ekstrakcji terminologii wykorzystuje prostą statystyczną analizę. *SSI* jest narzędziem rozpoznawania poprawnego znaczenia terminów, czyli służy do ekstrakcji synonimów oraz pojęć. Metoda wykorzystana w *SSI* doczekała się interaktywnego interfejsu *Valido* (Navigli, 2006c,a).

Test narzędzi z grupy OntoLearn wykazał, że autorzy skupili się na ekstrakcji synonimów oraz pojęć wprowadzając na przestrzeni lat 2001-2006 nowatorskie metody rozpoznawania poprawnego znaczenia terminów (problem niejednoznaczności terminów).

[4] `http://lcl.di.uniroma1.it/`
[5] Stan na dzień 29 grudnia 2007

TextToOnto	
Referencje:	Maedche i Staab (2000c)
Wykorzystuje:	—
Rozszerzone w:	Cimiano i Völker (2005); Kietz i in. (2000); Maedche i Staab (2000b,a)
Warstwa:	ekstrakcja terminologii
Powiązane narzędzia:	On-To-Knowledge (Fensel i in., 2000), TextToOnto (Maedche i Staab, 2000c,b,a)

2.3.3 TextToOnto

Metody zaproponowane wraz ze środowiskiem TextToOnto (Maedche i Staab, 2000c,b) były w kolejnych pracach tych samych autorów często wykorzystywane. Analiza wskazuje jednak na bardzo lakoniczne opisy samych metod. Artykuły dotyczące środowiska TextToOnto opisują jedynie tzw. *komponent uczenia i odkrywania*, który "zawiera różne metody ekstrakcji terminów" (Maedche i Staab, 2000c). Jedyną wskazówką jest niejasne odniesienie do miar współwystępowalności.

Analiza samego środowiska po zainstalowaniu wykazuje możliwość ekstrakcji terminów poprzez znane metody statystyczne, w tym miarę TFIDF.

Oprócz ekstrakcji terminologii z tekstu narzędzie TextToOnto umożliwia również ekstrakcję terminologii ze słowników dziedzinowych oraz ekstrakcję relacji nietaksonomicznych. Szczegółowy opis tych funkcjonalności przedstawiono odpowiednio w sekcjach 2.2.12 oraz 2.2.31.

2.3.4 Text2Onto

Text2Onto (Cimiano i Völker, 2005) stanowi kompleksowe środowisko służace do uczenia ontologii. Jest następcą narzędzia TextToOnto (Maedche i Staab, 2000c; Kietz i in., 2000; Maedche i Staab, 2000b,a), którego rozwoju zaniechano. W porównaniu z poprzednikiem wprowadzono kilka fundamentalnych zmian.

POM. Zmieniono model reprezentacji obiektów ontologii na tzw. *Probabilistic Ontology Model (POM)*. POM jest modelem wewnętrznym i nie jest bezpośrednio powiązany z żadnym powszechnym modelem formalnym ontologii. Powoduje to niezależność od konkretnego modelu, jednocześnie umożliwiając szereg dodatkowych funkcjonalności. Najważniejszym rozszerzeniem jest znormalizowana w przedziale $[0, 1]$ miara prawdopodobieństwa danej relacji. POM umożliwia zatem reprezentację rozmytą, np. stwierdzenie, że dany wyraz jest terminem z prawdopodobieństwem 50%.

Text2Onto	
Referencje:	Cimiano i Völker (2005)
Wykorzystuje:	Maedche i Staab (2000c); Kietz i in. (2000); Maedche i Staab (2000b,a)
Rozszerzone w:	—
Warstwa:	ekstrakcja terminologii, pojęć, relacji
Powiązane narzędzia:	On-To-Knowledge (Fensel i in., 2000), TextToOnto (Maedche i Staab, 2000c,b,a)

GUI. Narzędzie Text2Onto posiada bogaty interfejs użytkownika, co powoduje, że użytkownik ma istotny wpływ na proces ekstrakcji. Użytkownik może np. wskazać poprawność działania metod.

Zarządzanie zmianą. Narzędzie umożliwia przechwytywanie zmian w źródłach danych, tj. w dokumentach źródłowych. Dodanie dokumentu powoduje ponowne uruchomienie procesu uczenia. Zastosowanie POM oraz mechanizmów przyrostowych umożliwia przetworzenie tylko nowego dokumentu.

Zastosowanie POM jest wygodnym rozwiązaniem, lecz komplikuje proces serializacji. Dla każdego wspieranego modelu ontologii należy budować osobne konwertery. Wykorzystanie miar prawdopodobieństwa również nie jest niczym nowym, ponieważ praktycznie wszystkie modele ekstrakcji i tak generują miarę prawdopodobieństwa, która jest wykorzystywana do filtrowania zgodnie z progiem klasyfikacji.

W odróżnieniu od poprzednika, narzędzie Text2Onto dostarcza gotową i dobrze opisaną bibliotekę algorytmów ekstrakcji. Do ekstrakcji terminologii oraz pojęć wykorzystywane są miary częstości występowania tokenów, m.in. TFIDF oraz metoda wartości C/NC (sekcja 2.2.4). Do ekstrakcji relacji wykorzystywane są:

- eksploracja relacji taksonomicznych ontologii z rodziny WordNet,
- wzorce Hearsta (sekcja 2.2.22),
- wzorce przedstawione w Maedche i Staab (2000b,a) (sekcja 2.2.31) sformalizowane przy pomocy reguł JAPE (Cunningham i in., 2000).

Środowisko Text2Onto nie wprowadza żadnych nowych metod ekstrakcji, wykorzystuje raczej uznane metody przedstawione wcześniej. Nacisk autorów został położony na zarządzanie zmianą, a w konsekwencji na ewolucję ontologii (Leenheer i Mens, 2008; Bloehdorn i in., 2006).

2.4 Podsumowanie

Źródłem procesu uczenia ontologii w przeważającej części jest tekst w języku naturalnym. Inne podejścia, takie jak słowniki, źródła półustrukturyzowane lub relacyjne, nie reprezentują wystarczająco dużej części modelowanej w ontologii rzeczywistości. Z tego powodu znakomita część prac opiera się na ekstrakcji tekstu w języku naturalnym. Proces automatycznego tworzenia ontologii opiera się na szeregu powiązanych, następujących po sobie etapów: ekstrakcji terminologii, synonimów, pojęć, relacji taksonomicznych, relacji nietaksonomicznych oraz reguł.

Pierwsze etapy procesu uczenia, zwłaszcza ekstrakcja terminologii, związane są z problemamami z dziedziny analizy języka naturalnego (NLP — Natural Language Processing). Wiele metod jest opartych na analizie lingwistycznej. Specyfika analizowanego problemu natury biznesowej powoduje jednak, że ontologia, jako narzędzie reprezentacji wiedzy w Sieci Semantycznej, ma bardzo specyficzne wymagania. Powodują one, że analiza lingwistyczna nie tylko musi zostać "dostosowana" do nowego sposobu reprezentacji, ale również wykorzystać można inne, dotąd nieobecne metody. Zwłaszcza ekstrakcja pojęć wykorzystuje mechanizmy specyficzne z języków reprezentacji wiedzy, takie jak np. sposób definiowania abstrakcji. Dochodzi tutaj do klasycznych problemów projektowania — czy dany termin jest pojęciem czy instancją; na ile ważny jest dla modelowanej dziedziny, etc.

Systematyka przedstawionych metod została dokonana w postaci graficznej. Pozwala to na większą przejrzystość powiązań pomiędzy metodami, klasyfikacji metod w podobne grupy oraz naszkicowanie chronologicznych zależności i głównych nurtów, zarówno tematycznych, jak i środowiskowych. Zgodnie z przedstawionym procesem uczenia ontologii wyróżnia się sześć warstw. Wiele metod nie jest jednak zaklasyfikowanych tylko i wyłącznie do jednej warstwy. W skrajnych przypadkach metoda odnosi się do wszystkich warstw, np. Aussenac-Gilles i in. (2000b). Prowadzi to do problemów natury klasyfikacyjnej. Wydaje się, że najbardziej czytelnym sposobem prezentacji jest podział na dwie grupy. Pierwszą z nich stanowią metody ekstrakcji terminologii, synonimów oraz pojęć. W procesie analizy metod często trudno jest dokonać jednoznacznej klasyfikacji do jednej z tych grup. Klasyfikacja jest więc raczej subiektywna. Ponadto wiele metod dotyczy wszystkich trzech faz. Drugą grupę stanowią metody ekstrakcji relacji taksonomicznych oraz nietaksonomicznych. Ekstrakcja relacji taksonomicznych jest przecież specyficzną metodą ekstrakcji relacji nietaksonomicznych. Dlatego obie te fazy charakteryzują się podobnymi metodami. Ostatnia z faz wymienionych w procesie uczenia ontologii — ekstrakcja reguł jest bardzo rozmyta. Potwierdził to przegląd metod. Próby definicji reguł istnieją, lecz dotyczą w zasadzie wszystkich po-

zostałych faz i są często składową innych metod. Z tego powodu systematyka tej fazy nie obejmuje.

2.4.1 Metody ekstrakcji terminologii, synonimów oraz pojęć

Systematyka metod ekstrakcji terminologii, synonimów oraz pojęć przedstawiona jest na rysunku 2.5.

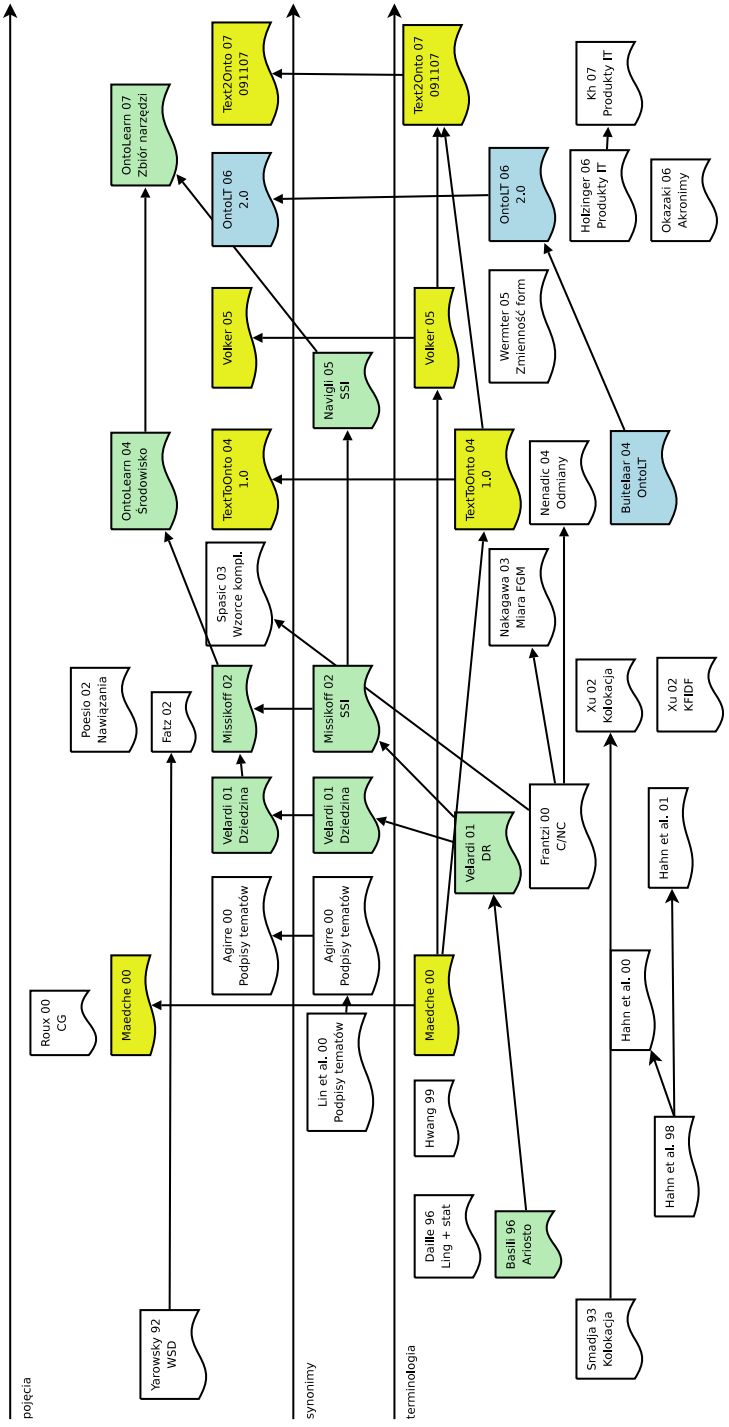
Trudność w oszacowaniu chronologicznie pierwszych metod ekstrakcji terminologii spowodowana jest faktem, że większość wczesnych metod wywodzi się z dziedziny przetwarzana tekstu naturalnego (NLP). Najbardziej popularnymi zagadnieniami lingwistycznymi wykorzystywanymi w procesie uczenia terminologii i pojęć są ujednolicanie znaczenia wyrazów (ang. *word sense disambiguation*) (Yarowsky, 1992), analiza kolokacji (Smadja, 1993) oraz rozwiązywanie nawiązań (ang. *anaphora resolution*) (Poesio i in., 2002). W zasadzie wszystkie prace lingwistyczne można zaklasyfikować jako metody ekstrakcji terminologii. W powyższym zestawieniu jako pierwsza wprowadzona jest praca łącząca analizę lingwistyczną i statystyczną (Daille, 1996). Połączenie obu tych typów analiz stanowi obecnie podstawowy nurt metod ekstrakcji terminologii, synonimów oraz pojęć.

Ekstrakcja terminologii, synonimów oraz pojęć dla ontologii to dziedzina, która zyskała odrębność dopiero ok. roku 2000. W roku tym pojawiła się mało znacząca wówczas praca Roux i in. (2000), która jako pierwsza strukturalizowała wynik analizy lingwistycznej w stosowaną w Sieci Semantycznej notację — grafy pojęciowe Sowy (Sowa, 1984). Metoda ta została rozwinięta przez środowisko uniwersytetu w Karlsruhe.

W większości metod wykorzystywane są zasoby językowe w postaci leksykonów, słowników, tezaurusów oraz ontologii. Do ostatnich lat XX w. najbardziej popularnym zasobem lingwistycznym był tezaurus Roget (Ozsu i Snodgrass, 2001). Wraz z pojawieniem się ontologii WordNet (Fellbaum, 1998) jego znaczenie spadło. Obecnie rodzina ontologii WordNet, czyli jego niemiecka wersja GermaNet (Hamp i Feldweg, 1997) oraz inicjatywa EuroWordNet (Vossen, 1998), spełniają podstawowe i fundamentalne znaczenie dla wielu metod. Inicjatywa EuroWordNet obejmuje swoim zasięgiem osiem języków europejskich: holenderski, hiszpański, włoski, niemiecki, francuski, czeski, estoński oraz nakładkę na brytyjski angielski[6].

Systematyka metod dowodzi, że szczególnie aktywne w kształtowaniu nowych metod były poszczególne ośrodki naukowe.

[6] `http://www.illc.uva.nl/EuroWordNet/finalresults-ewn.html`



Rysunek 2.5: Systematyka metod ekstrakcji terminologii, synonimów i pojęć

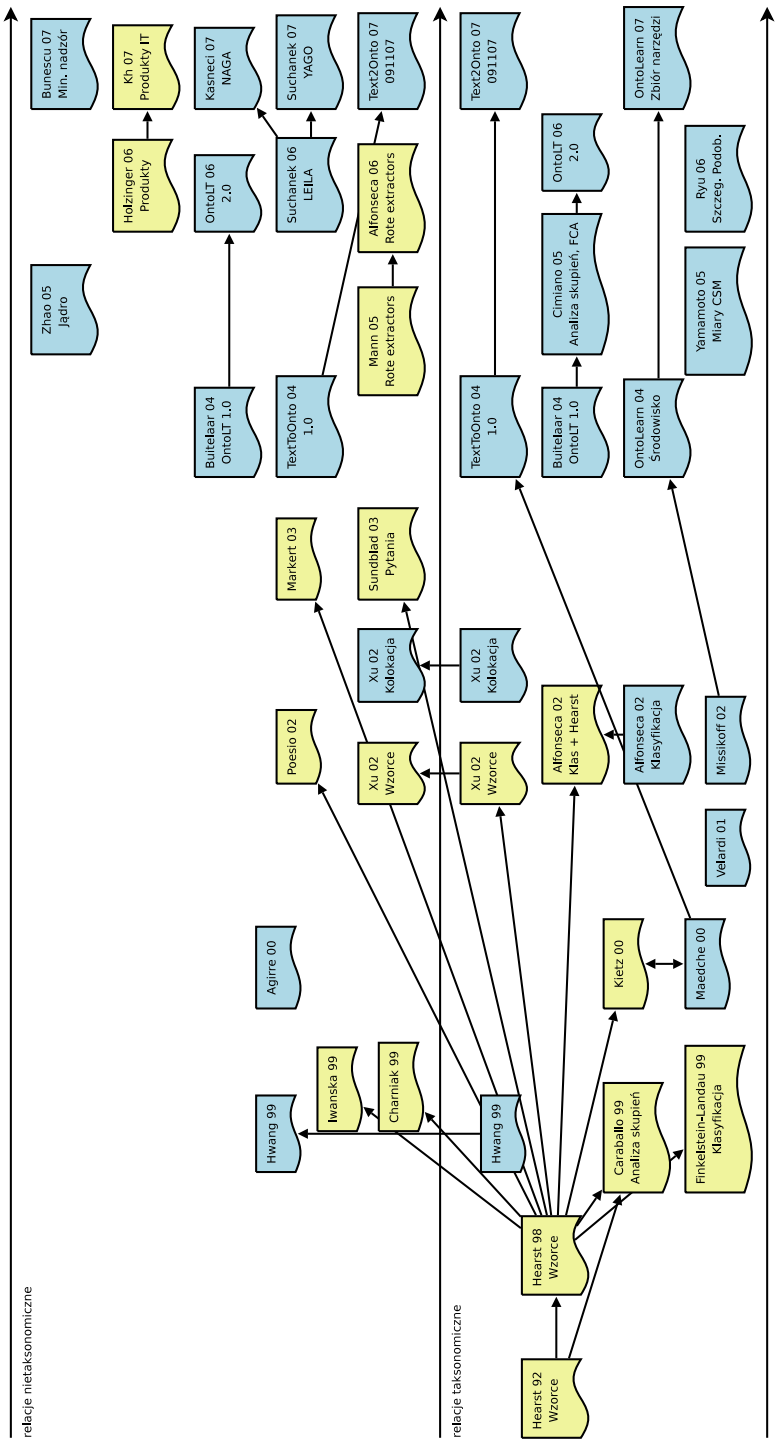
Pierwszym zespołem, który poważnie przyczynił się do opracowania kompleksowych metod ekstrakcji terminologii i pojęć byli pracownicy Uniwersytetu w Karlsruhe pod przewodnictwem prof. Rudi Studera. Zespół ten opracował w roku 2000 zestaw metod opartych na połączeniu analizy lingwistycznej, leksykalnej i statystycznej. Całość udostępniono w narzędziu On-To-Knowledge (Fensel i in., 2000), które było pierwszym publicznie dostępnym i umożliwiającym porównanie metod środowiskiem. Z tej przyczyny był to niewątpliwie przełom. Środowisko to jest o tyle ważne, że ewoluuje do tej pory — w linii bezpośredniej jest to TextToOnto (Maedche i Staab, 2000b,c) oraz Text2Onto (Cimiano i Völker, 2005). To ostatnie jest obecnie jednym z najpopularniejszych środowisk do budowy metod uczenia ontologii. Oczywiście rozwiązanie to w roku 2000 miało bardzo wiele niedociągnięć, na przykład służyło wyłącznie do ekstrakcji pojęć z dobrze ustruktukturalizowego źródła, czyli leksykonu organizacji Swiss Life (sekcja 2.2.12).

Kolejnym ośrodkiem, który niezwykle aktywnie uczestniczył w konstrukcji metod jest laboratorium *Linguistic Computing Laboratory* przy Uniwersytecie w Rzymie kierowane przez Prof. Paolę Velardi[7]. Zespół ten opiera swoje prace na niezwykle elastycznym narzędziu NLP Ariosto+Chaos (Basili i in., 1996). Prace zespołu związane są z definicją i wykorzystaniem relewancji dziedzinowej (Velardi i in., 2001a) oraz analizą znaczeniową terminów wieloczłonowych, a także z tworzeniem i utrzymywaniem popularnego środowiska OntoLearn (Navigli i in., 2004; Navigli i Velardi, 2004; Missikoff i in., 2002).

Pozostałe metody korzystają w procesie ekstrakcji z ogólnie znanych mechanizmów w postaci analizy lingwistycznej i statystycznej.

Przegląd metod ekstrakcji terminologii, synonimów oraz pojęć wskazuje, że można dokonać pewnego uogólnienia przedstawionych metod. Znacząca liczba przedstawionych podejść opiera się na metodach z użyciem następujących miar:

1. TFIDF opartą na podobieństwie dokumentów.
2. TFIDF opartą na podobieństwie korpusów (wyróżnienie korpusu na tle ogólnego słownictwa).
3. KFIDF opartą na wyróżnieniu dziedziny.
4. Metoda wartości C/NC.
5. Metody oparte na modelu n-gram.



Rysunek 2.6: Systematyka metod ekstrakcji relacji

2.4.2 Metody ekstrakcji relacji

Systematyka drugiej grupy metod, tj. metod ekstrakcji relacji taksonomicznych oraz nietaksonomicznych przedstawiona została na rysunku 2.6.

Klasyfikacji metod z tej grupy dokonać jest znacznie łatwiej. Podział pomiędzy relacje taksonomiczne i nietaksonomiczne jest precyzyjny. Warto jednak zaznaczyć, że niektóre metody dokonując ekstrakcji relacji nienazwanych, nie definiują jasnego podziału.

Metody przedstawione na rysunku podzielić można na dwie grupy:

- wywodzące się z analizy lingwistycznej i statystycznej oraz
- oparte na wzorcach syntaktyczno-leksykalnych.

Chronologicznie pierwsze metody do ekstrakcji relacji opracowane zostały w roku 1992 przez Marti Hearsta (Hearst, 1992). Praca ta przedstawiona została jeszcze raz, sześć lat później (Hearst, 1998), w większości w niezmienionej formie i dopiero wtedy doczekała się licznych rozwinięć w postaci różnorodnych wzorców opracowanych na podobnych zasadach. Drugą grupę stanowią metody oparte na analizie lingwistycznej lub statystycznej.

W centralnym punkcie schematu znajdują się dwie metody: Kietz i in. (2000) oraz Maedche i Staab (2000a). Są to prace środowiska związanego z produktem On-To-Knowledge, które jako pierwsze dokonało powiązania obu grup metod, tj. analizy statystyczno-lingwistycznej oraz wzorców Hearsta.

Analiza metod ekstrakcji relacji taksonomicznych wykazuje, że skuteczność tych metod mierzona miarą precyzji i zwrotu jest obecnie na zadawalającym poziomie. Połączenie analizy skupień wraz z analizą lingwistyczną zapoczątkowane przez Kietz i in. (2000); Maedche i Staab (2000a) i rozwinięte przez Cimiano (2006); Cimiano i in. (2005b); Cimiano i Staab (2005) powoduje, że trudno jest poprawić skuteczność obecnie stosowanych metod.

Analiza metod ekstrakcji relacji nietaksonomicznych wykazuje natomiast zdumiewające podobieństwo stosowanych podejść. W przypadku ekstrakcji dowolnych relacji wszystkie przedstawione metody wymagają dużego stopnia nadzoru ze strony użytkownika. W przypadku metod nienadzorowanych ekstrakcja ogranicza się do nazwanych wcześniej relacji.

2.4.3 Wnioski

Przedstawiona w niniejszym rozdziale analiza poszczególnych metod oraz narzędzi, a także ich synteza prowadzą do szeregu istotnych wniosków. Źródłem

[7] `http://lcl.di.uniroma1.it/people.jsp`

rozumowania jest problem biznesowy oraz badawczy określone w rozdziale 1. Przegląd obecnie istniejących metod oraz narzędzi pozwala wyprowadzić następujące wnioski:

1. Brak ogólnego modelu uczenia ontologii z tekstu. Istnieją tylko modele poszczególnych metod lub części, np. Sintek i in. (2004), natomiast całość procesu jest opisana wyłącznie koncepcyjnie, np. Cimiano (2006). Istnieje zatem potrzeba opracowania ogólnego modelu uczenia ontologii z tekstu.

2. Brak kompleksowej metody dla języka polskiego. Nie istnieje żadna metoda dedykowana dla języka polskiego, a większości z istniejących metod nie można przenieść, ponieważ odwołują się do zasobów lingwistycznych specyficznych dla danego języka naturalnego. Większość metod konstruowana jest dla języka angielskiego, istnieją również podejścia dla języków: francuskiego, hiszpańskiego, niemieckiego, japońskiego, chińskiego, włoskiego. Istnieje zatem potrzeba zbudowania metod oraz narzędzia dla uczenia ontologii z języka polskiego.

3. Nieefektywność i nieadekwatność metod ekstrakcji terminologii. Ekstrakcja terminologii jest kluczową częścią całości procesu uczenia ontologii z tekstu. Jednocześnie nie istnieją podejścia do ekstrakcji terminologii dedykowane dla zdefiniowanego problemu. Najbardziej podobna metoda, tj. ekstrakcja produktów IT przedstawiona w Holzinger i in. (2006); Khandelwal (2007) została sklasyfikowana jako problem uczenia ontologii ze źródeł półustrukturyzowanych. Oznacza to potrzebę opracowania metod ekstrakcji terminologii, które cechują się wyższą efektywnością niż obecnie stosowane.

4. Wszystkie metody ekstrakcji relacji wymagają dużego nakładu pracy eksperta. Istnieje jedna praca, która wskazuje ten problem (Bunescu i Mooney, 2007), ale ma ona swoje poważne wady. Należy zatem opracować metodę ekstrakcji relacji, która będzie minimalizować udział eksperta (nadzór ekstrakcji).

Wnioski płynące z przeprowadzonej analizy obecnie stosowanych metod wpłynęły na kształt niniejszej pracy. Pierwsze dwa wnioski są rozpatrywane w rozdziale 3., kolejne wnioski prowadzą do badań przedstawionych w rozdziałach 5. oraz 6.

Rozdział 3

Metamodel

Metamodel opisuje uogólnioną metodę uczenia ontologii z tekstu zawierającą modele ekstrakcji poszczególnych obiektów ontologii oraz funkcje przejścia pomiędzy modelami. Metamodel uczenia ontologii z tekstu M jest:

$$M = \{D, LA, T, S, C, A, TR, NTR\}. \tag{3.1}$$

D jest zbiorem dokumentów $D = \{d_1, \ldots, d_n\}$. Zbiór D definiuje korpus dla całości procesu uczenia ontologii z tekstu, określając między innymi dziedzinę, styl narracji oraz język naturalny. Dobór tego zbioru jest kluczowy dla uzyskanych wyników. Dyskusja nad pożądanymi właściwościami zbioru D oraz ich doborem przedstawiona została w rozdziale 4.

Zbiór D nie jest bezpośrednio wykorzystywany w procesach ekstrakcji z powodów wydajności oraz współoperatywności. Zbiór D jest sprowadzany do postaci anotacji lingwistycznych LA. LA jest zbiorem anotacji lingwistycznych dla zbioru D uzyskanych przy pomocy funkcji Λ:

$$\Lambda : D \times \lambda \times LR \times EC \rightarrow LA, \tag{3.2}$$

gdzie λ oznacza zbiór metod budowania anotacji, LR to zbiór zasobów lingwistycznych wykorzystanych w procesie budowania anotacji, a EC to tzw. klasyfikator wzorcowy.

Klasyfikator wzorcowy EC to anotacja dokonana przez eksperta, która uważana jest za wzorcową. Wykorzystywany jest w procesach ekstrakcji do ewaluacji efektywności metod. Klasyfikator wzorcowy jest parametrem optymalizacyjnym funkcji Λ. Do najczęściej spotykanych właściwości klasyfikatora wzorcowego należą:

- dobór liczebności oraz stosunku zbioru trenującego oraz testującego,
- jakość optymalizacji, np. pod względem zgodności z założeniami (zasady anotacji oraz tolerancja błędu),

- liczba wskazań ekspertów decydująca o klasyfikacji (przy założeniu, że dokument d_i jest anotowany przez więcej niż jednego eksperta).

Zbiór LA jest w teorii otwarty, w praktyce jednak stosować się powinno uznane standardy anotacji lingwistycznej. Dobór zbioru LA jest kluczowy dla porównywalności metod uczenia ontologii z tekstu. Abstrakcyjny model LA składa się z trzech części: anotacji morfosyntaktycznej odnoszącej się do poszczególnych tokenów, analizy wyrażeń oraz funkcji gramatycznych wyrażeń. Obecność oraz wykorzystanie każdej z tych warstw pozycjonuje metody ekstrakcji do odmiennych zastosowań. W przypadku ogólnym im bogatsza informacja lingwistyczna, tym skuteczność metod jest wyższa, lecz powszechność zastosowania niższa. Dyskusja nad pożądanymi cechami oraz doborem formatu anotacji znajduje się w rozdziale 4.

T jest zbiorem terminologii $T = \{t_1, \ldots, t_n\}$ będącej przedmiotem ekstrakcji ze zbioru LA przy pomocy funkcji Γ:

$$\Gamma : LA \times \gamma \times K_{term} \rightarrow 2^T, \tag{3.3}$$

gdzie γ oznacza zbiór metod ekstrakcji terminologii, a K_{term} stanowi zbiór progów klasyfikacji terminologii. Wynikiem funkcji Γ jest zbiór wszystkich podzbiorów zbioru T (2^T).

Zgodnie ze standardami normalizacyjnymi organizacji ISO (ISO 1087-1:2000, 2000; ISO 704:2000, 2000) oraz Polskiego Komitetu Normalizacyjnego (PN-ISO 1087-1:2004, 2004) terminy oznaczają desygnat lub etykietę pojęcia. Termin jest obecnym w zbiorze D ciągiem znaków identyfikującym nienazwane pojęcie. Pojęcie jest więc reprezentowane w dokumencie poprzez wystąpienie terminu. Definicja relacji pojęcia do terminu jest zgodna z obowiązującą definicją z reprezentacji wiedzy (Sowa, 2000a) oraz obowiązujących norm organizacji ISO (ISO 860:2007, 2007).

S jest zbiorem synonimów $S = \{s_1, \ldots, s_n\}$, w którym każdy s_n jest niepustym zbiorem terminów powiązanych relacją synonimiczności. Zbiór S uzyskuje się przy pomocy funkcji Φ:

$$\Phi : 2^T \times LR \times K_{sim} \times \phi \rightarrow S, \tag{3.4}$$

gdzie K_{sim} zbiór progów klasyfikacji synonimów, a ϕ zbiór metod ekstrakcji synonimów. Siła relacji synonimiczności pomiędzy terminami jest częścią np. popularnych tezaurusów z rodziny WordNet. Cechą wyróżniającą poszczególne modele jest wartość ze zbioru ϕ.

C jest zbiorem pojęć $C = \{c_1, c_2, \ldots, c_n\}$, który jest wynikiem ekstrakcji pojęć przy pomocy funkcji Δ:

$$\Delta : S \times LR \times \delta \times K_{con} \rightarrow C, \tag{3.5}$$

gdzie δ oznacza zbiór metod ekstrakcji pojęć, a K_{con} stanowi zbiór progów klasyfikacji pojęć.

A jest zbiorem aksjomatów dziedzinowych:

$$A = \{A_1 \Rightarrow B_{11}, B_{12}, \ldots B_{1m_1}; \ldots A_n \Rightarrow B_{n1}, \ldots B_{nm_j}\}, \tag{3.6}$$

gdzie każde A_n jest nagłówkiem wyrażenia n oraz każde $B_{n1}, \ldots, B_{nm_j}$ jest ciałem wyrażenia n. Zbiór $m_1 \ldots m_j$ składa się z elementów ciała dla których każde A_n może być różne.

TR jest zbiorem relacji taksonomicznych, które zachodzą pomiędzy elementami zbioru C:

$$TR = \{isa(c_{i1}, c_{j1}), \ldots isa(c_{in}, c_{jm})\}, \tag{3.7}$$

gdzie $c_{i1} \ldots c_{in}$ oznacza zbiór podmiotów relacji isa, natomiast $c_{j1} \ldots c_{jm}$ jest zbiorem obiektów relacji isa oraz $\forall\, isa\ i \neq j$.

NTR jest zbiorem nazwanych relacji nietaksonomicznych:

$$NTR = \{rel_1(c_{x_1}, c_{y_1}), rel_2(c_{x_2}, c_{y_2}), \ldots rel_n(c_{x_n}, c_{y_n})\}, \tag{3.8}$$

gdzie $rel_1, \ldots rel_n$ to zbiór nazwanych relacji nietaksonomicznych, $c_{x_1}, \ldots c_{x_n}$ to podmioty NTR, $c_{y_1}, \ldots c_{y_n}$ to obiekty NTR. Każdy element NTR oznacza relację dla danej dziedziny. Na przykład znaczenie nazwanej relacji nietaksonomicznej $worksAt(x, y)$ powinno być interpretowane jako binarna relacja nietaksonomiczna, której pierwszym argumentem (podmiotem) jest instancja pojęcia $Person$, a drugim argumentem jest element zbioru $Organization$.

3.1 Dualność modeli

Przedstawione modele ekstrakcji zawierają dwie postaci funkcji, jedna oznaczoną małą literą grecką jako argument funkcji oznaczonej wielką literą grecką. Na przykład, równanie 3.3 zawiera zarówno funkcję Γ, jak i γ.

Dualność modeli wynika z następujących przesłanek:

- Celem metamodelu jest uogólnienie istniejących metod ekstrakcji. Poszczególne modele ekstrakcji nie mogą odnosić się wyłącznie do jednej metody ekstrakcji, lecz pozwalać na ich abstrakcję.
- W każdym modelu ekstrakcji wyróżnić można elementy, które są niezmienne niezależnie od konkretnej postaci funkcji ekstrakcji.

Z tych powodów funkcja oznaczoną małą literą grecką stanowi konkretną metodę ekstrakcji, np. dla równania 3.3 — metodę ekstrakcji terminologii. Zarówno postać tych funkcji, jak i ich argumenty, będą inne przy każdym podejściu do ekstrakcji. Wielką literą grecką oznaczone są natomiast funkcje, które są niezależne od konkretnej postaci funkcji ekstrakcji. Uogólniona metoda ekstrakcji terminologii Γ ma taki sam kształt niezależnie od postaci funkcji γ, tj. zarówno zbiór LA, jak i K_{term} nie ulega zmianie.

Uzyskany poziom abstrakcji zapewnia, że zastosowanie różnych metod ekstrakcji nie wpływa na postać modeli ekstrakcji. Cecha ta jest przydatna m.in. przy ewaluacji metod.

W dalszych częściach pracy każdy z modeli ekstrakcji przedstawiony jest przy pomocy funkcji ekstrakcji oznaczonych małą literą grecką. W celu dalszej dyskusji zaprezentowany w przypadku metamodelu poziom ekstrakcji jest zbędny. Wykorzystana tym samym zostanie metoda abstrahowania wyodrębniającego (skupienie się na konkretnej metodzie, tj. elemencie modelu).

3.2 Modele ekstrakcji

Każdy z przedstawionych elementów metamodelu (3.1) wymaga funkcjonalnie odrębnego procesu ekstrakcji. Procesy te reprezentowane są przy pomocy funkcji oznaczonych jako metody ekstrakcji poszczególnych składowych metamodelu. Niniejszy rozdział dokonuje podziału na metody ekstrakcji znane z innych opracowań (głównie synteza metod przedstawionych w rozdziale 2.) oraz metody ekstrakcji wprowadzone przez autora Abramowicz i in. (2008); Abramowicz i Wisniewski (2008a,b).

3.2.1 Ekstrakcja terminologii

Ekstrakcja terminologii zgodnie z (3.3) jest funkcją odwzorowującą anotację lingwistyczną LA w zbiór wszystkich podzbiorów T przy pomocy metody ekstrakcji terminologii γ.

W skład funkcji γ nie wchodzą zadania:

- ekstrakcji kolokacji (Smadja, 1993; Xu i in., 2002) — por. sekcję 2.2.1 na stronie 36 oraz sekcję 2.2.15 na stronie 52,
- ekstrakcji nawiązań (Poesio i in., 2002; Vieira i Poesio, 2000) — por. sekcję 2.2.19 na stronie 58,
- ekstrakcji akronimów (Torii i in., 2006; Schwartz i Hearst, 2003; Okazaki i Ananiadou, 2006; Adar, 2004; Chang i Schutze, 2006) — por. sekcję 2.2.14 na stronie 52,

- ekstrakcji bytów nazwanych (Jurafsky i Martin, 2000; Hammerton i in., 2002; Manning i Schutze, 1999).

Powodem wyodrębnienia tych procesów z metod ekstrakcji terminologii jest ich odrębna charakterystyka i odrębny zestaw narzędzi, na co wskazują m.in. Wermter i Hahn (2006); Okazaki i Ananiadou (2006); Poesio i in. (2002). Ponadto metody wypracowane w ramach tych zadań mogą być z powodzeniem wykorzystane w ekstrakcji terminologii w odpowiadających im zagadnieniach, zwiększając efektywność każdej z metod. Ich zastosowanie nie ma więc wpływu na efektywność podstawowej postaci konkretnej implementacji funkcji γ.

Funkcja γ w niniejszej pracy jest funkcją klasyfikacji terminu dla wyrazu lub sekwencji wyrazów w_n znajdującego się w zbiorze LA, tj. funkcją prawdopodobieństwa warunkowego opartą na częściach mowy (pos_n) dla poszczególnych wyrazów lub sekwencji wyrazów:

$$P(X) = P(term_k|pos_1, \ldots, pos_n), \tag{3.9}$$

gdzie:
n oznacza liczbę węzłów modelu n-gram,
k oznacza pozycję węzła oznaczonego jako termin,
$1 \leqslant k \leqslant n$,
$term$ oznacza węzeł oznaczony jako termin,
pos oznacza węzeł reprezentowany poprzez znacznik części mowy.

Motywacja wyboru kształtu funkcji γ, wraz ze szczegółowym wyprowadzeniem, budową modelu oraz metodą szacowania wartości modelu przedstawione są w rozdziale 5.

3.2.2 Ekstrakcja synonimów

Ekstrakcja synonimów zgodnie z równaniem (3.4) jest funkcją odwzorowującą zbiór wszystkich podzbiorów T w zbiór synonimów S przy użyciu zasobów lingwistycznych LR ze zdefiniowaną relacją synonimiczności przy pomocy metody ekstrakcji synonimów ϕ. Tezaurusy z rodziny WordNet (Fellbaum, 1998), FrameNet (Narayanan i in., 2003), Roget (Kipfer, 2006), czy polski Słowosieć (Piasecki i Broda, 2007) posiadają wskazania podobieństwa synonimicznego o sprawdzonej jakości. Metoda ϕ polega na sprawdzeniu wszystkich kombinacji terminów i połączeniu tych o wartości relacji synonimicznej przekraczającej zdefiniowany próg k_{sim}. Algorytm sprowadza się do zastosowania klasycznej analizy skupień metodą aglomeracyjną.

Efektywność metody ekstrakcji synonimów Φ jest bezpośrednim przeniesieniem jakości zdefiniowanych relacji synonimicznych w zbiorze zasobów lingwistycznych LR.

3.2.3 Ekstrakcja pojęć

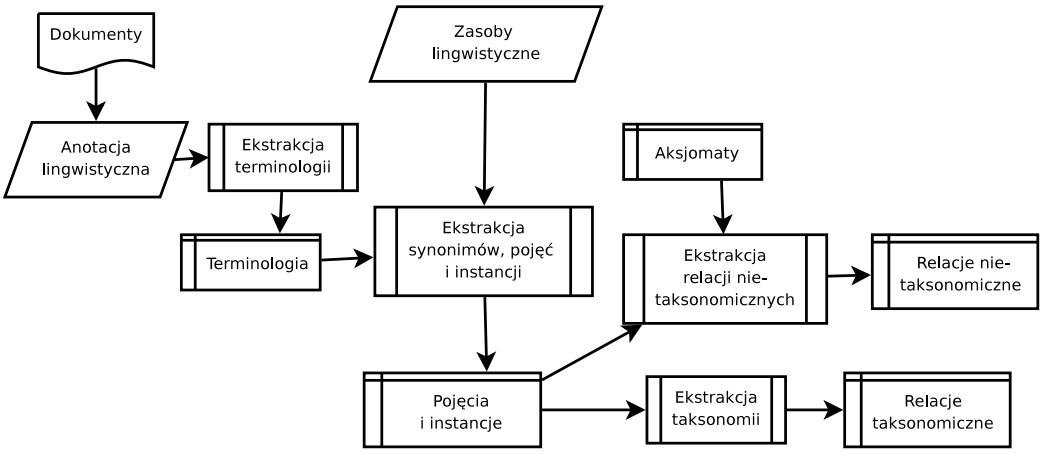
Ekstrakcja pojęć zgodnie z równaniem (3.5) jest funkcją odwzorowującą zbiór S składający się z elementów zbioru T w zbiór pojęć C przy użyciu zasobów lingwistycznych LR przy pomocy metody ekstrakcji pojęć δ. Funkcja δ jest funkcją klasyfikacji pojęć. W przedstawionym modelu ekstrakcji terminologii funkcja δ jest tzw. testem rozszerzenia (Buitelaar i Cimiano, 2006). Oznacza to, że δ klasyfikuje synonim s_n w przypadku realizacji leksykalnej s_n lub składających się na niego terminów. W przypadku przejścia testu rozszerzenia powstaje pojęcie.

Przy okazji ekstrakcji pojęć dokonywana jest prosta ekstrakcja instancji, która jest również funkcją klasyfikacji. W przypadku negatywnego wyniku testu rozszerzenia sprawdzana jest klasyfikacja POS synonimu lub terminu w zbiorze anotacji lingwistycznej LA. W przypadku klasyfikacji jako byt nazwany w jednym z dozwolonych znaczników wykorzystanego zbioru anotacji lingwistycznej LA analizowane wyrażenie jest klasyfikowane jako instancja. Na przykład podczas wykorzystania tezaurusa WordNet oraz ANNIE jako LR, klasyfikacja instancji następuje w przypadku nierealizowalności leksykalnej w WordNet oraz klasyfikacji POS jako: Person, Organization, Location, Address, Date lub Money.

Przedstawione metody klasyfikacji pojęć i instancji należą do klasy niewymagających pod względem zasobów lingwistycznych. Stanowią tym samym dobrą metodę do uzyskania zadowalających wyników dla celów ekstrakcji relacji.

3.2.4 Ekstrakcja relacji taksonomicznych

Ekstrakcja relacji taksonomicznych zgodnie z równaniem (3.7) jest funkcją odwzorowującą zbiór pojęć C w zbiór relacji taksonomicznych TR. W przedstawionym modelu zakłada się, że metoda ekstrakcji taksonomicznych jest zgodna z obecnie obowiązującymi rozwiązaniami opisanymi w rozdziale 2. Wynikiem działania metody ekstrakcji relacji taksonomicznych jest zbiór TR, który stanowi jeden z produktów końcowych procesu uczenia ontologii.



Rysunek 3.1: Ogólny podział funkcjonalny z podziałem na zdefiniowane procesy (zadania uczenia ontologii z tekstu) oraz wykorzystywane klasy zasobów

3.2.5 Ekstrakcja relacji nietaksonomicznych

Ekstrakcja relacji nietaksonomicznych zgodnie z równaniem (3.8) jest funkcją odwzorowującą zbiór pojęć C w zbiór relacji nietaksonomicznych NTR.

Przedstawiony metamodel składa się z nowej funkcji ekstrakcji relacji nietaksonomicznych opartej na założeniu sprzężenia zwrotnego pomiędzy aksjomatami dziedzinowymi a informacją lingwistyczną zawartą w zbiorze LA. Przedstawiona funkcja ekstrakcji relacji nietaksonomicznych odwzorowuje zbiór aksjomatów dziedzinowych A oraz zbiór anotacji lingwistycznych LA w zbiór relacji nietaksonomicznych NTR.

Szczegółowy opis metody ekstrakcji relacji nietaksonomicznych opartej na założeniu sprzężenia zwrotnego znajduje się w rozdziale 6.

3.3 Architektura

Do tej pory w rozdziale przedstawiono model teoretyczny procesu uczenia ontologii z tekstu. Metamodel posiada ponadto realizację modelu teoretycznego w postaci podziału funkcjonalnego i architektonicznego, które zawierają odpowiednio architekturę funkcjonalną oraz odwzorowanie podziału funkcjonalnego na konkretne technologie.

3.3.1 Podział funkcjonalny

Podział funkcjonalny polega na wyodrębnieniu funkcjonalnie jednorodnych komponentów w tzw. bloki architektoniczne oraz zdefiniowanie relacji zacho-

dzących pomiędzy nimi. Ogólny podział funkcjonalny metamodelu przedstawiono na rysunku 3.1. Ze względu na oznaczenie standardowego przepływu danych, diagram ten jest równocześnie modelem behawioralnym.

Komponenty znajdujące się po lewej stronie rysunku 3.1 oznaczają zbiór D, czyli korpus dokumentów oraz powstający na jego podstawie przy pomocy funkcji Λ zbiór anotacji lingwistycznych LA. Przedstawiony metamodel obejmuje przy realizacji funkcji Λ następujące etapy:

- określanie granicy zdań,
- tokenizacja,
- anotacja części mowy,
- analiza morfologiczna, w tym analiza fleksyjna, podstawy oraz złożoności,
- lematyzacja,
- rozpoznawanie bytów nazwanych.

Ponadto metamodel obejmuje analizę wyrażeń oraz funkcji gramatycznych wyrażeń. Analiza lingwistyczna przeprowadzona została przy użyciu narzędzi ANNIE (Cunningham i in., 2007), Ballie (Nadeau, 2005) dla języka angielskiego oraz SProUT (Piskorski i in., 2005) dla języka polskiego.

Format anotacji w zbiorze LA został dobrany w taki sposób, aby umożliwić rzetelną ewaluację oraz dostarczyć odpowiednich właściwości korpusu D dla założonego metamodelu. Wybranie formatu anotacji OntoLT (Buitelaar i in., 2004a) podyktowane jest jego dużą popularnością, co umożliwia weryfikację przeprowadzonej ewaluacji oraz strukturą mającą zastosowanie dla przedstawionych mechanizmów. Modułowa struktura anotacyjna umożliwia ponadto dobór zakresu informacji lingwistycznej dla różnych zastosowań i właściwości korpusu.

Szczegółowy zakres przeprowadzonych analiz dla funkcji Λ oraz formatu zbioru LA przedstawiony jest w rozdziale poświęconym eksperymentom (rozdział 7).

Zadanie ekstrakcji terminologii jest przeprowadzone przy pomocy funkcji Γ (3.3). Argumentem funkcji Γ jest korpus dokumentów D pod postacią zbioru anotacji lingwistycznych LA. Postać metody γ jest różny w zależności od zastosowanego podejścia. Metamodel zakłada zastosowanie jednej z metod przedstawionych w rozdziale 2., np:

- modele wywodzące się z teorii informacji, w tym analiza współwystępowalności,

- podejścia statystyczne: TFIDF, KFIDF, χ^2 lub metoda wartości C/NC (Frantzi i in., 2000),
- podejścia wywodzące się z uczenia maszynowego (ML), w tym modele Markova.

W przedstawionej architekturze proponuje się wykorzystanie nowej metody opartej na modelach Markova, której charakterystyczną cechą jest oparcie modelu na tzw. oknie kontekstowym dla terminu. Szczegółową dyskusję nad tą metodą przedstawiono w rozdziale 5. Na poziomie metamodelu postać funkcji γ nie ma jednak znaczenia.

Duża jednorodność funkcjonalna ekstrakcji synonimów, pojęć oraz instancji spowodowała umieszczenie wszystkich tych elementów w postaci jednego bloku architektonicznego na rysunku 3.1. Funkcja ekstrakcji synonimów Φ oraz ekstrakcji pojęć Δ działają głównie opierając się na zasobach lingwistycznych LR. Jakość tych zasobów w zdecydowanej mierze kształtuje jakość samych funkcji.

Zbiór pojęć C jest podstawą ekstrakcji relacji taksonomicznych oraz nietaksonomicznych. Funkcja ekstrakcji relacji taksonomicznych jest dowolną funkcją sklasyfikowaną w rodziale 2. jako metoda ekstrakcji relacji taksonomicznych. Funkcja ekstrakcji relacji nietaksonomicznych jest dowolną funkcją sklasyfikowaną w rozdziale 2. jako metoda ekstrakcji relacji nietaksonomicznych. W przedstawionej architekturze proponuje się wykorzystanie nowej metody ekstrakcji relacji nietaksonomicznych, której charakterystyczną cechą jest tzw. cykl zwrotny pomiędzy ogólnie przyjętymi prawami dziedziny (aksjomatami), a informacją lingwistyczną. Szczegółowy opis proponowanej metody znajduje się w rozdziale 6.

3.3.2 Podział architektoniczny

Podział architektoniczny służy do zaproponowania konkretnych technologii w celu dostarczenia funkcjonalności zawartych w podziale funkcjonalnym. Tabela 3.1 przedstawia odwzorowanie najistotniejszych funkcjonalności do konkretnych technologii. Odwzorowanie jest ogólne, co oznacza, że zaproponowane technologie umożliwiają budowanie rozwiązań uniwersalnych dla metamodelu, tj. np. dla różnych korpusów, w tym języków naturalnych.

Funkcjonalności przedstawione w tabeli 3.1 nie mają bezpośredniego przełożenia na wydzielone w ramach podziału funkcjonalnego bloki architektoniczne. Jedynym kryterium wydzielenia funkcjonalności była odrębność technologiczna.

Funkcjonalności	Wykorzystane technologie
Anotacja lingwistyczna LA	ANNIE, Ballie, SProUT
Format zbioru anotacji lingwistycznych LA	Format OntoLT
Ekstrakcja terminologii	Metoda ML
Klasyfikacja terminologii	ANNIE, WordNet, Słowosieć
Serializacja aksjomatów	SWRL
Budowa ontologii	Jena
Język ontologii	OWL DL
Wnioskowanie ze wsparciem SWRL	Pellet

Tabela 3.1: Podział architektoniczny

Wynik procesu	Zależność
Terminologia	—
Synonimy i pojęcia	Terminologia
Relacje taksonomiczne	Terminologia, pojęcia
Aksjomaty	—
Relacje nietaksonomiczne	Terminologia, pojęcia, aksjomaty

Tabela 3.2: Zależność pomiędzy wynikami procesu uczenia ontologii

3.4 Dobór zakresu badań

Podział funkcjonalny (rysunek 3.1) przedstawia oprócz identyfikacji bloków architektonicznych, także typy bloków architektonicznych oraz relacje pomiędzy nimi zachodzące. Najbardziej istotnym dla potrzeb analizy zakresu badań jest typ oznaczający wynik procesu. Podział funkcjonalny przedstawia pięć wyników procesu uczenia ontologii z tekstu: terminologię, pojęcia, relacje taksonomiczne, relacje nietaksonomiczne oraz aksjomaty. Większość z tych wyników jest efektem działania metod składających się na postać metamodelu. Tabela 3.2 przedstawia zależności pomiędzy wynikami procesu uczenia ontologii z tekstu.

Z wyjątkiem aksjomatów wszystkie wyniki zależne są od terminologii. Terminologia zatem charakteryzuje się dużym wpływem na postać innych wyników, co powoduje, że wyniki działania funkcji ekstrakcji terminologii Γ propagowane są na pozostałe wyniki. Każdy błąd popełniony w tej fazie ma swoje odzwierciedlenie w późniejszych wynikach. Na podstawie tabeli 3.2 można więc określić priorytet istotności wyników na podstawie zależności pomiędzy nimi. Wyniki, które nie są argumentami innych funkcji są najmniej istotne, ponieważ ewentualne błędy nie są propagowane na inne wyniki.

Podział funkcjonalny obrazuje również zależności pomiędzy wynikami procesu uczenia ontologii, a zasobami zewnętrznymi. W przypadku ekstrakcji

Wynik procesu	Zależność	Naukowość	Relewancja	Średnia
Terminologia	1	1	1	1
Synonimy	2	4	4	3,33
Pojęcia	3	5	3	3,66
Relacje taksonomiczne	4	1	5	3,33
Relacje nietaksonomiczne	4	1	1	2

Tabela 3.3: Priorytety dla doboru zakresu badań. Skala 1-5, przy czym 1 oznacza najbardziej istotny wpływ, a 5 oznacza brak istotności

synonimów oraz pojęć zależność od zasobów lingwistycznych LR jest znacząca. Większość problemów sprowadza się do jakości zasobu lingwistycznego, a nie do jakości samej metody ϕ lub δ. Nie jest to zatem problem naukowy, przynajmniej jeśli chodzi o zagadnienie uczenia ontologii. Ekstrakcja terminologii oraz relacji w ogólnym przypadku nie wymaga zasobów lingwistycznych, co przenosi cały ciężar jakości na same metody.

Ostatecznie decydujący wpływ na dobór zakresu badań w ramach metamodelu ma relewancja wyników dla użytkownika końcowego. To problem użytkownika końcowego decyduje o priorytecie wyników. Z przeprowadzonej w rozdziale 1. analizy biznesowej wynika, że najbardziej kluczowe dla użytkowników są terminologia oraz relacje nietaksonomiczne.

Tabela 3.3 przedstawia zestawienie priorytetów ze względu na omawiane cechy, tj. zależność wyników, naukowość zadania oraz relewancję problemu. Ostatnia kolumna przedstawia ważoną miarę istotności wyników. Wynika z niej, że zdecydowanie najbardziej istotnymi fazami uczenia ontologii są ekstrakcja terminologii oraz ekstrakcja relacji. Z tego tytułu niniejsza praca ma na celu zaproponowanie metod ekstrakcji terminologii oraz relacji.

Rozdział 4

Anotacja lingwistyczna

Pierwsze dwa elementy metamodelu (równanie 3.1) to zbiór dokumentów D oraz anotacji lingwistycznych LA. W niniejszym rozdziale przedstawiona zostanie funkcja Λ, a w szczególności jej elementy, czyli zbiór dokumentów D, zbiór anotacji lingwistycznych LA oraz metoda budowania zbiorów lingwistycznych λ.

W terminologii statystycznej korpus dokumentów stanowi reprezentatywną próbę populacji generalnej. Odpowiednio dobrany korpus, podobnie jak reprezentatywna próba, służy do estymacji właściwości populacji generalnej.

W kontekście zastosowań informatyki ekonomicznej, korpus jest kolekcją dokumentów przeznaczonych do estymacji cech populacji ogólnej, czyli wszystkich dokumentów znajdujących się w formie elektronicznej. Niestety, liczba ta jest niemierzalna, co najmniej z następujących powodów:

- rozmiar sieci Internet nie pozwala na szybkie zliczenie dostępnych dokumentów,
- zmienność sieci Internet powoduje, że w momencie ukończenie procesu liczenia dokumentów, liczba ta jest już nieaktualna,
- istnienie Ukrytego Internetu — wiele dokumentów jest po prostu niedostępnych dla mechanizmów, które mają na celu szacowanie liczby dokumentów (Kaczmarek, 2007).

Zgodnie z motywacją prowadzonych badań metamodel powinien posiadać cechę uniwersalności, tj. powinien mieć zastosowanie w jak najwyższej liczbie kontekstów. Kontekst wykorzystania opracowanych metod można wyróżnić w zależności od następujących cech:

- dziedzina,

- język naturalny,
- styl narracji.

Podejścia do uczenia ontologii, które są uniwersalne we wszystkich wymienionych typach kontekstów, nie istnieją. Problematyczne są w zasadzie wszystkie typy kontekstów, ponieważ powodują, że zarówno gramatyka języka, jak i używane słownictwo znacząco się różnią. Dla przykładu korpus Browna (Francis i Kucera, 1979) charakteryzuje się odmianą języka angielskiego powszechnie już zanikającą.

Dobrym sposobem do ewaluacji uniwersalności jest zastosowanie opracowywanych metod do różnych korpusów, najlepiej różniących się typem. Testowanie metod na korpusach z różnych dziedzin, w różnym języku oraz o zróżnicowanym stylu narracji jest niespotykane. Nie tylko z powodu obaw przez ewaluacją uniwersalności, ale również z powodu trudności, jakie przysparza konstrukcja korpusu.

Zadaniem anotacji jest dołączenie do tekstu w języku naturalnym zbioru anotacji lingwistycznych. Zbiór anotacji lingwistycznej zawiera metadane tekstu i jego struktury. Strukturę tekstu opisuje model statyczny anotacji, natomiast procesy składające się na anotację i uzupełnienie zbioru anotacji tworzą model behawioralny anotacji. Opis modelu behawioralnego oraz statycznego zostanie przedstawiony w kolejnych sekcjach.

4.1 Procesy anotacyjne

Model behawioralny anotacji tworzą procesy anotacyjne. Procesy anotacyjne mają na celu uruchomienie ściśle określonych procesów na tekście w języku naturalnym. Wejściem do procesu anotacji jest czysty tekst, niezawierający żadnych informacji dodatkowych oraz nieposiadający jawnie żadnej struktury. Tekst w języku naturalnym posiada oczywiście niejawną strukturę wynikającą z gramatyki danego języka. Właśnie ta ukryta warstwa jest bardzo interesująca dla metamodelu uczenia ontologii.

Procesy anotacyjne dążą do stworzenia zbioru anotacji w pewnym standardzie (struktura anotacji). Opis poszczególnych standardów anotacji znajduje się w sekcji 4.2.

Przedstawione procesy anotacyjne są oferowane przez znaczną liczbę narzędzi lingwistycznych. Ich zestawienie wykracza poza ramy niniejszej pracy; zainteresowani odesłani są do dobrze zapowiadającego się portalu informacyjnego projektu OntoWeb[1].

[1] `http://ontoweb-lt.dfki.de/langtech_index.htm`

Do najważniejszych procesów anotacyjnych należą w kolejności stosowania:

1. Części mowy i morfologia.
2. Analiza syntaktyczna.
3. Analiza semantyczna.

4.1.1 Części mowy i morfologia

Lingwiści grupują występujące w danym języku naturalnym wyrazy według klas charakteryzujących się podobnym zachowaniem syntaktycznym. Klasy te nazywają się kategoriami gramatycznymi lub syntaktycznymi, również znane jako części mowy (POS — Part of Speech). W warstwie syntaktycznej części mowy występują przeważnie w określonym kontekście zdaniowym, np. rodzajnik przed rzeczownikiem. W warstwie semantycznej części mowy przeważnie pełnią podobną rolę — czasowniki lub grupy czasowników przeważnie pełnią funkcję orzeczenia w zdaniu.

Analiza części mowy i morfologii składa się z dwóch procesów: analizy części mowy oraz analizy morfologicznej. Oba procesy wzajemnie się przenikają, tj. znajomość POS ułatwia analizę morfologiczną, natomiast sama analiza często jest najlepszą metodą poprawnej klasyfikacji POS.

Analiza części mowy jest problemem klasyfikacyjnym. Posiadając skończony zbiór klas części mowy problem polega na przyporządkowaniu każdego wyrazu do dokładnie jednej klasy. Istnieją dwie grupy metod analizy części mowy: metody regułowe oraz metody statystyczno-stochastyczne. Metody regułowe są oparte na regułach, np.: "każdy wyraz kończący się na -ing jest czasownikiem". Metody te są zależne od języka naturalnego. Niestety, większość narzędzi oparta jest w części na metodach regułowych, co tworzy poważne problemy w wykorzystaniu narzędzi w więcej niż jednym języku naturalnym. Metody statystyczno-stochastyczne oparte są bądź na częstości występowania wyrazów oraz części mowy w poszczególnych kontekstach zdaniowych, bądź na modelach prawdopodobieństwa szacujących prawdopodobieństwo wystąpienia danej klasy na podstawie cech kontekstu.

Tradycyjnie wyróżnia się osiem podstawowych części mowy. Dla celów analizy lingwistycznej potrzeba jednak znacznie bardziej szczegółowej klasyfikacji dlatego opracowano tzw. zbiory znaczników POS. Zbiory znaczników POS zawierają zamknięty zbiór znaczników określających klasę wyrazu. Wszystkie systemy znaczników POS opierają się na standardowym podziale części mowy, co oznacza, że podstawowe typy wyrazów mogą zostać przeniesione pomiędzy zbiorami wskaźników. Na przykład, czasownik zawsze

Korpus	Liczba znaczników
Brown (Francis i Kucera, 1979)	87
Penn Treebank (Marcus i in., 1993)	45
Susanne (Sampson, 1995)	353
CLAWS1 (Garside i Smith, 1997)	132
CLAWS2 (Garside i Smith, 1997)	166
CLAWS5 (Garside i Smith, 1997)	62
London-Lund (Greenbaum i Svartvik, 1990)	197

Tabela 4.1: Najpopularniejsze zbiory znaczników POS dla języka angielskiego

Klasa wyrazu	Wyraz	Brown	Penn	c5
przymiotnik	happy	JJ	JJ	AJ0
rzeczownik	data	NN	NN	NN0
czasownik	take	VB	VB	VVI
przyimek TO	to	IN	TO	PRP

Tabela 4.2: Porównanie trzech zbiorów znaczników POS

oznacza tę samą klasę wyrazów, wyróżnia się natomiast różne typy samego czasownika.

W tabeli 4.1 przedstawiono najpopularniejsze systemy znaczników POS dla języka angielskiego. Najstarszym systemem znakowania POS jest Brown. Najczęściej używanym jest Penn Treebank. Wywodzący się z Penn Treebank system Susanne posiada z kolei najbardziej szczegółowy podział znaczników. Dla celów uczenia ontologii najważniejszym czynnikiem jest popularność systemu znakowania. Im system jest bardziej powszechny, tym większa szansa na większą liczbę zaanotowanego tekstu oraz wykorzystujących go narzędzi.

W tabeli 4.2 przedstawiono porównanie nazewnictwa znaczników POS tych samych klas wyrazów dla języka angielskiego. Widoczny jest duży stopień unifikacji standardów znaczników POS. Najważniejsze klasy oznaczone są identycznymi znacznikami, zwłaszcza w przypadku zbiorów Brown i Penn Treebank. Im jednak bardziej szczegółowa klasa, tym różnic jest coraz więcej.

Analiza morfologiczna to dziedzina przetwarzania tekstu naturalnego zajmująca się różnymi formami wyrazów. Części mowy występują bowiem w różnych odmianach: rzeczowniki poddane są fleksji, czasowniki występują w różnych czasach, przymiotniki w różnych stopniach. Morfologia to dziedzina zajmująca się analizą różnych odmian danej części mowy.

Języki naturalne charakteryzują się znacząco różną morfologią. Język angielski jest pod tym względem językiem niezwykle prostym, np. czasownik w formie podstawowej może wystąpić tylko w czterech formach. W zasadzie wszystkie inne języki naturalne są znacznie trudniejsze w przetwarzaniu

morfologicznym. Zwłaszcza języki słowiańskie, w tym język polski, charakteryzują się dużą złożonością morfologiczną.

Analiza morfologiczna składa się z trzech głównych procesów: analizy fleksyjnej, derywacji oraz składania wyrazów. Fleksja to dział gramatyki zajmujący się odmianą wyrazów. Analiza fleksyjna powoduje zasilenie zbioru anotacji o informacje dotyczące modyfikacji formy podstawowej, czyli lematu. Fleksja nigdy nie zmienia klasy wyrazu, znaczenie modyfikuje bardzo nieznacznie. Częścią analizy fleksyjnej jest sprowadzenie wyrazu do formy podstawowej. Fleksja wyrazu jest zależna od części mowy, np. dla rzeczownika dotyczy:

- liczby — liczba pojedyncza lub mnoga,
- płci — rodzaj męski, żeński, nijaki,
- przypadku — mianownik, dopełniacz, celownik, biernik, ...

Derywacja to proces transformacji z formy podstawowej i znacznie częściej powoduje zmianę zarówno części mowy, jak i znaczenia. Analiza derywacyjna sprawdza, czy wyraz powstał z innej formy i nie jest jego odmianą. Na przykład przysłówek *szeroko* jest derywacją przymiotnika *szeroki*. Składanie wyrazów dotyczy sytuacji, w której dwa lub więcej wyrazy zostają połączone w jedną całość niosącą ze sobą odrębne znaczenie (np. *downtown*).

Wyzwania stawiane przez analizę części mowy i morfologię dla języka angielskiego ograniczają się do posiadania odpowiedno dużego słownika. Jedynym problemem wydaje się być wieloznaczność wyrazów, tj. sytuacji, w których składniowo ten sam wyraz ma różne znaczenia. Na przykład angielski wyraz *train* znaczy zarówno *trenować*, jak i *pociąg*, w związku z tym może przynależeć do dwóch różnych klas części mowy. Narzędzia oferujące analizę części mowy osiągają skuteczność na poziomie 98%.

4.1.2 Analiza syntaktyczna

Wyrazy nie pojawiają się w tekście w sposób całkowicie przypadkowy. Każdy język naturalny nakłada ograniczenia na porządek występowania klas wyrazów. Ponadto, pewne grupy klas wyrazów występują nader często razem tworząc struktury zwane wyrażeniami (ang. *phrases* lub *chunks*). Na przykład statystycznie często przed czasownikiem w formie podstawowej występują wyrażenia wskazujące na podmiot w zdaniu. Wyrażenia są więc grupami wyrazów o podobnych właściwościach syntaktycznych, występujących statystycznie często i pełniących w zdaniu łączną funkcję.

Analiza syntaktyczna zajmuje się analizą regularności oraz ograniczeń w porządku wyrazów oraz struktury wyrażeń. Analiza syntaktyczna pomaga zrozumieć znaczenie zdania biorąc pod uwagę znaczenie poszczególnych wyrazów. Na przykład oba zdania:

"UniCredito przejęło bank BPH",
"BPH przejęło bank UniCredito"

używają dokładnie takich samych wyrazów, a jednak ich znaczenie jest różne. To właśnie porządek wyrazów decyduje o znaczeniu. Przykład ten przedstawia również ograniczenia języka naturalnego. Pojawienie się rzeczownika po czasowniku jest dopuszczalne, podczas gdy pojawienie się kolejnego czasownika (np. *BPH przejęło sprzedało*) jest niedopuszczalne.

Porządek wyrazów w zdaniu względnie łatwo jest badać w języku naturalnym o ustabilizowanym porządku wyrazów (np. język angielski). Dużo trudniej jest badać zachowania w porządku wyrazów języka naturalnego, który charakteryzuje *dowolny szyk zdania*. Przykładami języków naturalnych o dowolnym szyku zdania są języki niemiecki i polski.

Na podstawie porządku wyrazów powstają wyrażenia. Do najważniejszych wyrażeń należą:

Wyrażenia rzeczownikowe. W zdaniu rzeczownik najczęściej jest otoczony przez inne wyrazy, które go uzupełniają (np. *chytry i przebiegły lis cętkowany*) i tworzą w ten sposób wyrażenie jednorodne znaczeniowo. Wyrażenia rzeczownikowe posiadają *głowę* wyrażenia, które stanowi sam rzeczownik oraz wyrazy go modyfikujące. Wyrażenie rzeczownikowe jest najczęściej argumentem orzeczenia zdania.

Wyrażenia przyimkowe. Zgrupowane wyrazy pełniące funkcję dopełnienia wyrażenia rzeczownikowego.

Wyrażenia czasownikowe. Grupa wyrazów otaczająca czasownik i pełniąca jednorodną funkcję w zdaniu. Podobnie jak wyrażenie rzeczownikowe, posiada głowę będącą czasownikiem.

Wyrażenia przysłówkowe. Rzadziej występujące, pełnią funkcję dopełnienia wyrażenia czasownikowego.

4.1.3 Analiza semantyczna

Analiza semantyczna pozwala zrozumieć znaczenie zdania na podstawie znaczenia poszczególnych wyrazów i wyrażeń. Wykorzystuje przy tym analizę syntaktyczną, której wynikiem są wyrażenia w zdaniu. Analiza semantyczna

polega na analizie wyrażeń i traktuje je jako jednorodną całość. Wyrażenia w zdaniu pełnią określone funkcje gramatyczne, np. wyrażenie czasownikowe jest często orzeczeniem zdania, a wyrażenie rzeczownikowe podmiotem. Wynikiem analizy semantycznej na poziomie procesu anotacyjnego jest zatem określenie funkcji gramatycznych wyrażeń.

4.2 Standardy anotacji

W podejściach do formalizacji i uczenia ontologii stosuje się różne statyczne modele reprezentacji tekstu. Tekst w swojej źródłowej postaci składa się wyłącznie ze znaków. Przetwarzanie tekstu w takiej postaci jest jednak wysoce nieefektywne. Przy wykonywaniu jakichkolwiek operacji na czystym tekście pojawia się dodatkowo potrzeba przechowywania metadanych tekstu. W ten sposób powstaje potrzeba tworzenia standardów anotacji.

Standard anotacji powinien umożliwiać przechowywanie oraz efektywny dostęp do co najmniej następujących elementów:

- tekst źródłowy, który może być odtworzony bez zbędnych mechanizmów dodatkowych,
- metadane wynikające ze struktury logicznej tekstu (np. podział na zdania, wyrazy) — struktura ta jest wynikiem działania procesów anotacyjnych (por. 4.1),
- metadane wynikające z właściwości poszczególnych elementów lingwistycznych (np. lemat tokena) — właściwości te są również wynikiem działania procesów anotacyjnych.

Standardy anotacji nie są ze sobą kompatybilne. Przenoszenie informacji pomiędzy nimi jest w praktyce możliwe, ale rzadko uzyskuje się satysfakcjonujące wyniki, zarówno z powodu różnic wynikających z dopuszczalnej struktury logicznej, jak i dopuszczalnych typów oraz zakresu właściwości lingwistycznych. Opracowując metody uczenia ontologii należy więc dokonać wyboru formatu anotacji. Jest to jeden z najbardziej kluczowych wyborów, ponieważ definiuje zarówno ekspresywność anotacji, jak i użyteczność w przetwarzaniu. Istnieje pięć kluczowych przesłanek wyboru standardu anotacji:

Saturacja standardu. Im standard jest bardziej rozpowszechniony, tym więcej narzędzi oraz metod na nim operuje — przestrzeń badawcza jest tym większa.

Ekspresywność definiuje jak wiele można wyrazić. Na przykład: czy standard pozwala tylko na określony typ anotacji? Czy można typy anotacji definiować w sposób dowolny? Czy standard umożliwia przedstawienie wszystkich niezbędnych do analizy danych?

Otwartość powoduje, że standard można wykorzystać w zasadzie z każdym otwartym narzędziem. Jest to kluczowa cecha, ponieważ par narzędzie-standard jest tyle ile narzędzi.

Możliwość ewaluacji. Standard anotacji musi umożliwiać rzetelną ewaluację przeprowadzanych prac. Jeśli np. jest to format prawnie zastrzeżony, a autor danej metody nie ujawni zakresu znajdujących się w nim informacji, to metoda w zasadzie jest nieporównywalna.

Dostosowanie do uczenia. Niektóre standardy wywodzą się ze środowisk do uczenia ontologii, niektóre z dziedzin pokrewnych. Dostosowanie do uczenia wyraża się możliwością skorzystania w prosty sposób z informacji w nim zawartych na specyficznych etapach procesu uczenia ontologii.

Najbardziej popularne formaty anotacji lingwistycznej dla uczenia ontologii wywodzą się ze środowisk do uczenia ontologii oraz ekstrakcji informacji:

GATE to najpopularniejsze środowisko do budowania aplikacji inżynierii tekstu (Kenter i Maynard, 2005). GATE dzieli format anotacji na trzy główne, równorzędne poziomy: zbiór właściwości, zbiór anotacji oraz tekst. Zbiór właściwości określa metadane dokumentu, np. źródło, czy typ MIME. Tekst zawiera tekst źródłowy wraz ze strukturą dokumentu. Zbiór anotacji zawiera dowolny zbiór typów anotacji wraz z właściwościami. Model zaprezentowany w GATE cechuje się bardzo dużą ekspresywnością — zbiory anotacji można definiować dowolnie. Format GATE jest mocno rozpowszechniony — głównie ze względu na popularność samego narzędzia. Jest on jednak nadmiarowy dla procesu uczenia ontologii, co wynika z faktu, że został stworzony na potrzeby ekstrakcji informacji, nie uczenia ontologii (Cunningham i in., 2002).

SProUT jest narzędziem do ekstrakcji informacji, w tym do ekstrakcji informacji z języka polskiego (Piskorski i in., 2005). Do analizy morfologicznej wykorzystuje narzędzie do analizy języka polskiego — Morfeusz (Woliński, 2006). Pozwala na swobodne definiowanie zbiorów anotacji. Narzędzie jest często wykorzystywane w ekstrakcji informacji z języka polskiego (np. Abramowicz i in. (2006)).

OI Model to model danych obejmujący ontologię oraz instancje stworzone na potrzeby środowisk TextToOnto (Maedche i Staab, 2004) oraz, w wersji rozszerzonej, Text2Onto (Cimiano i Völker, 2005). OI model jest modelem o dużej ekspresywności, nie jest natomiast standardem otwartym, ani dostosowanym do uczenia ontologii. Wersja ze środowiska Text2Onto, tzw. POM (Probabilistic Ontology Model), jest modelem stochastycznym przechowującym wyniki działania metod uczestniczących w procesie uczenia ontologii. Ze względu jednak na wspomniany brak otwartości, a także kłopoty z utrzymaniem jego stabilności, format ten jest mało popularny.

OntoLT jest formatem anotacji stworzonym w ramach projektu MuchMore (Vintar i in., 2001) oraz wykorzystywanym jako domyślny w narzędziu OntoLT (Buitelaar i in., 2004a; Buitelaar, 2003). Oparty na języku XML charakteryzuje się otwartością oraz dużą powszechnością, zwłaszcza w dziedzinie uczenia ontologii. Zakres informacyjny jest jednak zdefiniowany, dlatego ekspresywność formatu nie jest tak wysoka, jak w przypadku innych formatów. Dla celów uczenia ontologii jest on jednak w zupełności wystarczający. Format OntoLT jest opisany szerzej w punkcie 4.3.

Własne formaty. Oprócz standardowych formatów anotacji można stosować również własne modele. Większość badaczy w ten sposób podchodzi do problemu. Własny format daje możliwość wyboru ekspresywności, stopnia otwartości, czy stopnia dostosowania do własnych potrzeb uczenia. Niestety, formaty własne mają jedną cechę dyskwalifikującą przy ewaluacji — różne formaty notacji dają różne możliwości. Nawet ta sama metoda ekstrakcji terminologii sprawdzać się będzie różnie w zależności od języka anotacji (czyli np. różnego stopnia ekspresywności). Porównywać dwa różne formaty anotacji jest bardzo trudno, nie mówiąc już o procesie przygotowania dokładnie takiego samego korpusu przy pomocy dwóch różnych języków anotacji. Dwie najpopularniejsze odmiany własnych formatów opierają się na składni XML oraz na modelach relacyjnych przechowywanych w relacyjnych bazach danych. Oba te podejścia zostały przedmiotem porównania w tabeli 4.3.

Porównanie najpopularniejszych formatów anotacji przedstawiono w tabeli 4.3. Z przeprowadzonej analizy wynika, że dla celów niniejszej pracy najbardziej dogodnym wyborem jest standard anotacji zastosowany w narzędziu OntoLT. Format ten daje bardzo dużą przewagę badawczą ze względu na możliwość ewaluacji oraz dostosowanie do procesu uczenia ontologii.

	GATE	SProUT	OI Model	OntoLT	DB	XML
Saturacja	1	1	1	2	0	0
Ekspresywność	2	2	2	1	2	2
Otwartość	2	0	0	2	0	2
Ewaluacja	1	1	0	2	0	0
Dostosowanie	0	0	2	2	0	1

Tabela 4.3: Porównianie najczęściej stosowanych języków anotacji. Skala 0–2, przy czym 0 oznacza brak cechy, 1 to częściowe, a 2 pełne wsparcie

Możliwość ewaluacji jest tym istotniejsza, ponieważ uczenie ontologii jest dziedziną młodą, wymagającą rzetelnej ewaluacji przeprowadzonych badań.

4.3 Struktura formatu anotacji

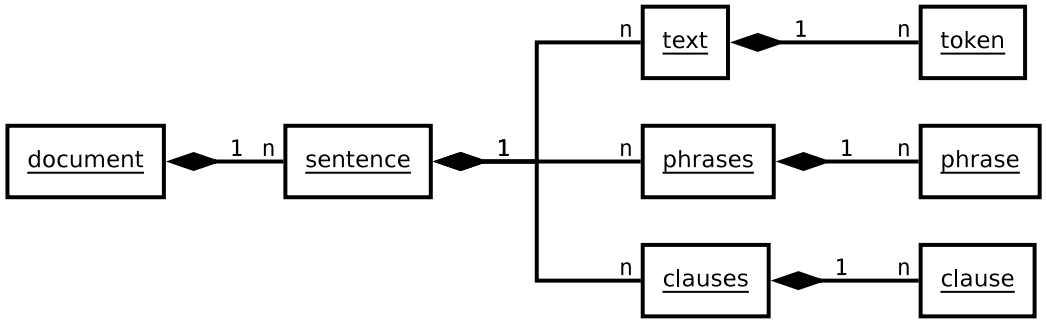
Format anotacji tworzy statyczną strukturę modelu tekstu oraz informacji lingwistycznej, czyli definiuje dopuszczalną strukturę, dopuszczalne typy danych oraz ich właściwości. Zgodnie z wyborem dokonanym w sekcji 4.2, zastosowano format anotacji OntoLT. Format ten jest standardem opartym na języku XML, definiuje więc schemat pliku XML i może być wyrażony w postaci pliku xsd.

Model domyślnego formatu anotacji składa się z dokładnie jednego elementu głównego *document*. Element główny *document* składa się z dowolnej liczby elementów *sentence*, czyli zdań. Na poziomie zdania następuje najistotniejszy podział modelu na trzy elementy: tekst (*text*), wyrażenia (*phrases*) oraz części zdania (*clauses*). Opis modelu w postaci diagramu UML przedstawiono na rysunku 4.1.

Podział na trzy warstwy jest zgodny z podziałem procesów anotacyjnych przedstawionych w sekcji 4.1. Poszczególne warstwy modelu statycznego obejmują więc informacje będące wynikiem działania odpowiednich procesów anotacyjnych.

W warstwie tekstu występują tokeny, czyli najmniejsze możliwe części tekstu posiadające osobne właściwości syntaktyczne. Każdy token opisany jest następującymi właściwościami:

- id — jednoznacznie identyfikuje tokena w dokumencie,
- str — przechowuje źródłowy tekst tokena,
- pos — identyfikuje część mowy tokena,
- lemma — przechowuje lemat tokena.



Rysunek 4.1: Struktura modelu anotacji OntoLT dla dokumentu tekstowego

Akronim	Wyrażenie	Tłumaczenie
PP	prepositional phrase	wyrażenie przyimkowe
NP	noun phrase	wyrażenie rzeczownikowe
SUBORD_CL	subordinate	wyrażenie podrzędne
AP	adjective phrase	wyrażenie przysłówkowe
AdvP	adverbial phrase	wyrażenie modyfikujące czasowniki
VG	verb group	wyrażenie czasownikowe
W	words	inne wyrażenia

Tabela 4.4: Dopuszczalne typy wyrażeń w domyślnym formacie anotacji

Przykładowy opis tokena dla wyrazu *usługi* wygląda w sposób następujący:

«token id=“21”, str=“usługi”, pos=“NNS”»
«lemma»usługa«/lemma»
«/token»

W warstwie wyrażeń oprócz unikatowego identyfikatora występują wskaźniki początkowego (*from*) oraz końcowego (*to*) tokena wyrażenia. Kluczową właściwością wyrażenia jest jego typ (*type*). Dopuszczalne typy wyrażenia przedstawione są w tabeli 4.4.

Przykładowe wyrażenie stworzone dla tekstu *have maintained* klasyfikuje wyrażenie jako wyrażenie czasownikowe i wygląda w sposób następujący:

«phrase id=“12”, from=“42”, to=“44”, type=“VG”»

W warstwie części zdań (*clauses*) model anotacji zawiera semantykę zdania. W przedstawionym modelu pojęcie *części zdania* nie jest tożsame ze znaczeniem w polskiej gramatyce. Oznacza funkcję gramatyczną grupy wyrażeń. Grupuje więc wyrażenia przy pomocy ich funkcji gramatycznych. W praktyce wyraża np. zdanie podrzędne lub zdanie nadrzędne. Wyrażenia mają ścisłe

Wyrażenie	Dopuszczalny typ
AP	PREDICATIVE_AP
AdvP	PREDICATIVE_ADVP
NP	SUBJ
	SUBJ/DEEP_OBJ
	AKK_OBJ
	DAT_OBJ
	GEN_OBJ
	NP_ADJUNCT_GEN
	PREDICATIVE_NP
PP	PP_ADJUNCT
	PP_OBJ
SUBORD_CL	XADJUNCT
	XCOMP
VG	—
W	—

Tabela 4.5: Dopuszczalne typy funkcji gramatycznych dla wyrażeń w domyślnym formacie anotacji

ograniczenia wywodzące się z gramatyki danego języka dotyczące możliwych typów funkcji gramatycznych. Tabela 4.5 przedstawia zestawienie wyrażeń oraz możliwych typów funkcji gramatycznych zgodnie z notacją przyjętą w Declerck (2002) i zastosowaną w domyślnym formacie anotacji.

Element *clause* zawiera następujące właściwości:

- id — unikatowy identyfikator elementu,
- from — identyfikator wyrażenia początkowego,
- to — identyfikator wyrażenia końcowego,
- predicate — identyfikator wyrażenia będącego orzeczeniem,
- arg — lista argumentów w części zdania; każdy z argumentów posiada: *id*, *identyfikator wyrażenia* oraz *typ wyrażenia* jak w tabeli 4.5.

Przykładowe element *clause* wygląda w sposób następujący:

> [PP_ADJUNCT: In an effort] [XADJUNCT: to retain backword-compatibility with earlier editions of KMi-News,] [SUBJ: I] [pred: have maintained] [DOBJ: links] [PP_OBJ: to the older KMi News Page].

Rozdział 5

Ekstrakcja terminologii

Metody ekstrakcji terminologii umownie dzielą się na metody nadzorowane oraz metody bez nadzoru. Umowność wynika z różnic w pojmowaniu granicy pomiędzy nadzorem, czyli interwencją użytkownika, a automatyzmem metody. Przez metody nadzorowane rozumie się najczęściej podejścia, w których użytkownik ręcznie buduje zbiór reguł gramatycznych dla danego języka naturalnego, danej dziedziny oraz typu ekstrakcji[1]. Metody nienadzorowane polegają na automatycznej konstrukcji modeli języka lub modeli statystycznych. Do najbardziej popularnych należą podejścia oparte na modelach statystycznych, które posługują się miarami częstości występowania określonych wyrazów w tekście. Konstrukcja modeli językowych oparta jest na modelach stochastycznych, które szacują miarę prawdopodobieństwa wystąpienia określonych stanów w tekście na podstawie informacji lingwistycznych.

Klasyczne metody nienadzorowane oparte są bezpośrednio na wyrazach lub klasyfikacji POS i wykorzystują model Markova (Manning i Schutze, 1999). Oznacza to, że klasyfikacja wyrazu zależy wyłącznie od wyrazów go poprzedzających. Chronologicznie pierwszą pracą wykorzystującą model Markova na podstawie klasyfikacji POS była praca przedstawiona w Brown i in. (1992).

Opracowana w ramach metamodelu (równanie 3.1) metoda ekstrakcji terminologii γ (równanie 3.3) jest metodą wykorzystującą modele lingwistyczne, dlatego została sklasyfikowana jako metoda nienadzorowana. Ze względu na charakterystykę podejścia metoda nazwana została metodą ekstrakcji terminologii wykorzystującą zmienne okno kontekstowe. Opracowana metoda ekstrakcji terminologii opiera się na następujących założeniach wynikających z obserwacji tekstu:

[1]Drugim często spotykanym rozróżnieniem jest konieczność trenowania modelu, przy czym ręczne budowanie reguł uznawane jest również jako trenowanie.

1. Klasyfikacja terminologii jest zależna od bezpośredniego kontekstu analizowanego wyrazu, bądź grupy wyrazów.

2. Na podstawie kontekstu, analizowanego wyrazu oraz warstwy części mowy tworzone jest tzw. okno kontekstowe.

3. Rozmiar kontekstu wyrazu może być dowolny, co oznacza, że pojemność informacyjna dla funkcji klasyfikacji wyrazów znajdujących się w kontekście jest zmienna.

4. Klasyfikacja przebiega na poziomie dowolnego łańcucha wyrazów bezpośrednio po sobie następujących, co gwarantuje przetwarzanie terminów wieloczłonowych.

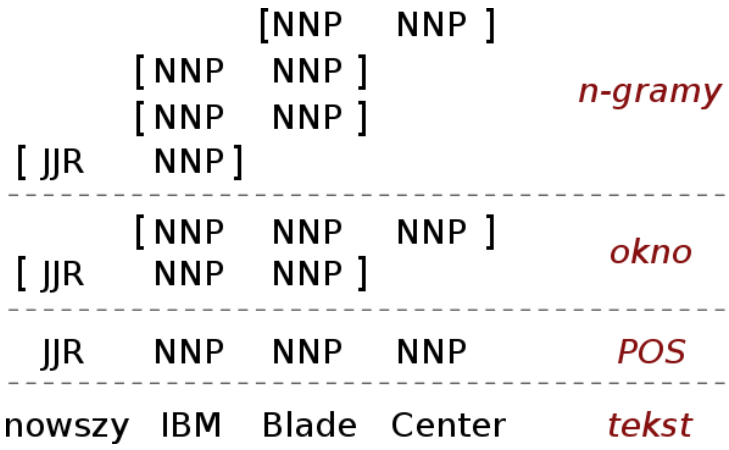
5. Wykorzystuje się wyłącznie podstawową informację lingwistyczną dotyczącą morfologii oraz części mowy.

Zależność od kontekstu jest kluczowym elementem przedstawionego modelu. Założenie to wynika z obserwacji, że obecność terminu w zdaniu może być spowodowana nie tylko wyrazem, bądź ciągiem wyrazów bezpośrednio go poprzedzających, ale również po nim następujących.

Obserwacja tekstu pozwala również postawić drugą hipotezę: pojemność informacyjna kontekstu dla funkcji klasyfikacji terminu jest zróżnicowana. Oznacza to, że wyrazy znajdujące się w analizowanym kontekście, mogą mieć różne znaczenie dla funkcji klasyfikacji. Nie zawsze warto analizować cały kontekst, ponieważ traci się częstość występowania określonych typów reguł zachowania się tekstu. Z kolei kontekst zbyt mały może dostarczyć zbyt mało informacji. Zarządzanie rozmiarem kontekstu powoduje problem jego optymalizacji. Przedstawiona metoda tworzy reguły oparte na dowolnej strukturze kontekstu w zależności od ich wartości statystycznych.

Ekstrakcja terminów wieloczłonowych oparta została w podobny sposób na analizie kontekstu, z tym wyjątkiem, że kontekst jest oparty na dowolnej sekwencji wyrazów oznaczających termin. Zatem, w przypadku analizy terminu wieloczłonowego całość łańcucha na niego się składającego traktowana jest jako potencjalny termin i dopiero na takiej sekwencji wyrazów budowany jest kontekst. Podejście takie umożliwia jednorodne traktowanie terminów jednoczłonowych i wieloczłonowych.

Zasadniczą różnicą jest wprowadzenie modelu opartego na oknie kontekstowym, a nie bezpośrednio na n-gramach. Warstwy modelu przedstawione są na rysunku 5.1. Dotychczasowe metody ograniczały się do budowania jednostronnych n-gramów bezpośrednio opartych na tekście lub znacznikach POS.



Rysunek 5.1: Przykładowe wyrażenie "nowszy IBM Blade Center" w rozkładzie na warstwy modelu, tj. tekst, części mowy (POS), okno kontekstowe oraz poszczególne n-gramy (w tym przypadku bigramy)

5.1 Budowa modelu

Pierwszym krokiem opracowanej metody jest konstrukcja modelu języka. Model budowany jest w celu wykorzystania go w fazie właściwej ekstrakcji terminologii z tekstu. Budowa modelu jest całkowicie zautomatyzowana, a jej wynikiem jest zbiór reguł ekstrakcji terminologii. Konstrukcja modelu składa się z szeregu następujących po sobie faz omówionych w kolejnych sekcjach.

5.1.1 Modele prawdopodobieństwa

Ekstrakcja terminologii jest funkcją prawdopodobieństwa (rozkładem prawdopodobieństwa) $P : F \rightarrow [0, 1]$ według aksjomatycznej definicji Kołmogorowa, ponieważ:

- $P(A_i) > 0$
- $P(\Omega) = 1$
- $P(\bigcup_{i=1}^{2} A_i) = \sum_{i=1}^{2} P(A_i)$

  gdzie:
  A_1 oznacza zdarzenie, w którym wyrażenie jest terminem,
  A_2 oznacza zdarzenie, w którym wyrażenie nie jest terminem,
  A_1 i A_2 są zdarzeniami rozłącznymi.

W tworzeniu modeli lingwistycznych dla ekstrakcji terminologii nie używa się jednak klasycznego modelu prawdopodobieństwa, ponieważ przeważnie posiada się częściową wiedzę na temat wyników próby. Wiadomo np., że analizowany wyraz należy do określonej klasy części mowy, wiadomo również, że znajduje się w pewnym kontekście, znany jest w końcu jego symbol. Dlatego w przypadku konstrukcji modeli dla ekstrakcji terminologii mowa jest o prawdopodobieństwie warunkowym, tj.:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{, dla } P(B) > 0. \tag{5.1}$$

Z powyższego równania oraz z zachodzącej relacji symetryczności pomiędzy zbiorami A i B ($A \cap B = B \cap A$) wynika, że:

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A). \tag{5.2}$$

Uogólnienie tego równania dla n zdarzeń jest podstawową zasadą stosowaną przy budowie modelu (tzw. *reguła łańcuchowa*):

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \ldots P(A_n| \cap_{i=1}^{n-1} A_i). \tag{5.3}$$

W wielu przypadkach wartość $P(A|B)$ jest trudna to oszacowania, podczas gdy wartość $P(B|A)$ jest albo dana, albo jej szacunek jest dużo prostszy. W takich przypadkach stosuje się twierdzenie Bayesa, które w postaci uproszczonej wynika bezpośrednio z równań 5.1 oraz 5.2:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}. \tag{5.4}$$

Uproszczona wersja twierdzenia Bayesa zakłada, że $P(A)$ jest znane. Nie zawsze jest to jednak łatwe, dlatego można oszacować wartość $P(A)$ przy pomocy równania 5.2. Zakładając, że zdarzenie $\overline{B}$ jest dopełnieniem zdarzenia B, czyli $B \cap \overline{B} = \oslash$ oraz $B \cup \overline{B} = \Omega$:

$$P(A) = P(A \cap B) + P(A \cap \overline{B}) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B}). \tag{5.5}$$

Uogólniając powyższe równanie do dowolnej liczby zdarzeń dzielących zdarzenie A, takich że $A \subseteq \cap_i B_i$ oraz B_i są rozłączne, stwierdzamy że:

$$P(A) = \sum_i P(A|B_i)P(B_i), \tag{5.6}$$

a to z kolei daje ostateczny kształt pełnemu twierdzeniu Bayesa:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i^n P(A|B_i)P(B_i)}. \tag{5.7}$$

gdzie:
$A \subseteq \cap_i^n B_i$, $P(A) > 0$, $B_i \cap B_j = \oslash \quad$ dla $i \neq j$.

W modelach opartych na teorii prawdopodobieństwa wysoce niepraktycznie jest mówić o konkretnych zdarzeniach, zwłaszcza w przypadku nieregularnych przestrzeni zdarzeń, które różnią się w zależności od stanu modelu. Zamiast więc mówić o grupie zdarzeń, bardziej praktycznie jest mówić o pewnych konkretnych wartościach, które te zdarzenia generują. Dla przykładu rzut sześcienną kostką dwa razy wygodniej jest określać prawdopodobieństwem wyrzucenia liczby 10, niż prawdopodobieństwem sumy wszystkich trzech składających się na to zdarzeń (tj. 4 i 6, 5 i 5, 6 i 4). Podobnie w przypadku ekstrakcji terminologii interesuje nas prawdopodobieństwo czy dane wyrażenie jest terminem czy nie. W tym stwierdzeniu niejako abstrahujemy od konkretnego stanu modelu, czyli wszystkich możliwych kombinacji zdarzeń mających na to wpływ.

Dlatego niezwykle pomocnym pojęciem w ekstrakcji terminologii jest *zmienna losowa*, która jest funkcją odwzorowującą zbiór Ω w zbiór liczb rzeczywistych, tj: $X : \Omega \rightarrow R^n$. W przypadku ekstrakcji terminologii mamy do czynienia z dyskretną zmienną losową, czyli funkcją: $X : \Omega \rightarrow S$, gdzie S jest policzalnym podzbiorem R, a konkretniej z funkcją: $X : \Omega \rightarrow \{0, 1\}$, gdzie 0 oznacza, że wyrażenie nie jest terminem, a 1 oznacza, że wyrażenie jest terminem. Taką postać zmiennej losowej nazywa się *wskaźnikową funkcją losową* lub *próbą Bernoullego*.

5.1.2 Modele Markova

W ekstrakcji terminologii nie wystarczy, że dokona się analizy konkretnej wartości zmiennej losowej. Analiza terminologii polega na sekwencyjnej analizie następujących po sobie zdań, wyrażeń, części mowy lub jakiejkolwiek charakterystyce lingwistycznej, która zostanie uznana za odpowiednio dyskryminującą. Dlatego dla celów ekstrakcji terminologii mowa jest o analizie sekwencji następujących po sobie zmiennych losowych. Sekwencja ta charakteryzuje się ponadto ograniczoną zależnością pomiędzy występującymi po sobie zmiennymi losowymi. Oznacza to, że konkretna zmienna losowa nie jest zależna od wszystkich poprzedzających wartości w sekwencji, lecz najczęściej tylko od zmiennej losowej bezpośrednio ją poprzedzającej.

Zakładając że $X = (X_1, \ldots, X_n)$ jest sekwencją zmiennych losowych przybierającą wartości w skończonym zbiorze przestrzeni stanów $S = s_1, \ldots, s_k$, *założenie Markova* lub *własność Markova* mówi, że:

Skończony horyzont zależności:

$$P(X_{n+1} = s_k | X_1, \ldots, X_n) = P(X_{n+1} = s_k | X_n). \tag{5.8}$$

Model niezmienny w czasie (stacjonarny):

$$= P(X_2 = s_k|X_1). \tag{5.9}$$

Równanie 5.8 własności Markova mówi o tym, że konkretna zmienna losowa jest zależna wyłącznie od zmiennej losowej bezpośrednio ją poprzedzającej. Równanie 5.9 mówi natomiast o niezmiennych charakterze modelu, czyli o tym, że parametry modelu nie ulegają zmianie. Oba te założenia są bardzo restrykcyjne, ponieważ w bezpośrednim przełożeniu zmiennych losowych na wyrazy występujące w tekście klasyfikacja terminów zależy od więcej niż tylko jednego poprzedzającego wyrazu. Własność tę jednak można łatwo uzyskać odpowiednią konstrukcją zmiennych losowych nie jako poszczególne wyrazy, lecz odpowiednie wyrażenia. Dalsza część tej dyskusji nastąpi w sekcji następnej, przy okazji omawiania budowym modeli n-gram. Druga z własności Markova jest w analizie lingwistycznej do zaakceptowania, ponieważ konstrukcje użyte w języku (nie samo słownictwo) zmienia się na tyle rzadko, że zmiany te można pominąć.

Modelem lub łańcuchem Markova nazywa się sekwencję zmiennych losowych, która spełnia własność Markova.

W analizie języka wyróżnia się dwa typy modeli Markova: widoczne modele Markova oraz ukryte modele Markova. Omawiane dotąd modele to widoczne modele Markova. Ukryte modele Markova różnią się jedynie istnieniem dodatkowej tzw. ukrytej warstwy, czyli zbioru parametrów, które nie są znane, a na ich podstawie musi zostać oszacowana zmienna losowa. Widoczny jest jedynie efekt ukrytej warstwy. W przypadku analizy języka, ukryte modele Markova mogą być użyte np. wtedy, gdy znana jest klasa części mowy (np. poprzez analizę składni wyrażenia), a na jej podstawie należy oszacować wyraz. W przypadku jednak ekstrakcji terminologii oraz istniejącej informacji lingwistycznej, ukrytej warstwy po prostu nie ma. Z tego powodu zastosowanie ukrytych modeli Markova nie zostało zbadane.

5.1.3 Model n-gram

Fundamentalnym problemem przy budowaniu modeli opartych na wnioskowaniu statystycznym jest odpowiedni dobór cech dyskryminujących. Problem polega na konieczności kompromisu pomiędzy przydatnością a stopniem dyskryminacji modelu. Na przykład w klasycznym zadaniu predykcji następnego wyrazu w zdaniu, klasy dyskryminujące można zbudować w oparciu na 10. poprzedzających wyrazach. Zdolność dyskryminacji takiego modelu spowoduje, że jeśli kiedykolwiek pojawi się taka sama sekwencja 10 wyrazów, to z dużym prawdopodobieństwem model sprawdzi się. Problem jednak w tym, że prawdopodobieństwo wystąpienia takiej sekwencji jest bardzo

niskie. Dzieląc zatem zbiór wszystkich możliwych sekwencji w ten sposób uzyskany model jest tyleż dyskryminujący, co zupełnie nieprzydatny.

Klasyczne zadanie predykcji następnego wyrazu w zdaniu jest funkcją prawdopodobieństwa:

$$P(w_n|w_1, \ldots, w_{n-1}). \tag{5.10}$$

W opracowanym modelu ekstrakcji terminologii zadanie klasyfikacji terminu dla wyrazu lub sekwencji wyrazów w_n, jest funkcją prawdopodobieństwa warunkowego opartą na częściach mowy (pos_n) dla poszczególnych wyrazów lub sekwencji wyrazów:

$$P(X) = P(term_k|pos_1, \ldots, pos_n), \tag{5.11}$$

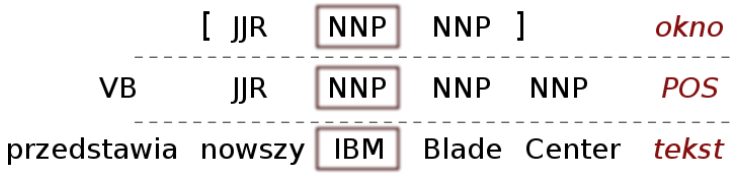
gdzie:
n oznacza liczbę węzłow modelu n-gram,
k oznacza pozycję węzła oznaczonego jako termin,
$1 \leqslant k \leqslant n$,
term oznacza węzeł oznaczony jako termin,
pos oznacza węzeł reprezentowany poprzez znacznik części mowy.

Ponadto opracowany model zakłada własność Markova (równanie 5.8 lub 5.9) w celu rozwiązania problemu konstrukcji odpowiednich klas dyskryminacyjnych. Założenie, że stan konkretnej zmiennej losowej zależy od wszystkich ją poprzedzających w analizowanej sekwencji (czyli np. dokumencie), jest może dopuszczalne, ale prowadzi do problemu nadmiernej dyskryminacji. Z tego powodu założono, że zmienna losowa zależy wyłącznie od wartości poprzedniej. Drugie założenie modelu Markova jest również prawdziwe, ponieważ w analizowanym okresie model jest niezmienny.

Model n-gram jest niczym innym jak modelem Markova stopnia $n - 1$. Stopień obu modeli oznacza długość analizowanej sekwencji, tzn. jeżeli analizowane są dwie zmienne losowe w sekwencji, mowa jest o modelu Markova stopnia 1. (jeden stan poprzedzający) lub modelu n-gram opartym na gramach składających się z 2 elementów. Notacja modeli n-gram jest na tyle popularna, że sekwencje zmiennych losowych modelu n-gram stopnia 1. nazywa się unigramami, stopnia 2. bigramami, stopnia 3. trigramami, itd.

5.1.4 Budowa okna

Z analizowanego tekstu dla każdego klasyfikowanego wyrazu pozyskiwane jest tzw. okno kontekstowe. Okno kontekstowe jest symetryczne w stosunku do terminu i składa się z parzystej liczby wyrazów kontekstowych plus termin. Termin jest zawsze elementem środkowym. Konstrukcja okna kontekstowego dla przykładowego wyrażenia przedstawiona jest na rysunku 5.2.



Rysunek 5.2: Budowa okna o rozmiarze 3 dla terminu KMi w wyrażeniu "przedstawia nowszy IBM Blade Center"

W związku z tym, że metoda umożliwia ekstrakcję terminów wieloczłonowych, sam termin może składać się z wielu wyrazów. Sposób reprezentacji oraz budowa okna nie zmienia się w porównaniu z rysunkiem 5.2, z wyjątkiem sposobu reprezentacji samego terminu (rysunek 5.3). Sposób reprezentacji kontekstu terminu nie ulega również zmianie, z wyjątkiem przesunięcia się okna o liczbę wyrazów równą liczbie wyrazów terminu wieloczłonowego - 1. Rozmiar okna jest więc funkcją:

$$f(n, size_{term}) = 2n + size_{term}, \tag{5.12}$$

gdzie:
n jest liczbą tokenów kontekstowych przylegających do terminu (z każdej strony),
$size_{term}$ oznacza rozmiar terminu, czyli liczbę wyrazów (tokenów).

Budowa okna kontekstowego umożliwia analizę wyrazów występujących zarówno przed analizowanym wyrazem, jak i po nim.

5.1.5 Budowa n-gramów

Zgodnie z równaniem 5.11 oraz jego rozwinięciem (równanie 5.13) n-gramy muszą być zbudowane dla dwóch typów sekwencji:

- $pos_1, \ldots, pos_n$,
- $term_k|pos_1, \ldots, pos_n$.

W powyższych sekwencjach n jest zarówno liczbą węzłów w n-gramach, co wynika z równania 5.11, jak również rozmiarem okna kontekstowego, co wynika z budowy modelu.

O ile pierwszy typ sekwencji jest standardowy, drugi przypadek niesie ze sobą wiele problemów. Wynikają one z tego, że sekwencja zawiera dwa poziomy właściwości analizowanej próby. Oprócz znaczników części mowy

pojawiają się wskazania eksperta związane z klasyfikacją terminów. Chcąc zachować jednorodność postaci, należy w taki sposób zbudować n-gramy, aby można było w jednorodny sposób traktować oba te poziomy.

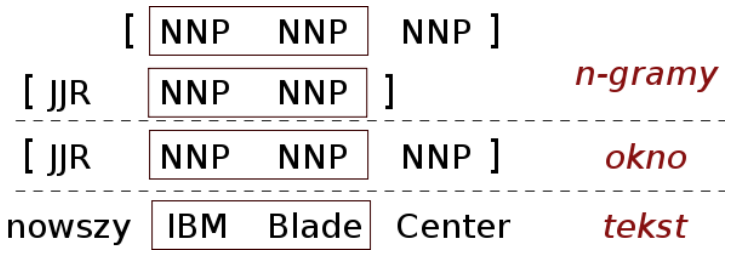
Zasadniczą właściwością budowanych n-gramów jest to, że są one oparte nie na wyrazach, lecz znacznikach części mowy. Oparcie n-gramów na wyrazach doprowadziłoby do problemów wynikających ze zbyt dużej dyskryminacji. Zastosowanie znaczników POS daje zamknięty zbiór cech reprezentujących tekst.

Istnieją dwie zasadnicze metody kontrukcji modelu n-gram dla danego okna kontekstowego. Model może być oparty na n-gramach jednorodnych lub niejednorodnych. Jednorodność oznacza ten sam stopień n-gramów. W przypadku modelu opartego na n-gramach jednorodnych powstaną n-gramy określonego poziomu. W przypadku modelu opartego na n-gramach niejednorodnych powstaną n-gramy od stopnia pierwszego (unigramy), aż po stopień równy rozmiarowi okna kontekstowego (zgodnie ze wzorem 5.12). W obu przypadkach problemem jest wybór odpowiedniego poziomu n-gramów. W modelu jednorodnym problem ogranicza się do stwierdzenia poziomu w momencie konfigurowania modelu. Dla różnych poziomów eksperymenty można powtarzać, aż do uzyskania najlepszych wyników. W modelu niejednorodnym problem przesuwa się do etapu ważenia wyników uzyskanych z n-gramów o różnych poziomach. W obu przypadkach należy parametr ten optymalizować. Do eksperymentów wykorzystano model oparty na n-gramach jednorodnych. Oznacza to, że dla okna kontekstowego o rozmiarze n powstają n-gramy poziomu *n-1*.

W związku z konstrukcją okna kontekstowego n-gramy posiadają różne umiejscowienie analizowanego terminu. W przypadku ogólnym termin może wystąpić na każdej pozycji w n-gramie. Dla okna kontekstowego o rozmiarze 3 powstaną dwa bigramy, jeden z terminem na drugiej pozycji, drugi z terminem na pozycji pierwszej.

W przypadku tworzenia modelu jednorodnego istnieje jeszcze jedna możliwość postępowania. Dla zadanego okna kontekstowego można stworzyć wiele modelów jednorodnych, a następnie zintegrować uzyskane wyniki, czyli prawdopodobieństwo klasyfikacji terminu. Stopień modelu n-gram, który zostanie uznany za dający wystarczające wyniki, może być dobierany dynamicznie, a wraz z nim rozmiar analizowanego okna kontekstowego. Stąd odmiana ta może zostać nazwana metodą dynamicznego okna kontekstowego.

W przedstawionym modelu wykorzystano n-gramy jednorodne z możliwością konfiguracji rozmiaru okna. Wynik budowy n-gramów dla przykładowego wyrażenia, w którym terminem jest wyrażenie *IBM Blade*, przedstawiono na rysunku 5.3. W wyniku analizy wyrażenia powstało okno trzy-elementowe



Rysunek 5.3: Budowa n-gramów dla okna kontekstowego o rozmiarze 3 oraz terminu wieloczłonowego *IBM Blade* w wyrażeniu "nowszy IBM Blade Center"

rozmiar okna	liczba wyrazów kontekst	max modelu n-gram	stopień modelu Markova	stopień n-gramów	n-gramy
1	0	1	0	1	pos_k
3	1	3	1	2	pos_{k-1}, pos_k
					pos_k, pos_{k+1}
5	2	5	2	3	$pos_{k-2}, pos_{k-1}, pos_k$
					$pos_{k-1}, pos_k, pos_{k+1}$
					$pos_k, pos_{k+1}, pos_{k+2}$

Tabela 5.1: Zależności pomiędzy podstawowymi właściwościami modelu opartego na oknie kontekstowym

(termin w całości stanowi jeden węzeł), a następnie dwa bigramy (ponownie termin jest traktowany jako jeden węzeł).

Ogólne zależności pomiędzy rozmiarem okna, poziomem n-gramów oraz modelu Markova, a także liczbą i strukturą n-gramów zostały przedstawione w tabeli 5.1.

5.1.6 Reprezentacja terminów wieloczłonowych

W przypadku terminów wieloczłonowych pojawia się problem ich reprezentacji w modelu. Reprezentacja terminu wieloczłonowego w modelu wymaga dodania informacji o statusie węzła modelu jako termin, tzw. flagi terminu oraz reprezentacji całości terminu. Operacja ta powoduje potrzebę zmiany sposobu reprezentacji modelu.

Jeśli węzłem w n-gramie nazywać będziemy najmniejszą jednostkę analizy w modelu n-gram, to termin wieloczłonowy można przedstawić na dwa sposoby:

- każdy token w terminie jako osobny węzeł lub
- cały termin jako jeden węzeł.

Przykładową sekwencję "W celu analiz ." można przedstawić przy pomocy modelu POS n-gram dla przypadku pierwszego jako:

$$IN, [term : NN], [term : NNS]$$

oraz dla przypadku drugiego:

$$IN, [term : NN, NNS], PUNCT$$

Pierwszy przypadek reprezentacji terminów wieloczłonowych upodabnia model do reprezentacji jednoczłonowych terminów, a przez to nie powoduje konieczności jego przebudowy. Drugi przypadek jest jednak bardziej intuicyjny, ponieważ:

- dokładniej odzwierciedla klasyfikację eksperta — to przecież sekwencja została oznaczona jako termin, a nie poszczególne wyrazy,
- nie powoduje zniekształcenia okna kontekstowego — po obu stronach terminu wieloczłonowego pozostaje taka sama liczba wyrazów kontekstowych,
- nie powoduje konieczności analizy długości samego terminu, a przez to upraszcza model — niezależnie od długości terminu pozostaje on tylko pojedynczym węzłem w modelu.

Ze wskazanych powodów dla przedstawionej metody wybrany został drugi przypadek reprezentacji terminów wieloczłonowych, tj. reprezentacja całego terminu jako jeden węzęł w modelu n-gram. Niestety, wybór ten implikuje, że porównanie elementów n-gramów nie jest oczywiste, na przykład:

$$[term : NN, NN] \neq [term : NN] \neq [term : NN], NN.$$

Dlatego dla porównywania węzła n-gramu oznaczonego jako termin należy wziąć pod uwagę długość terminu oraz znacznik POS z każdej jego pozycji.

5.1.7 Nazwy własne i byty nazwane

Sposób reprezentacji bytów nazwanych (ang. *Named Entities*) odbywa się przeważnie na jeden z dwóch sposobów. Po pierwsze, poprzez wydzielone elementy w strukturze pliku. Po drugie, poprzez odpowiednią klasę znaczników POS.

Reprezentacja poprzez wydzielone elementy jest niezależna od informacji morfosyntaktycznych. Jest to niezaprzeczalna zaleta w przypadku ekstrakcji informacji, gdzie źródłowa informacja morfosyntaktyczna może być przydatna na wielu etapach analizy. Z punktu widzenia ekstrakcji terminologii taka reprezentacja jest jednak nadmiarowa, ponadto komplikuje strukturę pliku.

Reprezentacja poprzez znaczniki części mowy polega na istnieniu klasy POS, która wskazuje na nazwy własne. Nazwy własne nie są jednak równoznaczne z bytami nazwanymi. W ekstrakcji informacji przyjęło się, że nazwy własne tylko wskazują na byty nazwane (Appelt i Israel, 1999). Różnica ta jest podobna do relacji pomiędzy terminem a pojęciem, ponieważ nazwa własna jest symbolem potencjalnie wskazującym na byt nazwany. Oczywiście w większości przypadków nazwa własna oznacza również byt nazwany i dla celów innych dyscyplin niż ekstrakcja informacji, założenie to jest powszechnie przyjmowane. Wynika to choćby z prostej przyczyny, że zbiory znaczników zawierają właśnie nazwy własne. Reprezentacja bytów nazwanych poprzez znaczniki POS jest znacznie prostsza w przetwarzaniu, ponieważ nazwa własna jest po prostu jedną z klas części mowy.

Przesądzającym argumentem na korzyść wykorzystania reprezentacji bytów nazwanych poprzez znaczniki POS jest fakt, że wszystkie popularne klasyfikatory POS znakują tekst do postaci nazw własnych. Wykorzystanie tej formy reprezentacji jest więc bardziej rozpowszechnione i zostało zastosowane w przedstawianej metodzie.

5.2 Szacowanie wartości modelu

W celu szacowania wartości modelu stosuje się estymatory. Zakładając reprezentatywność próby estymatory szacują wartości rozkładu prawdopodobieństwa populacji generalnej, czyli w tym przypadku modelu języka.

Zgodnie ze wzorem 5.11, prawdopodobieństwo wystąpienia terminu zgodnie z przedstawianym modelem wynosi:

$$P(term_k|pos_1,\ldots,pos_n) = \frac{P(pos_1,\ldots,pos_n,term_k)}{P(pos_1,\ldots,pos_n)}. \tag{5.13}$$

Prawdopodobieństwo terminu jest stosunkiem prawdopodobieństwa, że sekwencja znaczników $POS = pos_1,\ldots,pos_n$ zawiera termin na pozycji k, gdzie $1 \leqslant k \leqslant n$ do prawdopodobieństwa wystąpienia sekwencji $POS = pos_1,\ldots,pos_n$.

5.2.1 Metoda największej wiarygodności

Dla oszacowania wartości opracowanej metody ekstrakcji terminologii zastosowano estymator punktowy metodą największej wiarygodności (MLE). Zakładając, że $C(pos_1, \ldots, pos_n)$ oznacza częstość występowania danego n-gramu oraz N liczbę wszystkich n-gramów danego poziomu, MLE mówi, że:

$$P_{MLE}(pos_1, \ldots, pos_n, term_k) = \frac{C(pos_1, \ldots, pos_n, term_k)}{N}, \tag{5.14}$$

$$P_{MLE}(pos_1, \ldots, pos_n) = \frac{C(pos_1, \ldots, pos_n)}{N}, \tag{5.15}$$

stąd wynika, że prawdopodobieństwo wystąpienia terminu metodą największej wiarygodności wynosi:

$$P_{MLE}(term_k|pos_1, \ldots, pos_n) = \frac{C(pos_1, \ldots, pos_n, term_k)}{C(pos_1, \ldots, pos_n)}. \tag{5.16}$$

Metoda największej wiarygodności to najczęściej stosowana metoda w budowaniu modeli języka, przede wszystkim z racji swojej prostoty i relatywnie dobrej skuteczności. Posiada jednak fundamentalną wadę, która dyskwalifikuje ją z szeregu zastosowań, np. ze wszystkich modeli opartych na wyrazach. Metoda MLE jest podatna na tzw. problem rzadkości. Na przykład budując model służący do przewidywania następnego wyrazu na podstawie poprzedzających wyrazów, uzyskuje się model o wysokim stopniu dyskryminacji. Oznacza to, że n-gramy, które zostały stworzone podczas trenowania modelu będą dobrze klasyfikować znane przypadki. Problem jednak w tym, że wybór pomiędzy stopniem dyskryminacji a reprezentatywnością modelu jest klasycznym kompromisem. Dlatego stosując duży stopień dyskryminacji, model nie obejmuje wszystkich przypadków. Oznacza to, że relatywnie często model oparty na wyrazach będzie napotykał przypadki nieznane, tj. nieobjęte korpusem treningowym. Metoda największej wiarygodności przypisuje takim przypadkom wartość prawdopodobieństwa równą 0. Niespotkane kombinacje wyrazów pozostają więc dla modelu nie do rozpoznania.

Rozwiązaniem problemu nieznanych przypadków jest próba grupowania występujących wyrazów. Takim grupowaniem jest np. sprowadzanie wyrazów do formy podstawowej. Istotą przedstawionej metody jest to, że n-gramy zostały zbudowane na znacznikach POS, które charakteryzują się skończoną i względnie małą liczbą postaci. Właśnie ta cecha powoduje, że pomimo świadomości wad metody MLE, zastosowano ją w przedstawionym modelu.

Zgodnie z równaniem 5.16 należy więc dla każdego analizowanego wyrazu policzyć liczbę wystąpień n-gramów, w których wystąpił termin oraz liczbę wystąpień n-gramów, bez względu na wynik klasyfikacji eksperta.

5.2.2 Reprezentacja reguł

Wynikiem modelu jest zbiór reguł. Każda reguła posiada postać szacunku prawdopodobieństwa, że analizowane wyraz lub grupa wyrazów jest terminem. Reguła obejmuje n-gram o zadanym poziomie. Zgodnie z równaniem 5.16 należy policzyć dwie wartości:

- liczbę wystąpień n-gramów, w których wystąpił termin,
- liczbę wystąpień n-gramów,

przy czym n-gramy z obu wartości, z wyjątkiem obecności terminu, muszą być równe. Równość n-gramów bez klasyfikacji terminów oznacza, że:

- długość obu n-gramów (liczba wyrazów opisanych przy pomocy znaczników POS) jest równa,
- w obu n-gramach występują te same znaczniki POS,
- w obu n-gramach występuje ta sama kolejność występowania znaczników POS.

Liczbę wystąpień n-gramów bez klasyfikacji terminów można policzyć klasycznym zadaniem wyszukania wzorca w tekście (Baeza-Yates i Ribeiro-Neto, 1999). Przedstawiona metoda wykorzystuje algorytm Aho-Corasick (Aho i Corasick, 1975) w wersji operującej na bajtach. Zastosowanie tego algorytmu powoduje uzyskanie liniowej złożoności algorytmicznej.

Liczba wystąpień n-gramów, w których wystąpił termin pozyskiwana jest ze zbudowanego modelu. Ze względu na dużą dowolność w konstrukcji samego n-gramu oraz pozycji terminu przy sumowaniu n-gramów zastosowano następującą miarę równości n-gramów (tym razem z klasyfikacją terminu):

- długość obu n-gramów jest równa,
- występują te same znaczniki POS,
- występuje ta sama kolejność występowania znaczników POS,
- termin występuje na tej samej pozycji,
- termin obejmuje te same znaczniki POS.

Reprezentacja reguł następuje poprzez połączenie obu tych wielkości. Na przykład liczba wystąpień n-gramów, w których wystąpił termin może przybrać formę:

$$[NN, term : NNS, PUNCT] = 4,$$

natomiast liczba wystąpień n-gramów:

$$[NN, NNS, PUNCT] = 16.$$

Korzystając ze wzoru 5.16 prawdopodobieństwo, że w analizowanej sekwencji $NN, NNS, PUNCT$ terminem jest wyraz reprezentowane poprzez znacznik NNS wynosi:

$$P_{MLE}(term_2|pos_1, pos_2, pos_3) = \frac{C(pos_1, pos_2(term), pos_3)}{C(pos_1, pos_2, pos_3)} = \frac{4}{16} = 0,25. \tag{5.17}$$

5.3 Ekstrakcja terminologii z wykorzystaniem modelu

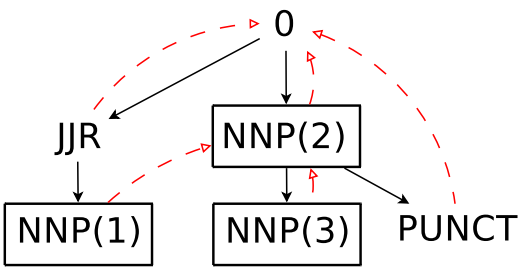
Ekstrakcja terminologii w przedstawionej metodzie składa się z trzech kroków:

1. Budowa modelu (sekcja 5.1).
2. Szacowanie wartości modelu (sekcja 5.2).
3. Ekstrakcja właściwa.

Ekstrakcja właściwa polega na zbudowaniu reprezentacji tekstu podatnej na wyszukiwanie węzłów w postaci n-gramów. Podatność oznacza wrażliwość mechanizmu wyszukiwawczego na specyficzne cechy przedstawionego modelu n-gram.

Punktem wyjścia dla opracowania mechanizmu ekstrakcji był algorytm Aho-Corasick (Aho i Corasick, 1975). Algorytm ten dostosowano do potrzeb reprezentacji węzłów w drzewie poprzez następujące modyfikacje:

1. Zmiana reprezentacji węzłów w drzewie z liter na węzły zawierające anotację lingwistyczną: wyraz, znacznik POS, flagę terminu.
2. Stany końcowe drzewa nie muszą zawierać wyników wyszukiwania — termin nie zawsze jest ostatni w n-gramie ($1 \leqslant k \leqslant n$ zgodnie z równaniem 5.11), co powoduje konieczność dodatkowej nawigacji po drzewie (oprócz funkcji przejścia) do rezultatów wyszukiwania. Stan końcowy jest tym samym tylko wyzwalaczem dla wyników; same wyniki mogą być w innym węźle.



Rysunek 5.4: Reprezentacja drzewa dla ekstrakcji terminów w wyrażeniu "nowszy IBM Blade Center ."

3. Stany końcowe mogą być wyzwalaczem dla więcej niż jednego wyniku.

4. Terminy wieloczłonowe muszą zachować taką samą formę jak pojedyncze tokeny, co oznacza powielenie mechanizmów z jednego węzła na n węzłów (np. reprezentacja wyników).

5. Stan końcowy dla ostatniego tokena terminu wieloczłonowego musi wyzwalać nawigację i zwrot wszystkich n poprzednich tokenów, gdzie n oznacza długość terminu.

Diagram 5.4 jest ilustracją niektórych z wykorzystanych mechanizmów. Przedstawia drzewo dla przykładowego wyrażenia

nowszy(JJR) IBM(NNP) Blade(NNP) Center(NNP).(PUNCT)

dla którego został zbudowany model oparty na następujących bigramach (przy założeniu, że *IBM* oraz *Center* są terminami):

[JJR, term: NNP],[term: NNP, NNP],[NNP, term: NNP],[term: NNP, PUNCT]

Na diagramie zaznaczono węzły drzewa przy pomocy znaczników POS, normalne przejścia przy pomocy czarnej, jednolitej linii oraz funkcje wyjścia (*fail()*) przy pomocy czerwonej, przerywanej linii. Stany końcowe są liśćmi drzewa. Węzły otoczone prostokątem oznaczają wystąpienie terminu. Przejście przez drzewo dla omawianej przykładowej frazy przedstawia tabela 5.2. Przejście 3. generuje dwa terminy: *IBM* oraz *Center* dla dwóch różnych wzorców. Dodatkowo *IBM* nie jest terminem przyporządkowanym do węzła końcowego NNP(3) tylko do węzła NNP(2).

Ze względu na przejrzystość przedstawiony diagram nie pokazuje reprezentacji terminów wieloczłonowych. Uzupełnienie modelu o przetwarzanie

Lp.	Wyraz	Węzeł(POS)	fail()	Wynik(termin)	Wynik(reguła)
1	nowszy	JJR	0	—	—
2	IBM	NNP(1)	NNP(2)	IBM(NNP1)	[JJR, term:NNP]
3	Blade	NNP(3)	NNP(2)	IBM(NNP2)	[term:NNP, NNP]
				Blade(NNP3)	[NNP, term:NNP]
4	Center	NNP(3)	NNP(2)	Blade(NNP2)	[term:NNP, NNP]
				Center(NNP3)	[NNP, term:NNP]
5	.	PUNCT	0	Center(NNP2)	[term:NNP, PUNCT]

Tabela 5.2: Przejście przez drzewo dla przykładowego wyrażenia

terminów wieloczłonowych wymagało dodatkowo zmiany odniesienia do terminu ze stanu końcowego z pojedynczego węzła na sekwencję węzłów.

Etykiety terminów ze zbioru LA oraz n-gramów nie mogą być dokładnie przeniesione do wyników w postaci nazw terminów. W nazwach terminów stosuje się najczęściej formę podstawową wyrazu. Zastosowane narzędzia do anotacji tekstu obejmują swym zakresem również formę podstawową, czyli lemat. Nie zawsze jednak należy go wykorzystać, np:

- w nazwach własnych lub
- w terminach wieloczłonowych.

Rozwiązanie nazw w n-gramach do nazw terminów wieloczłonowych odbywa się poprzez sprowadzenie ostatniego tokena w n-gramie do formy podstawowej, wszystkie pozostałe pozostawiając bez zmian. Wyjątek stanowi sytuacja, w której ostatni token jest nazwą własną.

Uzyskany w ten sposób model wykorzystano w celu ekstrakcji właściwej terminów.

5.4 Optymalizacja modelu dla dziedziny handlu elektronicznego

W celu uzyskania optymalnych wyników przedstawioną metodę ekstrakcji terminów należy dostosować do specyfiki tekstów z dziedziny handlu elektronicznego. Optymalizacja polega na uzyskaniu jak najwyższej efektywności metody mierzonej miarą F, poprzez ustalenie wartości następujących parametrów:

- rozmiar okna kontekstowego,
- stopień n-gramów w ramach okna,

- wygładzanie MLE (szacowanie przypadków nieznanych),
- metoda ważenia wyników dla terminu (np. prawo i lewostronne wyniki dla okna),
- przetwarzanie terminów wieloczłonowych.

Wzrost rozmiaru okna kontekstowego powoduje rozszerzanie analizowanego kontekstu, a więc m.in. wzrost precyzji analizowanych przypadków oraz wzrost złożoności obliczeniowej. Jednak z drugiej strony oznacza zwiększenie stopnia dyskryminacji modelu. Nadmierna dyskryminacja modelu jest niewskazana, ponieważ powoduje znaczący wzrost liczby przypadków niesklasyfikowanych. Kompromisem w przypadku wyznaczania rozmiaru okna kontekstowego jest taka wartość, która wyróżnia w znaczący sposób analizowane przypadki, jednocześnie nie powodując nadmiernej dyskryminacji. W drodze obserwacji korpusu dla dziedziny handlu elektronicznego oraz uzyskiwanych wyników pośrednich ustalono, że optymalny rozmiar okna kontekstowego dla badanego korpusu wynosi 5. Zgodnie z tabelą 5.1 oznacza to 2 analizowane wyrazy z każdej strony potencjalnego terminu, 3. stopień modelu n-gram, 2. stopień modelu Markova oraz maksymalnie 3. stopień n-gramów.

Kontynuując przykład wyrażenia "przedstawia nowszy IBM Blade Center" (por. rysunki 5.1, 5.2, 5.3 oraz 5.4) zastosowanie takiego rozmiaru okna kontekstowego oznacza budowanie następującej struktury okna w warstwie tekstu:

[przedstawia,nowszy,IBM,Blade,Center],

a to przekłada się na następującą strukturę w warstwie POS:

[VB,JJR,NNP,NNP,NNP].

Wyznaczenie optymalnej wartości rozmiaru okna kontekstowego implikuje maksymalny stopień n-gramów. Wartość tą można jednak optymalizować pod względem jednorodności oraz stopnia z zakresu od 1 do maksymalnej dozwolonej wartości przy danym rozmiarze okna. Jednorodność wskazuje na to, czy konstruowane n-gramy w ramach okna kontekstowego mają być takiego samego stopnia, czy nie. Niejednorodność powoduje, że n-gramy konstruowane są dla wszystkich dozwolonych stopni, czyli od maksymalnej wartości do unigramów. W przypadku jednorodnego modelu należy dokonać optymalizacji stopnia, np. dla wartości rozmiaru okna kontekstowego 5. oraz maksymalnie 3. stopnia n-gramów należy wyznaczyć wartość jednorodnego modelu n-gram wśród wartości: 3, 2, 1. korpusu W wyniku analizy korpusu z dziedziny handlu elektronicznego oraz uzyskiwanych wyników pośrednich zdecydowano się

na model jednorodny z wartością stopnia n-gramów odpowiadającą maksymalnej dozwolonej wartości.

Przy analizie terminu jednotokenowego "IBM", oznacza to, że dla przykładowego okna kontekstowego:

[VB,JJR,term:NNP,NNP,NNP]

powstaną 3 n-gramy jednorodne:

[VB,JJR,term:NNP],
[JJR,term:NNP,NNP],
[term:NNP,NNP,NNP].

Wzrost dyskryminacji modelu powoduje wzrost prawdopodobieństwa wystąpienia sekwencji części mowy nie objętej skonstruowanym modelem. Prawdopodobieństwo, że przypadek taki jest terminem wynosi 0. W celu objęcia przypadków nieznanych stosuje się tzw. wygładzanie modelu, np. metodą Laplace'a, Lidstone'a lub Jeffreys-Perksa (Manning i Schutze, 1999). W przedstawionych eksperymentach nie zastosowano wygładzania, nie zbadano również jego wpływu na model ekstrakcji. Obserwacja uzyskanych wyników dla korpusu z dziedziny handlu elektronicznego dostarcza jednak przesłanek do stwierdzenia, że nie istnieje wiele przypadków, w których termin jest odrzucony przez model ze względu na omawiany problem.

Zastosowanie okna kontekstowego przy jednorodnym modelu n-gram powoduje, że dla każdego potencjalnego terminu (oprócz przypadków skrajnych, np. koniec zdania, dokumentu) występuje więcej niż jeden szacunek prawdopodobieństwa. Metoda ważenia wielu wyników dla jednego potencjalnego terminu może być optymalizowana. W przedstawionych eksperymentach zastosowano wariant średniej arytmetycznej.

Kontynuując przykład analizy terminu "IBM", oszacowane prawdopodobieństwo dla omawianych sekwencji może być następujące:

[VB,JJR,term:NNP] = 0,4,
[JJR,term:NNP,NNP] = 0,2,
[term:NNP,NNP,NNP] = 0,6.

W przypadku analizy nowego tekstu, np. "pobranie starszego NYSE Stock Quotes", token "NYSE" jest przedmiotem wszystkich trzech reguł. Dotyczą go trzy różne szacunki prawdopodobieństwa. Wynik modelu musi być jednak jednoznaczny. Średnia arytmetyczna pozwala na oszacowanie prawdopodobieństwa $P = \frac{0,4+0,2+0,6}{3} = 0,4$.

Złożoność ekstrakcji terminów wieloczłonowych powoduje, że w niektórych przypadkach, na przykład przy tekstach charakteryzujących się minimalnym udziałem terminów wieloczłonowych, ich zastosowanie powoduje

spadek efektywności metod w porównaniu z klasyczną ekstrakcją terminów jednoczłonowych. Usunięcie z modelu funkcjonalności przetwarzania terminów wieloczłonowych spowoduje zatem wzrost efektywności metody. Dziedzina handlu elektronicznego charakteryzuje się znacznym udziałem terminów wieloczłonowych. Nazwy produktów i usług często są wyrażeniami wielowyrazowymi, np. wspomniany "A4-TECH Navigator Opto BW-5UP" (por. sekcję 1.1.1 na stronie 3).

Podsumowując, optymalizacja modelu dla dokumentów z dziedziny handlu elektronicznego implikuje wykorzystanie następujących wartości parametrów optymalizacyjnych:

- rozmiar okna kontekstowego: 5 wyrazów,
- model jednorodny z wartością stopnia n-gramów odpowiadającą maksymalnej dozwolonej wartości, czyli 3,
- brak wygładzania MLE,
- średnia arytmetyczna dla ważenia wyników dla terminu,
- włączenie mechanizmu przetwarzania terminów wieloczłonowych.

Rozdział 6

Ekstrakcja relacji

Opracowana w ramach metamodelu (równanie 3.1) metoda ekstrakcji relacji nietaksonomicznych NTR (równanie 3.8) oparta jest na zasadzie sprzężenia zwrotnego pomiędzy aksjomatami dziedzinowymi A (równanie 3.6) a cechami lingwistycznymi tekstu D zawartymi w anotacji lingwistycznej LA (równanie 3.2). Dwoma podstawowymi argumentami opracowanej metody jest zbiór anotacji lingwistycznej LA oraz zbiór aksjomatów A.

W obecnie istniejących rozwiązaniach ekstrakcja relacji nietaksonomicznych NTR następuje w rezultacie analizy cech lingwistycznych znajdujących się w zbiorze LA. Zarówno metoda ekstrakcji, jak i sam zbiór LA są w zdecydowanej większości przypadków różne, co nie zmienia faktu, że podejście jest podobne. OntoLT (Buitelaar i in., 2004a) jako reprezentatywne narzędzie tego typu podejścia umożliwia konstrukcję reguł lingwistycznych w języku XPath, które odnoszą się do cech lingwistycznych zawartych w zbiorze anotacji lingwistycznych LA. Przykładową regułę przedstawiono na rysunku 2.4 na stronie 93.

Problem polega na tym, że użytkownik takiego narzędzia musi dokładnie wiedzieć, czego szuka, jakie relacje są dla danej dziedziny ważne oraz w jaki sposób są skonstruowane. Wymaganie to jest nierealne i w konsekwencji powoduje, że pomimo istnienia mechanizmów temu służących, ekstrakcja relacji przynosi słabe efekty, przeważnie w postaci niskiej wartości zwrotu.

Wykorzystanie zbioru aksjomatów dziedzinowych A, które stanowią niepodważalną i ogólnie znaną wiedzę na temat dziedziny, daje możliwość rozwiązania tego problemu, tj. zwolnienia użytkownika od konieczności i trudu tworzenia skomplikowanych reguł lingwistycznych, których relewancja dziedzinowa może być niska.

```
Uslugodawca(x) => swiadczyUsluge(x,y) AND Usluga(y)

swiadczonaDla(x,y) => Usluga(x) AND Uslugobiorca(y)

Uslugodawca(y) => swiadczonaPrzez(x,y) AND Usluga(x)
```

Rysunek 6.1: Przykładowe reguły dla dziedziny handlu elektronicznego (korpus e-commerce)

6.1 Aksjomaty

W metodzie ekstrakcji relacji nietaksonomicznych przedstawionej w ramach metamodelu (równanie 3.1) użytkownik nie jest zobligowany do tworzenia reguł lingwistycznych. Reguły te tworzone są pośrednio przez sprzężenie zwrotne pomiędzy aksjomatami dziedzinowymi A (równanie 3.6) a cechami lingwistycznymi tekstu D zawartymi w anotacji lingwistycznej LA (równanie 3.2).

Zbiór aksjomatów A wykorzystany jest do wskazania bytów oraz relacji, które są istotne dla dziedziny. Na przykład, jeśli aksjomaty dziedzinowe zawierają relację "swiadczonaPrzez", celem metody jest ekstrakcja samej relacji oraz wszystkich jej składowych łącznie z klasami oraz instancjami. Ze zbioru aksjomatów A pozyskiwane są wszystkie elementy z wyjątkiem tych, które nie posiadają realizacji w analizowanym korpusie D.

Przykładowe reguły w zbiorze A przedstawia rysunek 6.1. Reguły te zostały opracowane w ramach analizy dziedziny handlu elektronicznego objętej korpusem *e-commerce*.

Reprezentacja aksjomatów odbiega od sposobu reprezentacji ontologii opartych na grafach RDF. W większości przypadków aksjomaty przedstawione są przy pomocy logiki pierwszego rzędu. Podstawą istniejących mechanizmów wnioskujących dla ontologii są wyrażenia Horna. Przedstawiona metoda ekstrakcji relacji dopuszcza więc aksjomaty dające się sprowadzić do wyrażeń Horna.

Językiem serializacji reguł dla Sieci Semantycznej jest Semantic Web Rule Language[1] (SWRL). W celach technicznej współoperatywności SWRL został wykorzystany do reprezentacji aksjomatów. Pomimo niektórych dodatkowych zmiennych, SWRL jest zgodny z logiką Horna[2]. Wybór SWRL jest również podyktowany faktem, że wykorzystywane mechanizmy wnioskujące wspierają reguły zapisane przy pomocy SWRL.

[1] `http://www.w3.org/Submission/SWRL/`

[2] `http://www.w3.org/Submission/WRL-related/`

6.2 Model uruchomieniowy

Przedstawiona metoda ekstrakcji relacji wykorzystuje model uruchomieniowy do udowadniania nowych faktów.

Uruchomienie modelu jest konsekwencją zapytania lub celu w postaci $\leftarrow B_1, \ldots B_n$. Dla danego zbioru aksjomatów, reguły A_1 (również oznaczającej program P_1) oraz zapytania $\leftarrow B_1, \ldots B_n$, celem uruchomienia jest weryfikacja przy jakich założeniach koniunkcja $B_1, \ldots B_n$ jest logiczną konsekwencją A_1, tj.: $P_1 \vdash B_1, \ldots B_n$.

Proces obliczeniowy oparty jest na dwóch mechanizmach, tj. podstawieniu i unifikacji (Robinson i Voronkov, 2001), dokładnie tak samo jak np. w Prologu.

6.3 Sprzężenie zwrotne

Sprzężenie zwrotne dla ekstrakcji relacji nietaksonomicznych jest cyklicznym procesem, który posiada następujące funkcjonalności:

1. Wykorzystanie cech zbioru anotacji lingwistycznych LA do tworzenia nowych wyrażeń w ontologii (elementów zbiorów C, TR, NTR).

2. Wykorzystanie wiedzy dziedzinowej w postaci aksjomatów do tworzenia nowych wyrażeń w ontologii (elementów zbiorów C, TR, NTR).

3. Umożliwienie wykorzystania cech anotacji lingwistycznych LA w aksjomatach dziedzinowych A.

4. Umożliwienie wykorzystania aksjomatów dziedzinowych A w procesie ekstrakcji elementów ontologii ze zbioru anotacji lingwistycznych LA.

Wykorzystanie pierwszych dwóch funkcjonalności nie jest niczym nowym — w zasadzie wszystkie obecnie istniejące rozwiązania do ekstrakcji relacji nietaksonomicznych oparte na tekście działają w ten sposób. Zarówno zbiór anotacji lingwistycznych LA, jak i zbiory aksjomatów A są naturalnymi źródłami wiedzy na temat dziedzinowych relacji nietaksonomicznych.

Pozostałe dwie cechy przedstawionej metody ekstrakcji relacji wyróżniają ją w sposób unikatowy i umożliwiają wykorzystanie obu źródeł naprzemiennie. Metoda umożliwia bowiem wykorzystanie cech anotacji lingwistycznej w aksjomatach oraz zależności aksjomatycznych w ekstrakcji elementów ontologii ze zbioru LA.

Powiązanie wszystkich czterech cech tworzy cykl sprzężenia zwrotnego, w którym informacje ze zbiorów LA oraz A uzupełniają się wzajemnie.

6.3.1 Wymagania

Przedstawione cechy metody wymagają znaczących zmian w sposobie funkcjonowania obecnie istniejących narzędzi. Należy przede wszystkim zidentyfikować cechy obu źródeł, które mogą być wykorzystywane w procesie sprzężenia zwrotnego.

Cechy zbioru aksjomatów A zostały zidentyfikowane w procesie analizy aksjomatów dziedzinowych dla testowych korpusów. Na podstawie wyników analizy proponuje się wprowadzenie mechanizmów umożliwiających definicję następujących cech lingwistycznych tekstu:

Wystąpienie danego tokena — wystąpienie w tekście etykiety instancji klasy z ontologii, np. etykiety "pl: Znak"[3] przyporządkowanej do instancji "CentrumKomputeroweZnak" klasy "InternetShop"[4].

Współwystępowanie danych dwóch tokenów w terminie — dowolne dwie etykiety instancji z ontologii współwystępują w terminie, np. "pl: komputronik" jako etykieta instancji klasy "Organization" oraz "pl: komputery osobiste" jako etykieta instancji klasy "ComputerArchitecture" współwystępują w terminie składającym się z trzech tokenów: "komputery osobiste komputronik".

Współwystępowanie dwóch terminów w dokumentach — dwie etykiety instancji współwystępują w dokumentach w kontekście całego korpusu D, np. "pl: komputronik" jako etykieta instancji klasy "Organization" oraz "wsparcie techniczne" jako etykieta instancji klasy "Service" współwystępują w dokumentach.

Uogólniona postać przedstawionych wymagań w stosunku do ekspresywności zbioru aksjomatów A wygląda w sposób następujący:

1. Wystąpienie danego tokena:

   ```
   includesTerm(x, klasa:ID)
   ```

   gdzie x oznacza dowolną instancję ontologii, a $klasa : ID$ oznacza ID klasy ontologii, np. "Organization".

[3] Pełna nazwa etykiety w języku OWL składa się z prefiksu oznaczającego język etykiety (np. pl) oraz nazwy w danym języku

[4] Nazwy klas oraz relacji zgodnie ze specyfikacją OWL powinny być pisane w języku angielskim. Dopuszcza się stosowanie innych języków naturalnych w przypadku nazw instancji (zarówno klas, jak i relacji)

2. Współwystępowanie dwóch tokenów w terminie:

   ```
   co-occurInTerm(x,y)
   ```

   gdzie x i y oznaczają dowolne instancje ontologii, np. Organization(Komputronik) AND ComputerArchitecture(KomputerOsobisty) AND co-occurInTerm(Komputronik,KomputerOsobisty).

3. Współwystępowanie dwóch terminów w dokumentach:

   ```
   co-occurInDocument(x,y)
   ```

   gdzie x i y oznaczają dowolne instancje ontologii, np. Organization(Komputronik) AND Service(WsparcieTechniczne) AND co-occurInDocument (Komputronik,WsparcieTechniczne).

Kluczowa w powyższych wyrażeniach jest relacja pomiędzy nazwami (ID) klas oraz instancji a etykietami. Definicja wyrażeń opiera się na unikatowych nazwach klas lub instancji (ID), natomiast wszelkie operacje związane z modelem uruchomieniowym wiążą się z interpretacją wszystkich skojarzonych etykiet. Powoduje to, że definiując np. "Organization(CKZnak)" model operuje na etykietach instancji "CKZnak", które dla języka polskiego to m.in.: "Znak", "Centrum Komputerowe Znak", czy "CK Znak".

Semantyka zdefiniowanych predykatów dodatkowych związana jest z cechami lingwistycznymi. W związku z tym, powyższe predykaty nazwane zostały *predykatami lingwistycznymi*.

6.3.2 Rozwiązanie

Predykaty lingwistyczne mogą być wykorzystane na 2 różne sposoby:

Anotacje kontekstowe (*tzw. in-line annotations*). Umieszczenie anotacji kontekstowych przed każdą regułą SWRL w pliku SWRL. Anotacja kontekstowa oznacza więc informację na temat przetwarzania danej reguły SWRL. Na przykład predykat lingwistyczny co-occur(x,y) nie występuje w ciele elementu swrl:Imp w definicji reguły, tylko w specyficznym znaczniku przed regułą, np. "Processing information". Przed uruchomieniem mechanizmu wnioskującego odpowiedni procesor przetwarza informacje w postaci anotacji kontekstowych i przetwarza ją.

Przetwarzanie polega na przeniesieniu deklaratywnych instrukcji w anotacji kontekstowej w ciało reguły. Następuje przepisanie reguły do postaci odwzorowującej żądane właściwości. Główny wysiłek to przygotowanie procesora, który poprawnie odwzoruje anotację kontekstową na reguły gotowe do przetwarzania standardowymi mechanizmami wnioskującymi.

Właściwości obiektowe. Metoda polega na przedstawieniu predykatów jako standardowych relacji obiektowych ontologii (*tzw. object properties*). Głównym problemem jest wydajny sposób odwzorowania wszystkich predykatów na poprawne elementy ontologii.

Druga z wymienionych metod jest bardziej odpowiednia, ponieważ nie ma potrzeby definiowania składni i semantyki anotacji kontekstowej, a także nie jest konieczne opracowanie procesorów przetwarzania. Traktowanie predykatów lingwistycznych jako elementów ontologii przenosi cały ciężar ich interpretacji na metodę ich budowy oraz mechanizmy wnioskujące. Z tych powodów proponuje się zastosowanie drugiej możliwości, która składa się z następujących kroków:

1. Każda poprawna ontologia[5] zawiera dodatkowe właściwości obiektowe będące odwzorowaniem predykatów lingwistycznych[6]:
   (a) includesTerm(owl:Thing, owl:Class),
   (b) coOccurInTerm(owl:Thing, owl:Thing),
   (c) coOccurInDocument(owl:Thing, owl:Thing).
2. Dla danego korpusu D dodatkowe właściwości obiektowe muszą zostać rozwiązane tak, aby wskazywały na poprawne obiekty w zbiorze anotacji lingwistycznych LA. Na przykład w celu rozwiązania predykatu lingwistycznego *includesTerm*, należy wyszukać w zbiorze LA wszystkie wystąpienia etykiet klasy dla listy tokenów składających się na terminy. W rezultacie znane są wszystkie instancje, w których wystąpiły etykiety danej klasy.
3. Model ontologii jest uzupełniany realizacjami predykatów lingwistycznych dla instancji.
4. Uruchamiany jest mechanizm wnioskujący na podstawie stworzonej listy reguł zawierającej realizacje predykatów lingwistycznych.

[5]Terminem *poprawna ontologia* określa się ontologię, która spełnia wymagania modelu formalnego.

[6]Przestrzeń nazw *owl* oznacza `http://www.w3.org/2002/07/owl`.

5. Nowe fakty są dodane do niemonotonicznego modelu ontologii zgodnie z modelem uruchomieniowym.

Dziedzina i zasięg predykatów lingwistycznych *owl:Thing* powinna zostać dostosowana do faktycznie wykorzystywanych zasobów w konkretnej aplikacji metody. Operacja ta znacząco zmniejsza liczbę analizowanych obiektów i rozmiar samej ontologii przetwarzanych przez mechanizm wnioskujący, a w konsekwencji obniży złożoność obliczeniową metody oraz zwiększy wydajność.

Każdy z predykatów lingwistycznych posiada odrębną logikę i w związku z tym potrzebuje oddzielnej implementacji. Nakład implementacyjny jest różny w zależności od predykatu. Wystąpienie danego tokena jest najprostszym do zaimplementowania predykatem, ponieważ wymaga tylko sprawdzenia wszystkich tokenów w korpusie pod kątem wystąpienia jednego z elementów zbioru składającego się z etykiet szukanej klasy. Złożoność relacji współwystępowalności jest znacznie większa, ponieważ wymaga zastosowania miar współwystępowalności, np. standardowych miar Jaccarda, Dice'a lub cosinusa (Kuropka, 2005). W szczególności w przypadku obliczania miary współwystępowalności w dokumentach złożoność obliczeniowa jest wysoka. W tego powodu sugeruje się przeprowadzenie obliczeń wstępnych w oderwaniu od właściwego procesu ekstrakcji relacji.

Model ontologii oraz serializacja aksjomatów znajduje się w plikach przeważnie o dużym rozmiarze. Umieszczanie modelu ontologii wraz z regułami jest podatne na błędy oraz kosztowne w zarządzaniu. W warstwie fizycznej proponuje się więc wydzielenie obu obszarów. Plik reguł SWRL, jako poprawnie zbudowany plik OWL, powinien importować model ontologii przechowywany w fizycznie odrębnym pliku. W rezultacie zwiększona jest spójność obu modeli, ponieważ każdy z nich jest wykorzystywany w odrębnych fazach — model ontologii zmienia się w wyniku ekstrakcji relacji, natomiast aksjomaty ulegają zmianie tylko w przypadku zmiany ogólnie przyjętych reguł w danej dziedzinie.

6.4 Optymalizacja modelu

W celu uzyskania optymalnych wyników przedstawioną metodę ekstrakcji relacji należy poddać optymalizacji. Optymalizacja przedstawionej metody ekstrakcji relacji polega na dążeniu do jak największej efektywności metody mierzonej np. miarą F lub dążeniu do przetwarzania jak największej liczby zagadnień lingwistycznych (np. współwystępowalność, nawiązania, akronimy, skróty, etc.). Optymalizacja jest dokonywana poprzez ustalenie wartości następujących parametrów:

- liczba zdefiniowanych predykatów lingwistycznych,
- liczba cykli sprzężenia zwrotnego,
- próg klasyfikacji dla niebinarnych predykatów lingwistycznych, np. miary współwystępowalności.

Zwiększanie liczby zdefiniowanych predykatów lingwistycznych prowadzi do wzrostu liczby przetwarzanych zagadnień lingwistycznych, tj. właściwości zbiorów LA. Wpływ na efektywność metody jest jednak trudny do przewidzenia i zależy od jednostkowych efektywności dodawanych predykatów. Wzorcowa implementacja metody przedstawiona w rozdziale 7. wykorzystuje predykaty lingwistyczne zidentyfikowane powyżej, które w drodze obserwacji korpusu uznane zostały za najczęściej spotykane.

W miarę rozwoju mechanizmów wnioskujących można zwiększać liczbę cykli sprzężenia zwrotnego. Kryterium optymalizacji jest stosunkowo proste: liczbę cykli należy zwiększać o jednostkę, aż do stanu, gdy liczba faktów przestaje wzrastać (przy zachowaniu niemonotoniczności modelu).

Standardowym parametrem optymalizacyjnym jest próg klasyfikacji ustalany dla każdego niebinarnego predykatu lingwistycznego osobno. W przypadku możliwości zastosowania wielu miar dla tego samego predykatu (np. dla współwystępowalności miary Jaccarda, Dice'a lub cosinusa) optymalizować można również wagę poszczególnych miar.

Rozdział 7

Ewaluacja

Niniejszy rozdział przedstawia warsztat oraz wyniki przeprowadzonych eksperymentów dla wszystkich objętych zakresem pracy faz cyklu uczenia ontologii z tekstu. Wyniki przedstawione są w kolejności zgodnej z porządkiem w metamodelu (równanie 3.1).

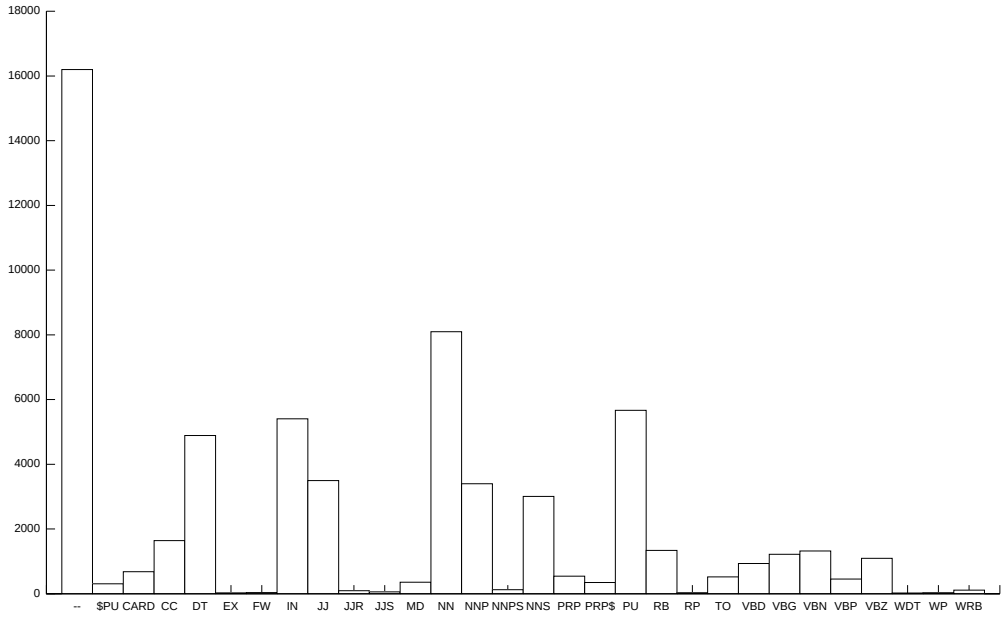
7.1 Korpusy testowe

W niniejszej pracy zastosowano dwa korpusy. Różnią się one przede wszystkim dziedziną — pierwszy korpus charakteryzuje ofertę sklepu internetowego, drugi z nich zawiera informacje bieżące z działalności uniwersytetu. Oba korpusy różnią się również językiem naturalnym (język polski a język angielski) oraz stylem narracji (opis techniczny a krótkie wiadomości).

Podstawowym korpusem dla ewaluacji metod w niniejszej pracy jest korpus e-commerce, drugi zastosowano pomocniczo, aby zweryfikować ogólność metody. Chronologicznie pierwszym zebranym w całości korpusem był jednak korpus z informacjami bieżącymi z działalności uniwersytetu — zostanie więc przedstawiony jako pierwszy.

7.1.1 Korpus KMi-News

Korpus KMi-News jest kolekcją dokumentów rozprowadzaną wraz ze środowiskiem do uczenia ontologii OntoLT (Buitelaar i in., 2004a; Buitelaar i Sintek, 2004; Buitelaar i in., 2004b). Zawiera 273 zaanotowane dokumenty, co daje łącznie liczbę 62303 tokenów. Tematyka korpusu to aktualności z działalności instytucji Knowledge Media Institute przy uniwersytecie Open University w Milton Keynes (Wielka Brytania). Wiadomości mają postać notek prasowych z roku 2004.



Rysunek 7.1: Rozkład części mowy korpusu KMi w wersji źródłowej. W celu zachowania czytelności rysunku zrezygnowano ze wszystkich części mowy, których częstość występowania jest mniejsza niż 20

Format korpusu jest zgodny z formatem MuchMore (Vintar i in., 2001) oraz OntoLT (Buitelaar, 2003). Zwłaszcza zgodność z OntoLT umożliwia rzetelną ewaluację prowadzonych badań oraz umożliwia zastosowanie opracowanych reguł językowych w narzędziu stosowanym powszechnie, tym samym uzyskując powtarzalność prowadzonych w ten sposób eksperymentów (również przez innych badaczy).

Korpus KMi-News jest korpusem zaanotowanym na wszystkich trzech poziomach anotacji: tekstu (ang. *tokens*), wyrażeń (ang. *phrase*) oraz logicznych części zdania (ang. *clause*). Stanowi więc kompletną bazę do stosowania metod wykorzystujących zgromadzoną informację lingwistyczną.

W warstwie tekstu, tj. właściwości tokenów, najistotniejszą charakterystyką korpusu jest zastosowany zbiór znaczników części mowy (POS) oraz jego rozkład. Rozkład części mowy dla pełnego korpusu KMi-News przedstawiono na rysunku 7.1. Po pierwsze, znacząca część tokenów jest w ogóle nie zaanotowana i stanowi 26,33% ogólnej liczby tokenów. Rodzi to poważne konsekwencje dla jakichkolwiek operacji wykonywanych na korpusie. Poza tym rozkład jest zgodny z intuicją, tzn. znaczące grupy stanowią:

- rzeczowniki oraz ich formy (NN, NNS, NNP, NNPS),

- przyimki oraz zaimki (IN, DT),
- przymiotniki (JJ),
- znaki interpunkcyjne (PUNCT),
- czasowniki oraz ich formy (VBD, VBG, VBN, VBP, VBZ).

Bezpośrednią przesłanką do dokonania szczegółowej analizy korpusu był fakt braku anotacji części mowy (26,33% całości korpusu). Obserwacja wykazała, że następne ok. 20% tokenów jest zaanotowanych niepoprawnie. Najczęstszymi błędami w anotacji domyślnej były dwie kategorie błędów. Przede wszystkim, bardzo często anotacja błędnie rozpoznawała znaczenie rzeczownika jako czasownik i odwrotnie (np. *stay*). Zwłaszcza w języku angielskim jest to problem wymagający dodatkowych operacji, najczęściej analizy kontekstu wystąpienia. Drugą najczęstszą grupą błędów stanowią wieloznaczności na linii NN-NNP (rzeczownik a nazwa własna, np. *institute*).

Brak anotacji oraz błędna anotacja tokenów stanowią razem prawie 50% całości tokenów. Pomimo więc względnej dostępności samego korpusu, dużego dostosowania jego struktury do potrzeb uczenia ontologii oraz dużych możliwości ewaluacji, korpus nie nadaje się bezpośrednio do wykorzystania. Wszelkie próby jego zastosowania w źródłowej postaci spowodują olbrzymie zniekształcenia w uzyskanych wynikach.

Korpus KMi-News potrzebuje zatem ponownej anotacji, która obejmować będzie zarówno brakujące części mowy, jak i części mowy zaanotowane błędnie. Pierwszą metodą anotacji jest metoda ręczna polegająca na zaanotowaniu tekstu przez eksperta. Korpus w ten sposób zaanotowany staje się korpusem referencyjnym, którego anotacja jest porównywana jako wzorcowa przy późniejszej ewaluacji.

Anotacja ręczna

Anotację ręczną przeprowadza się na reprezentatywnej próbie najczęściej w dwóch celach. Po pierwsze, dla celów ewaluacyjnych — ekspert zrobi to najlepiej (przynajmniej tak się z reguły zakłada). Korpus zaanotowany przez eksperta stanowi więc korpus wzorcowy. Gdyby metoda powieliła wynik korpusu z anotacją ręczna, jej skuteczność wynosi 100%. Po drugie, anotację ręczną stosuje się wtedy, gdy nawet najmniejsze zniekształcenie korpusu może mieć znaczenie dla skuteczności badanych metod. Ma to znaczenie zwłaszcza w przypadku opracowywania modelu i testowania różnych jego postaci, czyli w początkowej fazie pracy badacza. Później, gdy model jest już opracowany, nawet jego trenowanie może odbywać się już na jakościowo gorszych danych.

Do tych dwóch najczęściej występujących przesłanek, dochodzi w tym przypadku jeszcze jedna funkcja — poprzez ręczną anotację można się bardzo wiele dowiedzieć o prawidłowościach znajdujących się w tekście. Już na poziomie anotacji można więc próbować wychwytywać reguły stanowiące późniejszy model języka.

Z powyższych przyczyn wynika, że szczególnie w przypadku opracowywania modeli dobrze zaanotowany korpus jest niezmiernie ważny. Z tego powodu zdecydowano się na ręczną anotację reprezentatywnej próby korpusu KMi-News, tj. 1132 tokenów składających się na 11 pierwszych dokumentów korpusu. Niniejsza wersja korpusu została nazwana w skrócie KMi-11.

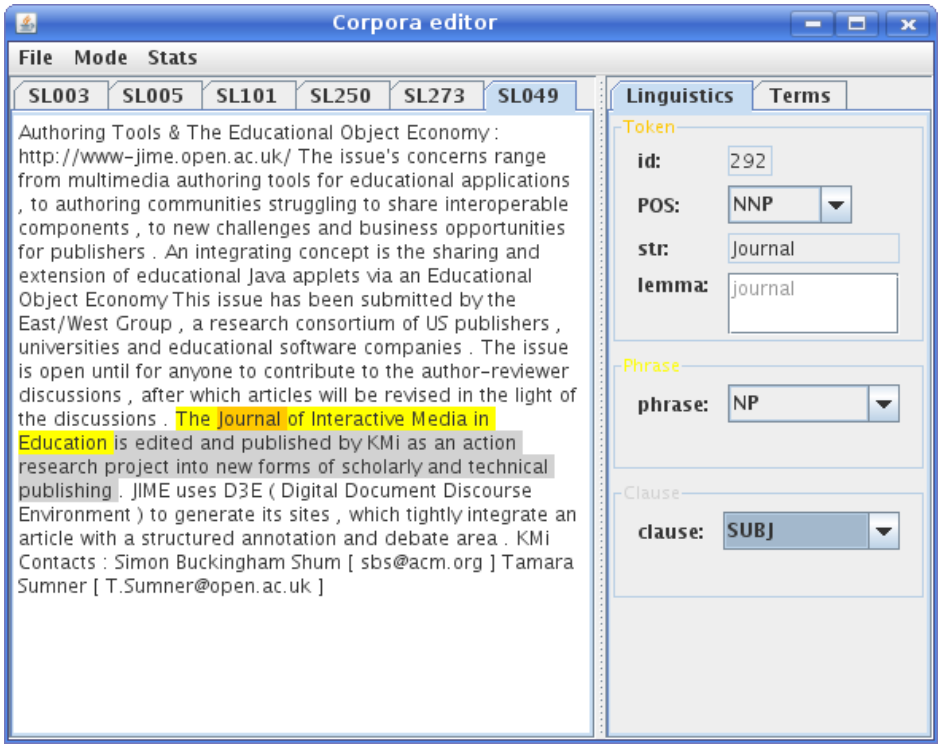
Anotację ręczną przeprowadzono z wykorzystaniem opracowanego prototypu narzędzia, który umożliwia m.in.:

- odzwierciedlenie struktury tekstu na wszystkich trzech poziomach: tekstu, wyrażeń oraz części zdania,
- wyświetlenie oraz modyfikację właściwości anotacyjnych,
- zarządzanie zbiorami znaczników POS, wyrażeń oraz funkcji gramatycznych,
- anotację terminów jednotokenowych oraz wielotokenowych.

Przykładowe okno prototypu aplikacji do anotacji ręcznej przedstawiające strukturę anotacji widoczne jest na rysunku 7.2.

Anotacja ręczna przebiegała dwuetapowo. W pierwszym etapie zaanotowano części mowy, tj. uzupełniono brakujące wpisy oraz zmieniono wpisy błędne. W praktyce oznaczało to przejrzenie wszystkich 1132 tokenów. W drugim przejściu zaanotowano terminy. Przed przystąpieniem do anotacji ręcznej przyjęto następujące założenia dotyczące anotacji terminologii:

1. Anotowane są terminy jednotokenowe, które nadają się do jakichkolwiek późniejszych faz cyklu uczenia ontologii. Mogą to być pojęcia, instancje, byty nazwane, akronimy, skróty, itd.
2. Anotowane są terminy, które charakteryzują się znaczną istotnością dziedzinową, np. termin *contrary* w wyrażeniu *on the contrary* nie jest dobrym terminem i nie podlega anotacji.
3. Anotowane są terminy wielotokenowe (wieloczłonowe), które mają znaczenie jako całość. Na przykład *Open University* jako nazwa uczelni traktowana jest w całości jako termin, podczas gdy w wyrażeniu *open university*, w którym token *open* występuje jako przymiotnik, zaanotowany zostanie tylko termin *university*.



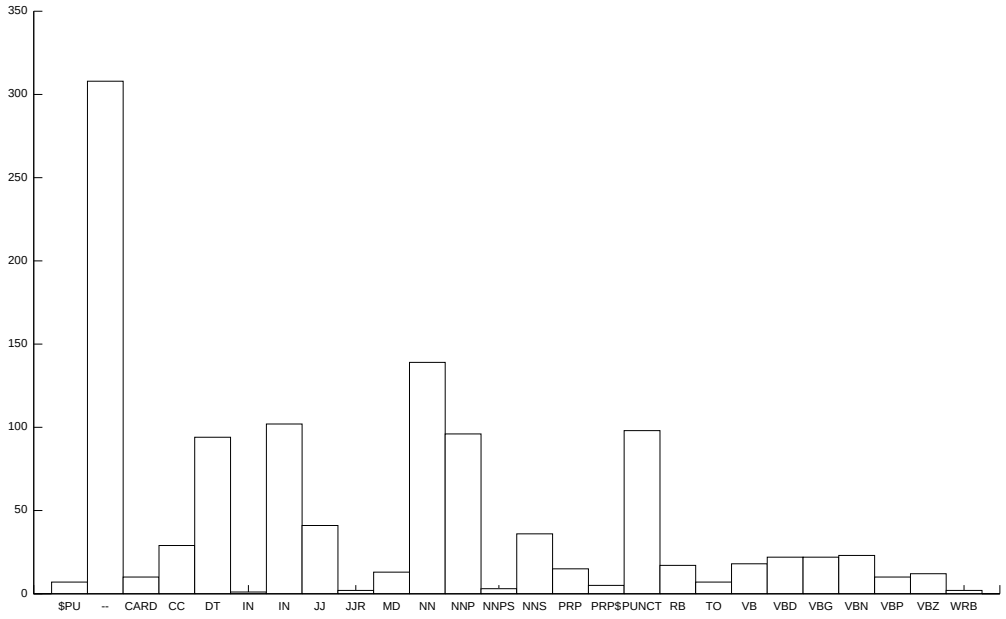
Rysunek 7.2: Główne okno prototypu aplikacji do anotacji ręcznej oraz przeglądania korpusów i ich właściwości lingwistycznych. Aplikacja została stworzona przez autora wyłącznie na potrzeby pracy i została opracowana przy użyciu języka Java Standard Edition i standardowych bibliotek Swing

4. Nazwy własne, w tym osoby anotowane są w ten sam sposób co terminy wielotokenowe. Na przykład wyrażenie *Dr. Hans Geisler* powoduje anotację dwóch terminów: *Dr.* oraz *Hans Geisler*.

5. Nazwy własne anotowano w trzech kategoriach: osoby (PER), byty geograficzne (LOC) oraz organizacje (ORG). Podział nazw własnych wynika ze znaczącego udziału nazw własnych w całości anotowanego korpusu.

6. Wyrażenia temporalne nie zostały anotowane, chyba że spełniają założenia ogólne dla terminologii.

7. Tokeny błędnie zbudowane nie zostały anotowane. Błędna anotacja jest wynikiem działania takich procesów jak tokenizacja i pozostaje poza zakresem anotacji ręcznej. Na przykład wyrażenia *he'll* lub *KMi/IBM* nie można jednoznacznie poddać klasyfikacji POS.



Rysunek 7.3: Rozkład części mowy korpusu KMi-11 w wersji źródłowej

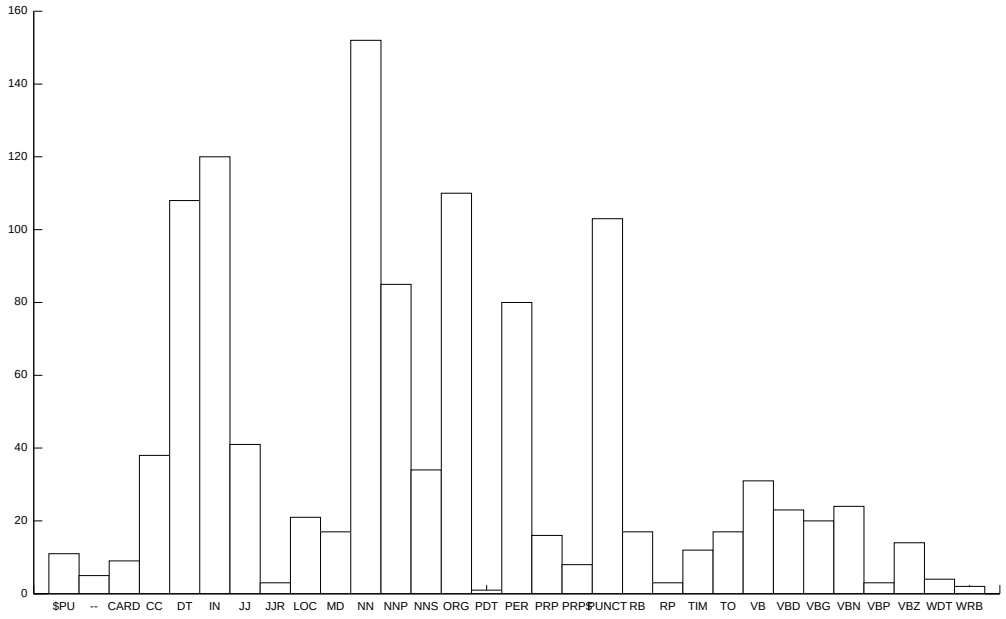
8. Do anotacji wykorzystano zbiór znaczników Penn Treebank.

W wyniku anotacji ręcznej z zastosowaniem powyższych założeń powstało 313 terminów oraz kompletny zestaw anotacji części mowy. Rozkład części mowy przed i po anotacji ręcznej przedstawiono odpowiednio na rysunkach 7.3 oraz 7.4. Uzyskano następujące wyniki:

1. Znaczący spadek tokenów bez anotacji (z 308 do 5, co oznacza redukcję o 98%). 5 pozostałych niezaanotowanych tokenów wynika z błędnej tokenizacji tekstu.

2. Znaczący wzrost udziału tokenów o klasyfikacji nowych typów nazw własnych, zwłaszcza osób (PER) i organizacji (ORG), a w mniejszym stopniu lokalizacji (LOC) i czasu (TIM). Udział organizacji przewyższa nawet udział wszystkich innych niesklasyfikowanych do nowych grup nazw własnych. Pomimo więc bezwzględnego spadku udziału niesklasyfikowanych inaczej nazw własnych (NNP), w przypadku połączenia wszystkich nowych typów nazw własnych (NNP + ORG + PER + LOC + TIM), ich udział w całym rozkładzie jest największy ze wszystkich grup.

3. W ogólności wzrost udziału wszystkich innych grup części mowy.



Rysunek 7.4: Rozkład części mowy korpusu KMi-11 po anotacji ręcznej

Zakładając, że wynik anotacji ręcznej jest wzorcowy, można policzyć skuteczność anotacji analizowanej próby. Zakładając, że N_{tokens} oznacza liczbę wszystkich tokenów oraz n_{source} oznacza liczbę poprawie zaanotowanych tokenów w korpusie źródłowym, wzrost poprawności anotacji POS dla anotacji ręcznej wynosi:

$$E_{manual} = \frac{N_{tokens} - n_{source}}{n_{source}} * 100\% = \frac{1132 - 654}{654} * 100\% = 73,09\%. \quad (7.1)$$

Anotacja automatyczna

Anotacja ręczna części mowy oraz terminologii dokonana wyłącznie na 11. pierwszych dokumentach korpusu KMi-News okazała się bardzo czasochłonna. Szacunkowe zaanotowanie całości korpusu KMi-News wynosi ok. 2 osobomiesiące. Podczas tak długiego procesu bardzo trudno jest zachować niezmienność założeń. Ponadto jakakolwiek zmiana korpusu wymaga ponownej anotacji. Dlatego zastosowano drugą metodę — anotację automatyczną metodą prototypu.

Dla celów anotacji automatycznej zastosowano narzędzie GATE, które jest rozbudowanym środowiskiem służącym do budowania aplikacji inżynierii tekstu (Tablan i in., 2004). ANNIE (Cunningham i in., 2007) jest aplikacją GATE służącą do anotacji części mowy oraz towarzyszących procesów. W ce-

lu anotacji korpusu dla uczenia ontologii tandem GATE + ANNIE można wykorzystać na trzy różne sposoby:

Domyślnie GATE rozprowadzany jest razem ze środowiskiem graficznym służącym m.in. do wczytywania plików, uruchamiania zasobów i procesów anotacyjnych oraz zapisywania wyników. Jest to domyślny sposób wykorzystania narzędzia GATE. Na załadowanym korpusie uruchamiane są standardowe metody przetwarzania lingwistycznego ANNIE. Zapis następuje w formacie GATE XML. Scenariusz domyślny rodzi jednak poważne problemy:

- przy każdym wykorzystaniu narzędzia należy je uruchomić oraz poprzez GUI zainicjować. Przy dużej liczbie różnych korpusów oraz wielu miejscach wykorzystania narzędzia jest to niezwykle uciążliwe,
- podczas ładowania dokumentów z korpusu KMi-News zostaje zaburzona pozycja tokenów. GATE błędnie interpretuje spacje pomiędzy znacznikami dokumentu XML. Powoduje to utratę powiązania zbioru anotacji z pozycją tokena w dokumencie. Staje się to problemem na przykład w sytuacji, gdy dokument zawiera dwa jednakowe wyrazy. Nie wiadomo wówczas, który zbiór anotacji odnosi się do danego tokena,
- format wyjściowy GATE nie jest zgodny z domyślnym formatem anotacji. Powoduje to konieczność budowania mediatorów. Implementacja mediatorów jest czasochłonna oraz podatna na błędy. Każda zmiana w strukturze formatów oznacza ponowną implementację mediatorów.

Stworzenie zasobów GATE. Model danych środowiska GATE opiera się na zastosowaniu trzech typów zasobów: zasoby lingwistyczne (dokumenty, korpusy, słowniki, itd.), zasoby obliczeniowe (tokenizatory, lematyzatory, dzielenie zdań, itd.) oraz zasoby graficzne (elementy GUI). Problemy z formatem ładowania plików oraz ich zapisu do formatu zgodnego z domyślnym formatem anotacji można zlikwidować poprzez implementację własnych zasobów lingwistycznych GATE. Należy stworzyć dwa typy zasobów lingwistycznych — format wejściowy, który poprawnie odczyta tekst dokumentów oraz format wyjściowy zgodny z domyślnym formatem anotacji. Implementacja zasobów w narzędziu trzecim w stosunku do metamodelu jest jednak nieefektywna. Nieefektywność wynika z następujących powodów:

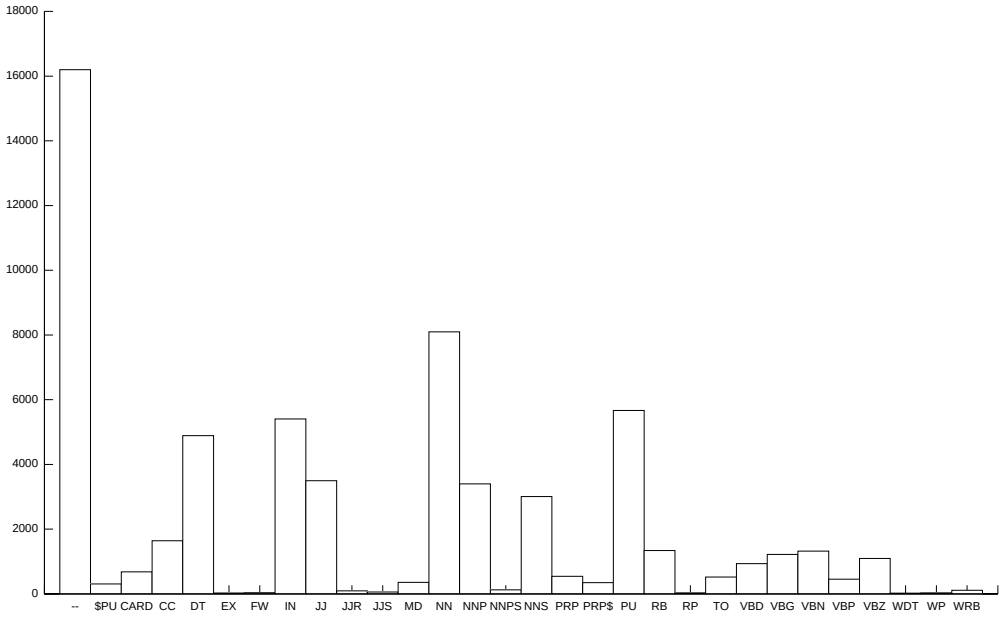
- GATE ulega ciągłemu rozszerzaniu — nie ma zatem pewności, że nie zmieni się samo narzędzie, co spowoduje konieczność ponownej implementacji zasobów,
- nie ma pewności, że GATE pozostanie jedynym możliwym do wykorzystania narzędziem. W przypadku przejścia na inne narzędzie anotacji inwestowanie w rozszerzanie niewykorzystywanego narzędzia jest zmarnowaniem zasobów,
- GATE poprzez swój model zasobów narzuca własne rozwiązania, co powoduje ograniczenie ekspresywności formatu oraz możliwych do zastosowania operacji,
- jest to rozwiązanie doraźne, tzn. nadal nie ma żadnej integracji z prototypem dla metamodelu,
- istnieje potrzeba uruchamiania i inicjowania środowiska graficznego GATE oraz postępowania zgodnie z domyślną procedurą wykorzystania zasobów.

Zewnętrzne API. Ostatnia metoda anotacji przy użyciu narzędzia GATE polega na wykorzystaniu zewnętrznego API GATE (Kenter i Maynard, 2005). Jest to jedyna metoda, która umożliwia pominięcie uciążliwego uruchamiania i inicjowania środowiska graficznego, a także dowolną kolejność wykorzystania zasobów lingwistycznych, co daje dużą elastyczność anotacji. W tym celu należy włączyć bibliotekę GATE do prototypu metamodelu oraz skorzystać z instrukcji API GATE. Tym samym uzyskuje się możliwość wykorzystania funkcjonalności oferowanej przez GATE w dowolnych punktach cyklu życia metamodelu. Pominięty zostaje również problem implementacji mediatorów do i z różnych formatów, ponieważ sposób przetwarzania zbiorów anotacji leży w gestii programisty.

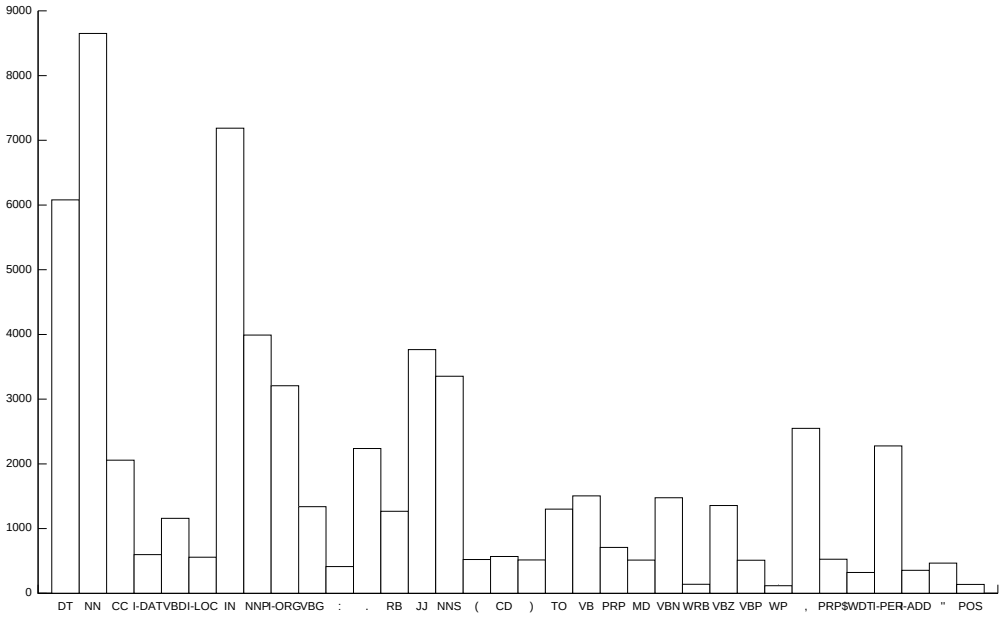
W celach anotacji automatycznej wykorzystano trzeci rozpatrywany wariant, tj. stworzono mechanizm pobierania zbiorów anotacji dla tekstu przy wykorzystaniu API GATE. W związku z automatyzacją procesu, rozmiar anotowanego korpusu jest w zasadzie dowolny. Czas potrzebny do zaanotowania częściami mowy całości korpusu KMi-News wynosi ok. 4 minut[1].

Rozkład części mowy dla korpusu źródłowego (KMi-News) oraz korpusu po anotacji automatycznej przy pomocy API GATE (KMi-Gate) przedstawiono na rysunkach 7.5 oraz 7.6 na stronie 164. Rysunek 7.5 jest identyczny

[1]Pomiaru dokonano na komputerze klasy PC, 1GB RAM, Centrino 1.8M, pracującego pod kontrolą systemu operacyjnego Linux 2.6.



Rysunek 7.5: Rozkład części mowy korpusu KMi w wersji źródłowej



Rysunek 7.6: Rozkład części mowy korpusu KMi-News po anotacji automatycznej. W celu zachowania czytelności rysunku zrezygnowano ze wszystkich części mowy, których częstość występowania jest mniejsza niż 100

z rysunkiem 7.1 na stronie 156. W miejscu tym został powtórzony w celu porównania z wynikami anotacji automatycznej.

Korpusu źródłowego oraz korpusu KMi-Gate niestety nie można porównać w sposób bezpośredni. Pomimo zapewnień autorów, że oba formaty korzystają ze zbioru znaczników Penn Treebank (Marcus i in., 1993), analiza wykryła nieścisłości w zbiorze używanych znaczników. Zwłaszcza domyślny format anotacji posiadał znaczniki odbiegające od norm przyjętych w Penn Treebank. Do najistotniejszych odchyleń należy zaliczyć:

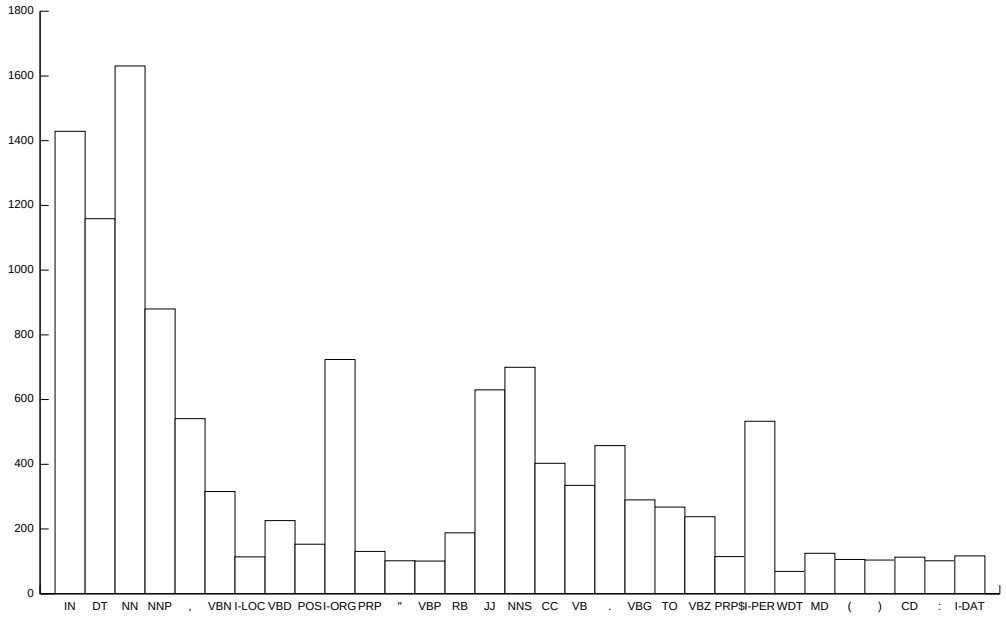
- różnicę pomiędzy znacznikiem dotyczącym typów liczbowych — *CARD* w formacie domyślnym oraz *CD* w GATE,
- sposób anotacji znaków interpunkcyjnych — domyślny format posiada jeden znacznik *PUNCT* na oznaczenie wszystkich znaków interpunkcyjnych, podczas gdy GATE używa większego stopnia granulacji typu,
- brak wykorzystania znacznika *POS* w domyślnym formacie anotacji.

Najistotniejszym osiągnięciem anotacji automatycznej jest spadek liczby tokenów bez anotacji z 16196 do 361. Ważną zmianą jest również względny wzrost udziału nazw własnych (*NNP*). Najprawdopodobniej wynika to z dużej skuteczności zastosowanej w GATE metody klasyfikacji nazw własnych. Wśród pozostałych grup znaczników nie zauważono znaczących zmian.

Zakładając całkowitą poprawność anotacji automatycznej można policzyć jej efektywność (podobnie jak w równaniu 7.1). Niestety, założenie to jest zbyt ostre, dlatego można policzyć co najwyżej liczbę oznaczającą procentową zmianę znaczników POS. Zakładając, że N_{tokens} oznacza liczbę wszystkich tokenów w korpusie oraz n_{source} oznacza liczbę niezmienionych anotacji tokenów w korpusie źródłowym, zmiana wynikająca z anotacji POS wynosi:

$$C_{auto} = \frac{N_{tokens} - n_{source}}{n_{source}} * 100\% = \frac{62303 - 34403}{34403} * 100\% = 81,10\%. \quad (7.2)$$

Istnieją jednak przesłanki pozwalające przypuszczać, że anotacja automatyczna jest bardzo dokładna i dąży do całkowitej poprawności. Po pierwsze, można porównać liczbę tokenów bez anotacji. Dokonana analiza wskazuje, że liczba tokenów bez anotacji wynosi 361, co stanowi 0,61% ogólnej ich liczby. Wartość ta jest więc bardzo niska. Po drugie, dokonano metodą obserwacji analizy wybranych dokumentów będących wynikiem anotacji automatycznej. Podczas analizy nie stwierdzono żadnych błędów. Istnieje zatem duże prawdopodobieństwo, że obliczony procent zmiany części mowy jest równocześnie skutecznością anotacji automatycznej.



Rysunek 7.7: Rozkład części mowy korpusu KMi-70. W celu zachowania czytelności rysunku zrezygnowano ze wszystkich części mowy, których częstość występowania jest mniejsza niż 50

Powstały korpus KMi-Gate nie zawiera anotacji w warstwie terminologii. W celach uczenia lub ewaluacji ten typ anotacji zawsze należy pozostawić do ręcznej anotacji eksperckiej. Wcześniej anotowany korpus KMi-11 zawiera już dokumenty zaanotowane w warstwie terminologii. Nie jest to jednak jeszcze korpus reprezentatywny, dlatego postanowiono rozszerzyć anotację terminologii. W sumie w warstwie terminologii zaanotowano pierwszych 70 dokumentów korpusu zgodnie z wcześniejszymi założeniami. Powstały w ten sposób korpus obejmuje anotację części mowy wszystkich dokumentów oraz anotację w warstwie terminologii dla pierwszych 70. dokumentów. W celu jednoznacznej identyfikacji w eksperymentach został on nazwany KMi-70.

Rozkład części mowy korpusu KMi-70 (rysunek 7.7) charakteryzuje się podobnym rozkładem, co całość korpusu po anotacji automatycznej (rysunek 7.6). Zawiera 12601 tokenów oraz 3415 zaanotowane terminy.

Porównanie wersji korpusu KMi-News

W trakcie analizy korpusu KMi-News pojawiły się trzy jego wersje o różnych wartościach wybranych cech. W tabeli 7.1 przedstawiono zestawienie analizowanych wersji w rozkładzie na liczbę terminów, liczbę dokumentów

Korpus	Tokeny	Dokumenty	POS	Terminologia	NE
KMi-News	62303	273	SCHUG	Nie	Nie
KMi-11	1132	11	Ręcznie	Tak	Tak
KMi-Gate	62303	273	GATE	Nie	Tak
KMi-70	12601	70	GATE	Tak	Tak

Tabela 7.1: Zakres anotacji wykorzystywanych wersji korpusu KMi-News

oraz właściwości anotacji. W zestawieniu wskazano, czy dana wersja korpusu obejmuje anotację części mowy (POS), terminów jednotokenowych (terminy) oraz nazw własnych (NE).

7.1.2 Korpus e-commerce

Polskojęzyczny korpus z dziedziny handlu elektronicznego nie istniał w chwili uruchamiania eksperymentów. Konieczne było zatem stworzenie własnego korpusu o charakterze reprezentatywnym. Wśród analizowych problemów biznesowych znaczące miejsce zajmują dostawcy sprzętu IT, np. Komputronik lub Centrum Komputerowe Znak. Oba podmioty posiadają serwisy internetowe z opisami produktów, nadają się więc do zgromadzenie pożądanego korpusu.

Analiza opisów produktów w serwisie Komputronik[2] prowadzi do wniosku, że serwis ten, niestety, nie nadaje się do stworzenia korpusu o odpowiedniej jakości, co najmniej z następujących powodów:

- opisy powtarzają się, tj. podobne produkty (ale nie takie same) posiadają dokładnie taki sam opis,
- teksty charakteryzują się albo stylem czysto marketingowym, np. "Jest to z pewnością przełomowy moment i proponowana, nowa konsola ustanowi nowe standardy łącząc ogromne możliwości z miniaturyzacją ...", albo zupełnym brakiem interpunkcji (teksty wielozdaniowe bez kropek, zdania złożone bez przecinków, itd.),
- wielkość korpusu jest znacznie mniejsza w porównaniu z serwisem Znak (1445 opisów w Komputronik do 6789 w Znak, co stanowi tylko 21%).

Analiza tekstów z serwisu Znak[3] nie wykazała aż tak znacznych przeciwskazań, chociaż ich jakość nadal pozostawia wiele do życzenia. Charakter

[2] `http://www.komputronik.pl/`
[3] `http://www.znak.pl/`

opisów jest często zbyt techniczny (np. lista specyfikacji). Biorąc jednak pod uwagę dziedzinę, jest to do zaakceptowania.

Korpus e-commerce powstał więc poprzez zgromadzenie wszystkich dostępnych opisów produktów IT w postaci tekstu w języku polskim, dnia 23 października 2007 roku w sklepie komputerowym Centrum Komputerowe Znak[4]. Operacja ta została przeprowadzona przy użyciu opracowanego oprogramowania przy pomocy języka transformacji XSL, jest powtarzalna i odporna na zmiany struktury witryny. Uzyskany korpus zawiera 6789 dokumentów, przy czym dokument zawiera kompletny opis jednego produktu. Stan opisów przedstawiony jest na dzień 23 października 2007 roku.

Zgromadzony korpus nie posiada żadnej informacji lingwistycznej, ani struktury domyślnego formatu anotacji. Transformacja do pożądanego formatu anotacji wraz z pozyskaniem informacji lingwistycznej jest przedmiotem anotacji automatycznej.

Anotacja automatyczna

Anotacja automatyczna została wykonana przy użyciu narzędzia SProuT (Piskorski i in., 2005). Wykorzystano dostępne w narzędziu standardowe gramatyki dla języka polskiego dla anotacji części mowy i morfologii.

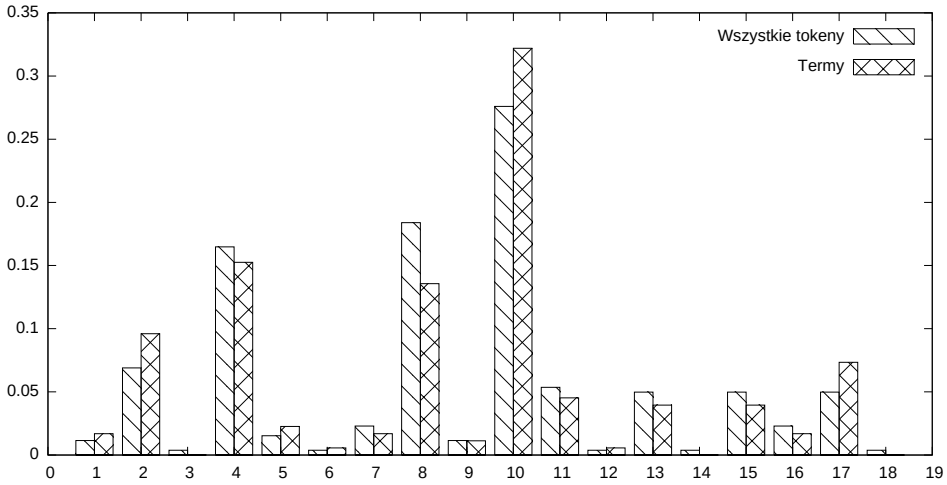
Anotacja lingwistyczna SProuT składa się zestawu informacji morfologicznych dotyczących tokenów. Informacja morfologiczna jest dostosowana dla języka polskiego i składa się m.in. z:

- rodzaju części mowy (wyłącznie w postaci podstawowej, np. rzeczownik, czasownik, przymiotnik, ...),
- deklinacji części mowy (np. dla rzeczownik odmiana przez przypadki, dla czasownika czas),
- liczba (pojedyncza, mnoga),
- rodzaj (męski, żeński, nijaki).

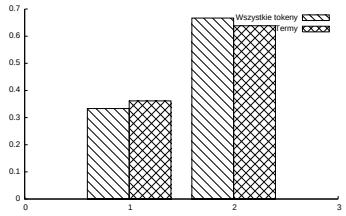
Złożoność morfologiczna języka polskiego powoduje, że wynikowa informacja jest znacznie bogatsza niż wykorzystany domyślny format anotacji. W celu oszacowania właściwości anotacji mających rzeczywisty wpływ na klasyfikację terminologii posłużono się następującym eksperymentem.

Przykładowe dokumenty zaanotowano pod kątem klasyfikacji terminu jednotokenowego. Następnie zliczono z całego korpusu oraz z tych tokenów, które zostały wskazane jako terminy, następujące elementy anotacji lingwistycznej:

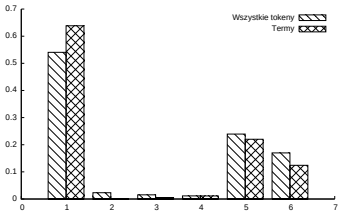
[4] `http://www.znak.pl/`



(a) Deklinacja



(b) Liczba

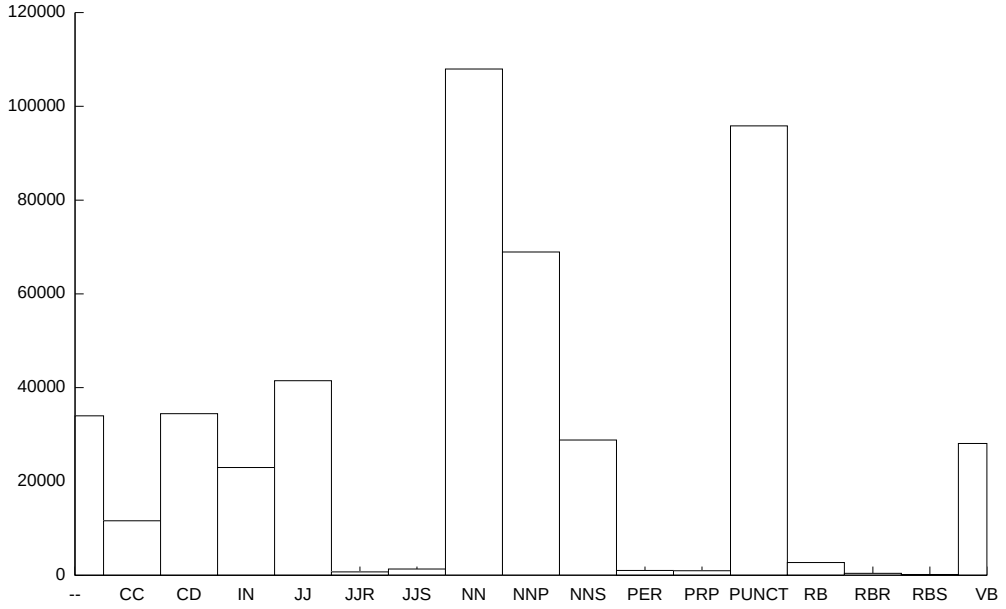


(c) Rodzaj

Rysunek 7.8: Rozkład badanych cech dla korpusu e-commerce w zależności od klasyfikacji terminologii

- deklinację przez przypadki,
- liczbę (pojedyncza/mnoga),
- rodzaj.

Dla każdej z tych właściwości sporządzono rozkład dla całości korpusu oraz dla próby wskazanych terminów. Oba te rozkłady porównano. Jeżeli rozkłady są podobne do siebie, oznacza to, że obserwowana cecha nie jest wrażliwa na klasyfikację terminologii. Oznacza to, że dany element anotacji lingwistycznej nie jest statystycznie ważny dla celów klasyfikacji i może zostać pominięty podczas analizy. Jeżeli natomiast rozkłady są znacząco różne od siebie, oznacza to, że rozkład danej cechy jest wrażliwy na klasyfikację terminologii i cecha ta powinna zostać uwzględniona.



Rysunek 7.9: Rozkład części mowy korpusu e-commerce

Eksperyment przeprowadzono metodą prototypu. Wyniki prototypu widoczne są na wykresach przedstawionych na rysunku 7.8. Dotyczą one odpowiednio deklinacji, liczby oraz rodzaju analizowanych tokenów.

Przedstawione porównania nie wskazują na znaczący wpływ jakiejkolwiek z badanych cech na klasyfikację terminologii. Tym samym można je pominąć. Dla zachowania spójności z przyjętym domyślnym formatem anotacji zdecydowano się zachować te same klasy części mowy.

W wyniku przeprowadzenia procesów anotacyjnych z wykorzystaniem narzędzia SProuT powstał korpus e-commerce zawierający 6789 dokumentów, w tym 481455 tokenów. Rozkład części mowy korpusu przedstawiono na rysunku 7.9.

Anotacja terminologii

Anotację terminologii przeprowadzono metodą ekspercką. Zorganizowane zostały specjalne warsztaty, w których uczestniczyło trzynastu ekspertów. Każdy z nich dokonywał anotacji terminologii zgodnie z przyjętym schematem zastosowanym w korpusie *KMi*.

Z łącznej liczby 6789 dokumentów do anotacji wybrano metodą losową 1000 dokumentów. Każdy ekspert do anotacji otrzymywał metodą losową jeden dokument. Dany dokument mógł zostać wylosowany tylko raz przez

Liczba anotacji	Liczba dokumentów
0	90
1	247
2	277
3	226
4	160

Tabela 7.2: Rozkład liczby anotacji eksperckich wśród 1000 dokumentów dla korpusu e-commerce

	KMi	e-commerce
Dziedzina	uniwersytet	produkty IT
Język naturalny	angielski	polski
Styl narracji	swobodny, e-mail	specyfikacje techniczne
Liczba dokumentów	273	6789
Liczba tokenów	62303	481455
Anotacja POS	GATE/ANNIE	SProuT
Byty nazwane	GATE/ANNIE	—
Format anotacji	OntoLT	OntoLT
Anotacja terminologii	ekspert	grupa ekspertów

Tabela 7.3: Porównanie korpusów *KMi* i *e-commerce*

danego eksperta. Każdy dokument mógł być zaanotowany maksymalnie przez 4 ekspertów. Strategia taka umożliwiała przeprowadzenie maksymalnie 4000 anotacji.

W wyniku przeprowadzonego warsztatu otrzymano 2183 anotacji eksperckich. Rozkład liczby anotacji eksperckiej wśród 1000 dokumentów przedstawiono w tabeli 7.2. W sumie zaanotowano 910 dokumentów, średnio każdy z tych dokumentów został zaanotowany przez 2,4 eksperta.

7.1.3 Podsumowanie

Wynikiem przeprowadzonych prac są dwa korpusy o zróżnicowanym charakterze. Poglądowe porównanie korpusów przedstawiono w tabeli 7.3.

7.2 Wyniki ekstrakcji terminologii

Eksperymenty przeprowadzone w fazie ekstrakcji terminologii obejmowały szereg metod ekstrakcji terminologii γ, m.in. metodę okna kontekstowego omówioną w rozdziale 5.

7.2.1 Zakres eksperymentów

W ramach eksperymentów wykorzystano dwa korpusy omawiane w sekcji 7.1, tj. KMi oraz e-commerce.

Następnie zaimplementowano poniżej wymienione metody ekstrakcji terminologii γ:

- standardową metodę TFIDF opartą na podobieństwie dokumentów,
- metodę TFIDF opartą na podobieństwie korpusów (wyróżnienie korpusu na tle ogólnego słownictwa),
- KFIDF (Xu i in., 2002),
- metodę wartości C/NC (Frantzi i in., 2000),
- klasyczną metodę opartą na modelu POS n-gram (Manning i Schutze, 1999).

Eksperymenty obejmowały metody uznane jako reprezentatywne, co wynika bezpośrednio z przeglądu metod z rozdziału 2. Dodatkowo zaimplementowano przedstawioną w ramach metamodelu metodę opartą na oknie kontekstowym.

Implementacja metod reprezentatywnych miała na celu zniwelowanie luki pomiędzy materiałami wtórnymi a pierwotnymi. Metody przedstawione w rozdziale 2. stanowią w toku pracy materiał wtórny. Niestety nie da się odtworzyć dokładnie takich samych warunków, w których metody te zostały ewaluowane. W związku z tym postanowiono odtworzyć metody i dokonać ewaluacji na zgromadzonych korpusach, uzyskując w ten sposób materiały pierwotne.

Każda z metod została następnie uruchomiona na tym samym zbiorze LA. W celach porównawczych format wyjściowy każdej z nich został ujednolicony do postaci:

```
<termin, prawdopodobieństwo>
```

Lista w taki sposób przygotowanych wyników dla każdej metody została porównana ze wskazaniami eksperta. Anotację ekspercką (tzw. klasyfikator wzorcowy) ustalono dla każdego z korpusów w następujący sposób:

- korpus KMi — warsztaty z udziałem eksperta z instytucji *Knowledge Media Institute* działającego w ramach *Open University*. Tym samym uzyskano anotację o wysokiej wiarygodności,

- korpus e-commerce — warsztaty z udziałem grupy studentów Informatyki Ekonomicznej na Wydziale Informatyki i Gospodarki Elektronicznej Uniwersytetu Ekonomicznego w Poznaniu. Uzyskany korpus wzorcowy zawiera anotację objętą weryfikacją jakości w postaci usunięcia wskazań skrajnych, wynikających ze słabej jakości anotacji niektórych z uczestników lub anotacji znacząco odstających od pozostałych.

7.2.2 Miary ewaluacji

Uzyskane wyniki zostaną przedstawione przy pomocy standardowych miar precyzji i zwrotu (Baeza-Yates i Ribeiro-Neto, 1999), które w przypadku ewaluacji ontologii wykorzystywane są najczęściej (Dellschaft i Staab, 2006).

Badane metody wykazały różną wrażliwość na wartość progu klasyfikacji k oznaczającego wartość prawdopodobieństwa, powyżej której analizowane wyrażenie klasyfikowane jest jako termin. Metody statystyczne osiągały swoje wartości optymalne przy niskich wartościach progu, podczas gdy metody oparte na modelach n-gram, w tym opracowana metoda okna kontekstowego, zachowywały się lepiej przy wyższych wartościach progu.

Z powodu różnej wrażliwości metod na próg klasyfikacji dokonano:

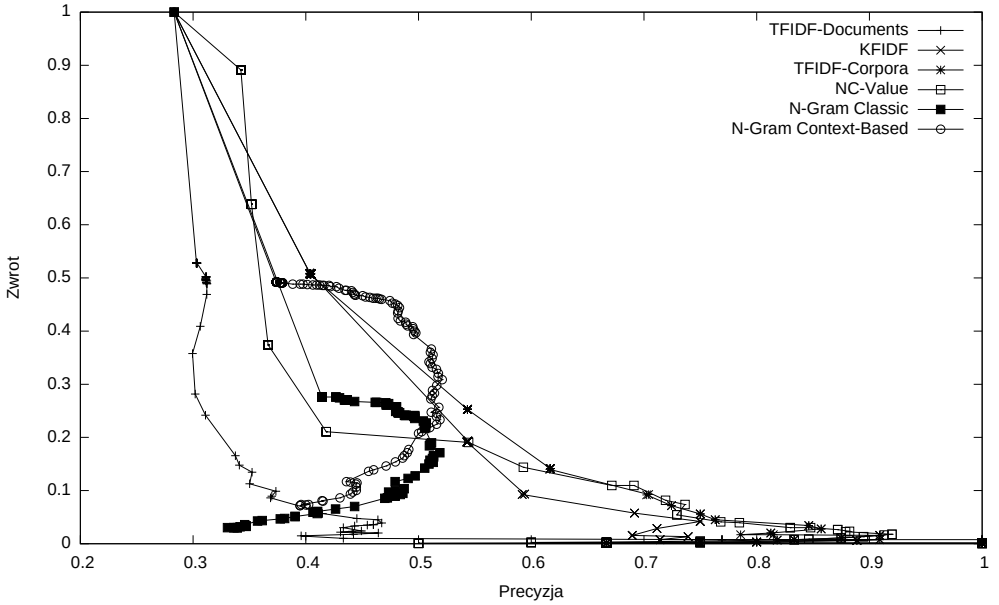
- dla korpusu KMi 28 pomiarów wszystkich metod przy następujących wartościach progu klasyfikacyjnego k: 0; 0,01; 0,02; ...0,1; 0,15; 0,2; 0,25; ...0,95,

- dla korpusu e-commerce 146 pomiarów wszystkich metod przy następujących wartościach progu klasyfikacyjnego k: 0; 0,001; 0,002; ...0,010; 0,011; ...0,05; 0,06; 0,07; ...0,1; 0,11; 0,12; 0,13; ...0,99.

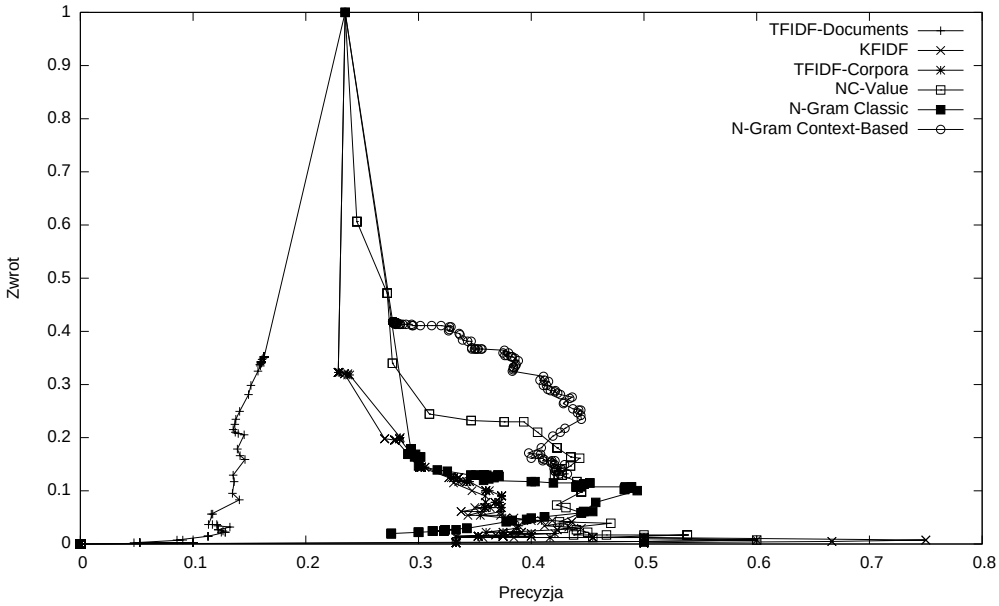
W celu wytrenowania metod opartych na modelu n-gram zbiór uczący podzielono na podzbiór trenujący i testujący. Dla obu korpusów zbiór trenujący stanowi 25% losowo wybranych dokumentów. Reszta zbioru uczącego stanowi podzbiór testujący. Motywacja i przesłanki takiego doboru wyjaśniono w sekcji 7.2.3.

Wyniki wszystkich przeprowadzonych pomiarów dla sześciu porównywanych metod przedstawia rysunek 7.10.

Metody oparte na częstości występowania (TFIDF, KFIDF oraz wartości C/NC) dały zróżnicowane wyniki. Zdecydowanie najgorzej wypada miara TFIDF obliczona klasycznie na zasadzie wyróżniania dokumentów wśród kolekcji. Wyróżnienie elementu zbioru D koncepcyjnie jest niepoprawne, dlatego wynik ten nie jest zaskoczeniem. Przewidując słabe wyniki klasycznego TFIDF zaimplementowano odmianę TFIDF, która wyróżnia cały korpus



(a) Korpus KMi



(b) Korpus e-commerce

Rysunek 7.10: Precyzja i zwrot wszystkich sześciu porównanych metod ekstrakcji terminologii na dwóch testowanych korpusach

D na tle ogólnego słownictwa danego języka. W tym celu metoda przetwarza rzeczowniki oraz przymiotniki z korpusu BNC (zgodnie z rozwiązaniem zaproponowanym przez Sintek i in. (2004)). Kolejną metodą jest KFIDF (Xu i in., 2002), która oparta jest na wyróżnieniu dziedziny reprezentowanej przez korpus. Obie metody prezentują wyniki znacznie lepsze od klasycznego TFIDF, z dobrym kompromisem pomiędzy precyzją i zwrotem. Ostatnia z miar statystycznych — metoda wartości NC (Frantzi i in., 2000) generuje równie dobre wyniki, w szczególności przy wyższych poziomach precyzji, dla których daje najwyższe wartości zwrotu. Oprócz miary wartości NC, zaimplementowano również miarę wartości C (Frantzi i in., 2000), lecz zrezygnowano z jej przedstawienia z powodu bardzo podobnych wyników. Jedyne odstępstwa w porównaniu do wartości NC polegały na kilku gorszych wynikach przy niektórych pomiarach. Biorąc pod uwagę, że metoda wartości NC jest rozwinięciem metody wartości C, fakt ten nie budzi zastrzeżeń.

Metoda oparta na klasycznym modelu n-gram osiąga wyniki zróżnicowane w zależności od wartości progu klasyfikacyjnego. W najlepszych przypadkach osiąga wyniki podobne do metod KFIDF oraz wartości NC, w pozostałych znacząco odstaje od pozostałych wyników.

Metoda okna kontekstowego przedstawiona w rozdziale 5. wyróżnia się najlepszymi wynikami, które osiągają najlepszy kompromis pomiędzy precyzją i zwrotem. Niestety, zarówno przy niskich, jak i wysokich poziomach progu klasyfikacji metoda generuje wyniki poniżej średniej.

Metody wykorzystujące model n-gram wyróżniają się charakterystycznym zakrzywionym kształtem. Kształt ten jest wynikiem spadku precyzji wraz ze spadkiem zwrotu w pewnym przedziale progu klasyfikacji. Wzrost progu klasyfikacji prowadzi do nadmiernego filtrowania pozytywnych wyników w stosunku do negatywnych wskazań. Jest to charakterystyczna właściwość metod opartych na modelu Markova, której nie zaobserwowano w przypadku metod statystycznych, gdzie wraz ze wzrostem progu klasyfikacji i spadkiem zwrotu, precyzja zachowuje się stabilnie. Własność ta powoduje, że istnieje tylko specyficzny przedział progu klasyfikacji, dla którego zastosowanie metod opartych na modelu n-gram jest wskazane. Problem ten jest klasycznym zagadnieniem optymalizacyjnym, w którym argumentem jest próg klasyfikacji, a wartością funkcji celu miara ewaluacji, np. połączona miara precyzji i zwrotu — miara F (Dellschaft i Staab, 2006).

W analizowanych przypadkach zakres wartości progu klasyfikacji dający najlepsze rezultaty wynosił dla metody klasycznego modelu n-gram 0,25 – 0,55, natomiast dla metody opartej na oknie kontekstowym 0,2 – 0,65.

W celu jednoznacznego stwierdzenia efektywności porównywanych metod policzono miarę F dla wszystkich pomiarów. Dla korpusu KMi zanotowano 168 pomiarów (28 pomiarów dla 6 metod). Fragment wyników dla korpu-

Pozycja	Metoda	Miara F
1	Nowa metoda	46,6187%
2	Nowa metoda	46,2396%
3	Nowa metoda	46,1312%
...	...	...
9	Wartość NC	43,3735%
...	...	...
13	KFIDF	42,7746%
...	...	...

Tabela 7.4: Najlepsze jednostkowe wyniki dla korpusu KMi wśród 168 pomiarów (28 eksperymentów dla 6. metod)

Pozycja	Metoda	Miara F
1	Nowa metoda	37,0186%
2	Nowa metoda	36,9347%
3	Nowa metoda	36,7041%
...	...	...
41	Wartość NC	34,9296%
...	...	...
153	TFIDF-Corpora	27,2823%
...	...	...
156	KFIDF	27,0661%
...	...	...

Tabela 7.5: Najlepsze jednostkowe wyniki dla korpusu e-commerce wśród 876 pomiarów (146 eksperymentów dla 6. metod)

su KMi posortowanych względem malejącej miary F przedstawia tabela 7.4. Dla korpusu e-commerce zanotowano 876 wyników (146 pomiarów dla 6 metod), których fragment według posortowanych malejąco miar F, przedstawiono w tabeli 7.5.

7.2.3 Dobór zbioru trenującego

Efektywność podejść opartych na uczeniu maszynowym jest podatna na wielkość zbioru trenującego dobranego do trenowania modelu. W celu oszacowania siły oraz kierunku wpływu tego parametru na metodę okna kontekstowego postanowiono przeprowadzić następujące eksperymenty.

Z obu badanych korpusów wydzielono zbiory uczące stanowiące podzbiór korpusów wynikający z przeprowadzonej klasyfikacji wzorcowej. Zbiór uczący należy następnie podzielić na podzbiór trenujący oraz testujący. Przyjęto system, w którym podział ten następuje w wartościach procentowych stanowiących wartość procentową całości zbioru uczącego. Na przykład, dla korpusu KMi-70, próg o wartości 10% oznacza wydzielenie metodą pseudolosową 7. dokumentów do zbioru trenującego i pozostałych 63. dokumentów do zbioru testującego.

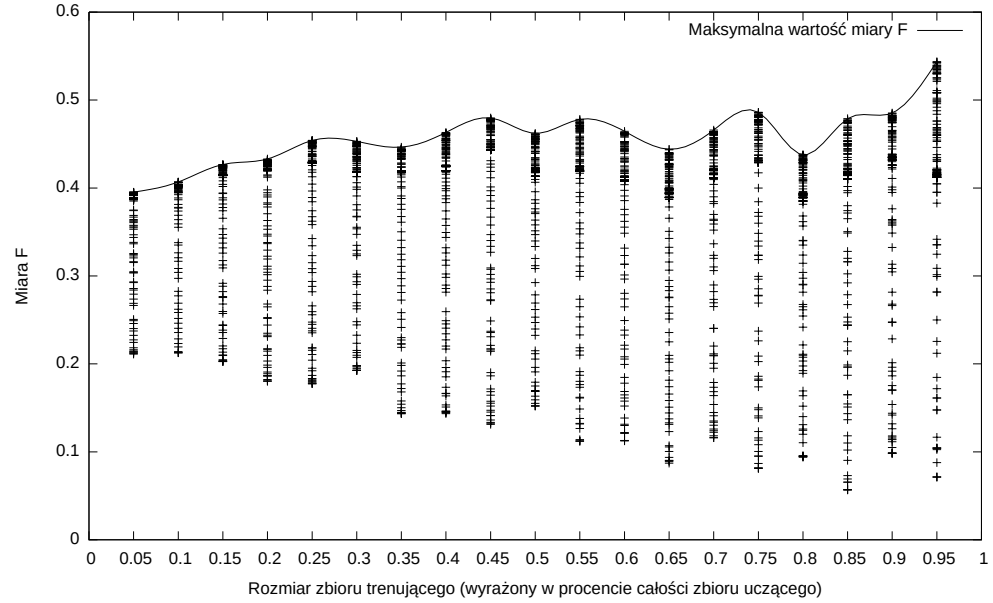
Wybrano 19 progów procentowych, tj. 5%, 10%, 15% ... 95%. Dla każdego progu przeprowadzono eksperymenty dla wszystkich 146. wartości progu klasyfikacyjnego, co dało łącznie 2774 pomiarów.

Wyniki przeprowadzonego eksperymentu przedstawione są na wykresie 7.11. Dla każdej wartości progu procentowego widoczny jest rozkład efektywności metody liczonej miarą F. Na wykresie dodatkowo połączono maksymalne wartości miary F dla każdego progu procentowego.

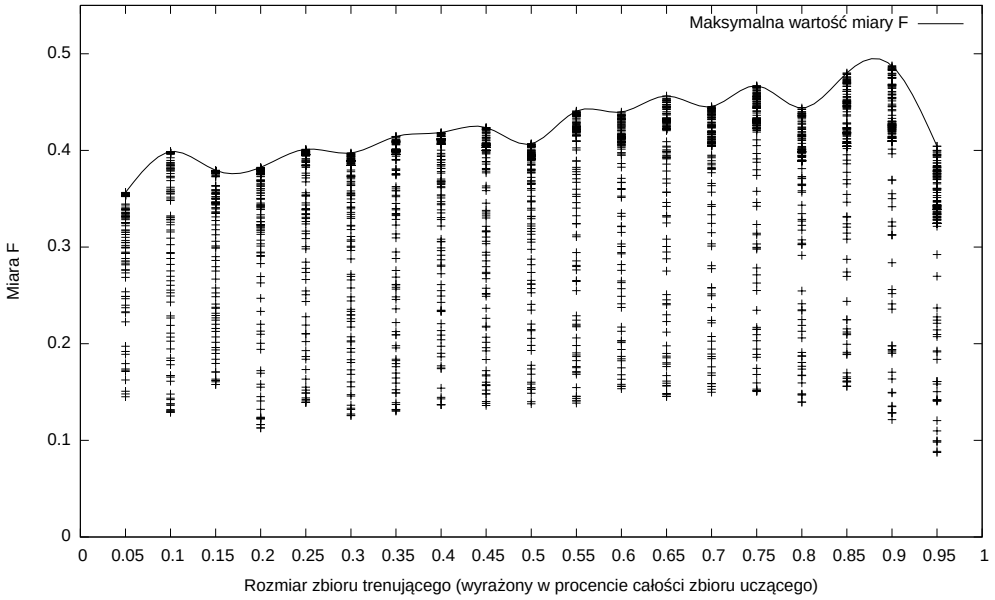
Wykres 7.11 pokazuje, że metoda ekstrakcji terminologii oparta na oknie kontekstowym przy odpowiednio zoptymalizowanym progu klasyfikacji osiąga efektywność w przedziale ok. 37–50%. Rozkład efektywności jest w miarę równomierny o tendencji lekko rosnącej, co gwarantuje stabilność jej działania. Istotną zaletą jest fakt, że przy niższych liczebnościach zbioru trenującego metoda nie traci drastycznie na efektywności. Często jest wadą metod opartych na uczeniu maszynowym.

W celu szczegółowego zbadania kształtowania się wartości precyzji i zwrotu, z danych przedstawionych na wykresie 7.11 postanowiono wybrać cztery reprezentatywne progi procentowe. Dobrany zbiór progów zawiera:

- pierwszą wartość progu, dla której metoda uzyskuje dobry wynik,
- wartość progu, dla której metoda uzyskuje najlepszy wynik,
- dwie wartości pośrednie.

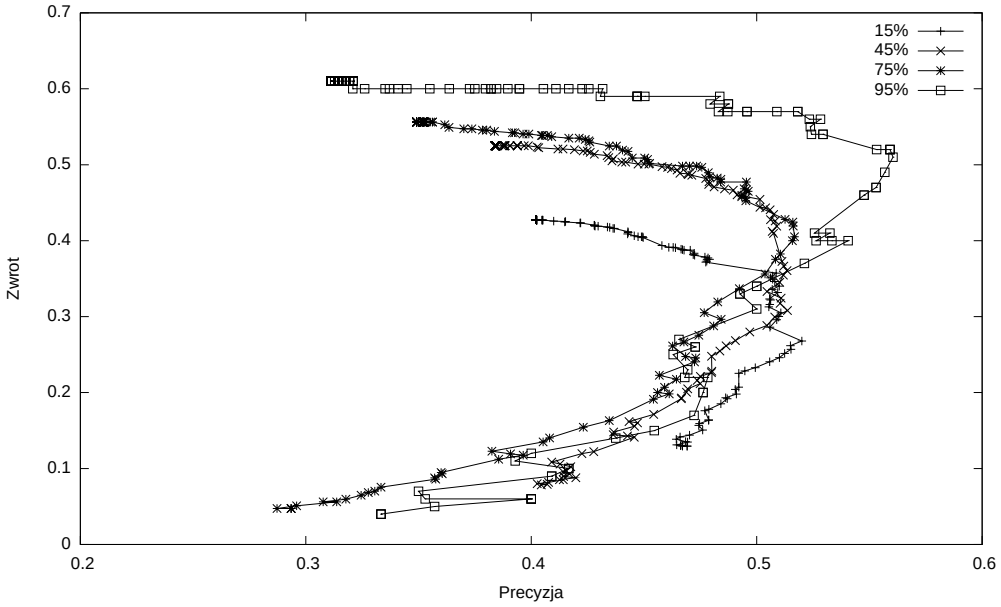


(a) Korpus KMi

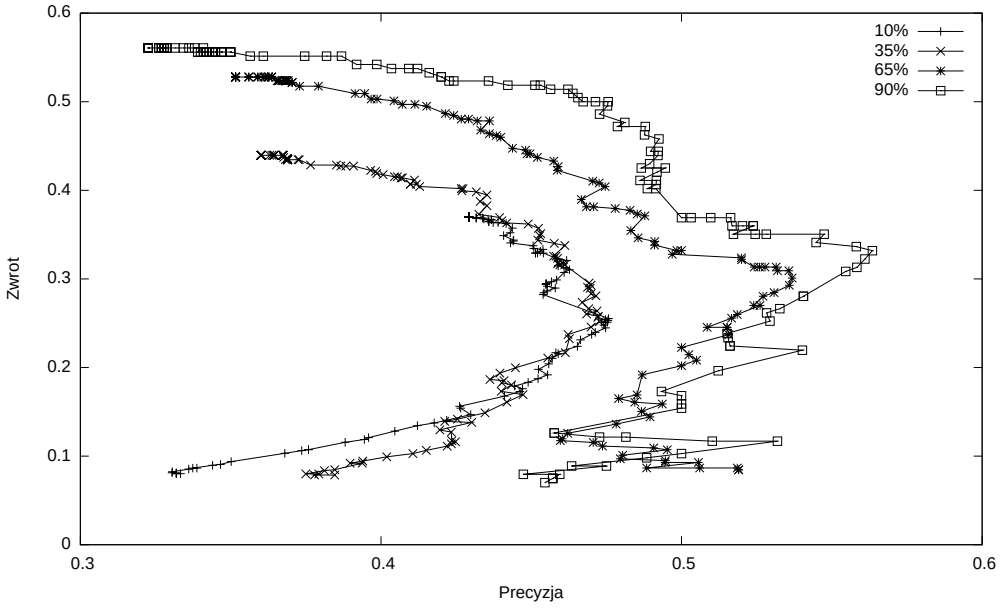


(b) Korpus e-commerce

Rysunek 7.11: Wpływ wielkości zbioru trenującego na efektywność metody ekstrakcji terminologii opartej na oknie kontekstowym

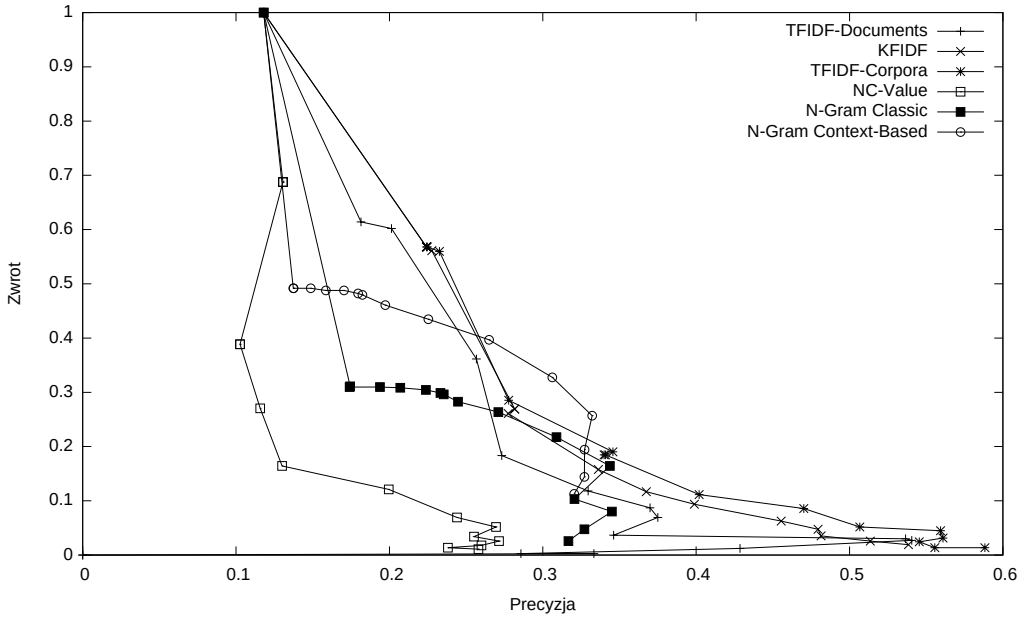


(a) Korpus KMi



(b) Korpus e-commerce

Rysunek 7.12: Szczegółowy wpływ wielkości zbioru trenującego na efektywność metody ekstrakcji terminologii opartej na oknie kontekstowym przy wybranych wielkościach procentowych zbioru uczącego



Rysunek 7.13: Precyzja i zwrot metody ekstrakcji terminologii opartej na oknie kontekstowym dla korpusu komputronik o bardzo niskiej jakości danych

Wyniki szczegółowego porównania miar precyzji i zwrotu dla wybranych progów procentowych przedstawiono na rysunku 7.12, który pokazuje stabilny wzrost wartości precyzji i zwrotu wraz ze wzrostem liczebności zbioru trenującego. Tendencja ta jest lekko zniekształcona przy wyższych wartościach progu klasyfikacji (ale już po osiągnięciu wartości optymalnej).

Na podstawie przeprowadzonych badań, do testowania efektywności metody wykorzystano próg 25%. Gwarantuje on uzyskanie dobrej efektywności, jednocześnie nie powoduje przetrenowania modelu.

7.2.4 Wnioski

Podejścia oparte na modelach Markova są wrażliwe na jakość danych. Stabilna, dobra jakość danych powoduje, że model jest lepiej wytrenowany. Niska jakość danych zarówno korpusu *KMi*, jak i korpusu *e-commerce* nie sprzyja dobrym wynikom przedstawionej metody. Pomimo tego przeprowadzone eksperymenty wskazują na uzyskanie modelu dobrze radzącego sobie z danymi o niskiej jakości.

Szczególnym przypadkiem korpusu o niskiej jakości danych jest korpus pozyskany z witryny sklepu komputerowego Komputronik, który został od-

rzucony w trakcie doboru zbioru D na rzecz korpusu pozyskanego ze sklepu Znak. W eksperymencie pomocniczym postanowiono dokonać ekstrakcji terminologii również na tym korpusie. Wyniki uzyskane na wykresie 7.13 są rezultatem eksperymentów na 25% zbiorze trenującym. Zbiór uczący został przygotowany na tym samym warsztacie i na podstawie tych samych reguł, co korpus pozyskany ze sklepu Znak. Przedstawione wyniki są na zdecydowanie niższym poziomie, najlepsze z nich nieznacznie przekraczają 30% miary F. Przeprowadzone badania dowodzą jednak ogólności metody ekstrakcji terminologii opartej na oknie kontekstowym.

Przegląd literatury z zakresu badań dowodzi, że samo porównanie wyników działania różnych metod ekstrakcji terminologii jest cenne i spotykane niezwykle rzadko. Dzieje się tak z powodów analizowanych w rozdziale 4., wynikających z różnej, niekompatybilnej ze sobą reprezentacji anotacji lingwistycznej i modelów ekstrakcji. Jedyne obszerne zestawienie metod ekstrakcji kolokacji i terminologii znaleźć można w Wermter i Hahn (2006), które obejmuje obszerną analizę 3. miar statystycznych. Samo zatem porównanie efektywności 6. metod jest cenne.

Porównanie wyników, uzyskanych z uruchomienia metod dla dwóch znacząco różnych korpusów, wskazuje na stabilność uzyskiwanych wyników. Pomimo że wartości precyzji i zwrotu dla tych samych metod w obu korpusach różnią się, to podstawowe cechy nie ulegają zmianie. Obie wartości miar efektywności są nieznacznie wyższe w przypadku korpusu KMi. Charakterystyczny, zakrzywiony kształt metody okna kontekstowego występuje w obu korpusach. Metody oparte na częstości występowania również zachowują podobne tendencje, np. zdecydowanie najgorsze wyniki metody opartej na mierze TFIDF lub wysoki poziom precyzji wyników uzyskanych przy pomocy metody wartości NC.

Ważnym czynnikiem sukcesu dla metody okna kontekstowego oraz metody opartej na mierze wartości NC jest fakt, że analizowane korpusy zawierają znaczącą liczbę terminów wieloczłonowych. Inne metody, ze względu na fakt, iż terminów wieloczłonowych nie przetwarzają, wiele tracą. Można zatem zarzucić, że słaba efektywność tych metod jest spowodowana brakiem funkcjonalności przetwarzania wieloczłonowości terminów. Zarzut ten można łatwo odrzucić analizując sposób konstrukcji metody wartości NC (sekcja 2.2.4 na stronie 40). Oryginalna metoda wartości NC nie przetwarza terminów jednotokenowych. W przedstawionym porównaniu do klasycznej metody wartości NC dodano funkcjonalność przetwarzania terminów jednotokenowych poprzez połączenie jej z metodą TFIDF wyróżniającą dokument wśród korpusu. Analizując konstrukcję porównanej metody NC można więc stwierdzić, że jest to metoda TFIDF rozszerzona o przetwarzanie terminów wielotokenowych przy pomocy metody wartości NC. Uzyskana me-

toda jest więc możliwie najlepszym przypadkiem dodania funkcjonalności przetwarzania terminów wielotokenowych do metody przeznaczonej wyłącznie do przetwarzania terminów jednotokenowych. W wielu zaobserwowanych przypadkach dodanie funkcjonalności przetwarzania terminów wielotokenowych prowadzi nawet do spadku efektywności metody. Zależność ta widoczna jest podczas porównania efektywności metody okna kontekstowego z metodą opartą na klasycznym modelu n-gram. Przy danych poziomach progu klasyfikującego k precyzja metody okna kontekstowego jest niższa.

Kolejnym rozpatrywanym wielokrotnie problemem jest liczebność zbioru uczącego. Niestety, pomimo ambitnych planów w postaci organizacji warsztatów z 30. uczestnikami w roli ekspertów oraz systemem weryfikacji jakości anotacji, udało się zgromadzić zbiór dokumentów ledwie wystarczający na przeprowadzenie rzetelnych badań. W celu zachowania obiektywności klasyfikatora wzorcowego starano się, aby anotacja ekspercka nie była wykonywana przez autora metod.

7.3 Wyniki ekstrakcji relacji

Eksperymenty przeprowadzone w fazie ekstrakcji relacji nietaksonomicznych obejmowały wyłącznie przedstawioną w rozdziale 6. metodę. Bezpośrednich konkurentów nie stwierdzono, tak więc metoda nie jest bezpośrednio porównywalna pod względem skuteczności. Niemniej jednak przedstawione są podstawowe miary precyzji i zwrotu dla samej metody.

7.3.1 Zakres i warsztat eksperymentów

W ramach eksperymentów wykorzystano korpus KMi przedstawiony w sekcji 7.1, ten sam, który zastosowano podczas eksperymentów z metodą ekstrakcji terminologii. Korpus KMi został zaanotowany przy użyciu narzędzia GATE oraz standardowych zasobów lingwistycznych ANNIE (Cunningham i in., 2007). Zbiór anotacji lingwistycznych LA został przygotowany zgodnie z metamodelem i jest tożsamy z modelem zastosowanym w OntoLT (Sintek i in., 2004). Pozyskano tylko płytką informację lingwistyczną, z tego powodu domyślny format anotacji zawierał jedynie sekcje $<text>$.

Ekstrakcji terminologii dokonano przedstawioną w ramach metamodelu metodą opartą na oknie kontekstowym (rozdział 5).

Ekstrakcja synonimów oraz pojęć została przeprowadzona zgodnie z metodą ϕ oraz δ przedstawionymi w metamodelu (3.4) oraz (3.5). W wyniku testu rozszerzenia powstał również zbiór instancji.

Ontologia wstępna

Wynikiem ekstrakcji terminologii, synonimów oraz pojęć jest tzw. ontologia wstępna, która została umieszczona w pojedynczym pliku owl i składa się z:

- 417 klas (elementów zbioru C), w tym np.: Colleague, PilotProject, Enterprise, University, ...,
- 579 instancji klas zdefiniowanych jako byty nazwane, tj. Person (pl: Osoba), Organization (pl: Organizacja), Location (pl: Lokalizacja), Date (pl: Data), Money (pl: JednostkaMonetarna), Address (pl: Adres),
- etykiety instancji, które odzwierciedlają ich realizacje w tekście.

Aksjomaty

Przygotowanie przykładowych reguł dla dziedziny objętej korpusem KMi oparto na wiedzy dziedzinowej eksperta[5]. Z powodu braku odpowiednich źródeł konieczne było przygotowanie zbioru A ręcznie. Wynik obserwacji oraz formalizacji reguł dziedzinowych przedstawiono na rysunku 7.14.

Reguły dziedzinowe zostały następnie sformalizowane w języku SWRL i umieszczone w jednym pliku. Zestaw reguł zawiera 25 reguł z rysunku 7.14 przedstawionych przy pomocy składni języka SWRL. W rozróżnieniu na predykaty lingwistyczne zestaw zawiera:

- 12 reguł ze standardowymi predykatami,
- 6 reguł z predykatami lingwistycznymi includesTerm,
- 6 reguł z predykatami lingwistycznymi coOccurInDocument,
- 1 reguła z predykatem lingwistycznym coOccurInTerm.

Zgodnie z rekomendacjami modelu ekstrakcji relacji, zarówno dziedzina jak i zasięg predykatów lingwistycznych, zostały ograniczone do rzeczywiście wykorzystanych typów danych.

Predykaty lingwistyczne

Zgodnie z założeniami metody ekstrakcji relacji, ontologia została wzbogacona o definicje oraz realizacje predykatów lingwistycznych. W wyniku tego procesu ontologia została poszerzona o następujące właściwości obiektowe:

[5]W pracach nad eksperymentami uczestniczyła dr Maria Vargas-Vera z Knowledge Media Institute.

Relacje standardowe:

```
AcademicStaff(x) => Professor(x)
AcademicStaff(x) => Doctor(x)
AcademicStaff(x) => Assistant(x)
AcademicStaff(x) => Mr(x) AND  worksAt(x,y) AND University(y)
AcademicStaff(x) => Mrs(x) AND  worksAt(x,y) AND University(y)
UniversityStaff(x) => worksAt(x,y) AND University(y)
Publisher(x) => publish(x,y) AND Organisation(x)
employs(x,y) => worksAt(y,x)
Employer(x) => employs(x,y) AND Organisation(x)
Employee(x) => Person(x) AND worksAt(x,y) AND Organisation(y)
Supervisor(x) => Person(x) AND supervise(x,y)
Visitor(x) => visited(x,y)
```

Relacje z wykorzystaniem predykatu coOccurInDocument:

```
worksAt(x,y) => Professor(x) AND University(y)
AND co-occurInDocument(x,y)
worksAt(x,y) => Doctor(x) AND University(y)
AND co-occurInDocument(x,y)
worksAt(x,y) => Assistant(x) AND University(y)
AND co-occurInDocument(x,y)
studiesAt(x,y) => Student(x) AND University(y)
AND co-occurInDocument(x,y)
collaborate(x,y) => Organisation(x) AND Organisation(y)
AND co-occurInDocument(x,y)
locatedIn(x,y) => Organisation(x) AND Location(y)
AND co-occurInDocument(x,y)
```

Relacje z predykatem coOccurInTerm:

```
hasTitle(x,y) => Person(x) AND Title(y) AND co-occurInTerm(x,y)
Title(x) = {Professor, Doctor, ...}
```

Relacje z predykatem includesTerm:

```
Professor(x) => Person(x) AND
includesTerm(x, oneOf(Professor,Prof,Prof.))
Doctor(x) => Person(x) AND includesTerm(x, oneOf(Doctor,Dr,Dr.))
Project(x) => includesTerm(x, "Project")
Department(x) => Organisation(x) AND includesTerm(x, "Department")
Institute(x) => Organisation(x) AND includesTerm(x, "Institute")
University(x) => Organisation(x) AND includesTerm(x, "University")
```

Rysunek 7.14: Przykładowy zbiór reguł dziedzinowych

próg	liczba
0.0	26304
0.1	5550
0.2	2136
0.3	1235
0.4	976
0.5	931
0.6	591
0.7	556
0.8	551
0.9	548
1.0	547

Tabela 7.6: Liczba predykatów lingwistycznych coOccurenceInDocument w zależności od progu klasyfikacji

- 110 predykatów lingwistycznych includesTerm,
- 28 predykatów lingwistycznych coOccurInTerm,
- 547-26304 predykatów lingwistycznych coOccurInDocument w zależności od progu klasyfikacji mierzonego miarą współwystępowalności Jaccarda.

Pierwsze dwa predykaty zostały stworzone na podstawie leksykalnych realizacji ich argumentów. Proces ten nie był złożony, a jego wynik był binarny. Predykat lingwistyczny coOccurInDocument wymagał wykorzystania miar współwystępowalności, np. zastosowanej miary Jaccarda. W toku eksperymentów wyliczona została miara współwystępowalności pomiędzy wszystkimi pojęciami, co spowodowało dużą złożoność obliczeniową i konieczność ich wyliczenia przed właściwym procesem ekstrakcji. Szczegółowe dane dotyczące liczby predykatów lingwistycznych coOccurInDocument zostały przedstawione w tabeli 7.6.

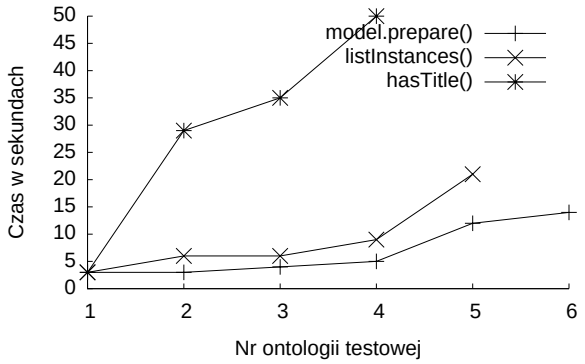
7.3.2 Dostosowanie ontologii

Ontologia stworzona dla całości korpusu KMi zawiera znaczną liczbę obiektów oraz dodatkowych predykatów lingwistycznych w postaci właściwości obiektowych. Testy[6] różnych wersji ontologii (tabela 7.7), których wyniki

[6]Testy przeprowadzono na komputerze: Linux 2.6, Centrino 1.8, 1 GB RAM, JVM 1.6 512 MB.

Nr	Liczba klas	Liczba relacji	Liczba instancji
1	2	wszystkie	2
2	wszystkie	wszystkie	2
3	wszystkie	wszystkie	50
4	wszystkie	wszystkie	100
5	wszystkie	wszystkie	200
6	wszystkie	wszystkie	250
7	wszystkie	wszystkie	300
8	wszystkie	wszystkie	wszystkie

Tabela 7.7: Charakterystyka ośmiu wersji ontologii testowej z różną liczbą klas, instancji oraz relacji



Rysunek 7.15: Wydajność mechanizmu wnioskującego Pellet mierzona czasem wykonania trzech podstawowych funkcjonalności, tj. przygotowania modelu ontologii, wylistowania wszystkich instancji oraz wyszukania instancji z zadaną relacją nietaksonomiczną

przedstawione zostały na rysunku 7.15, wykazały, że wykorzystywany mechanizm wnioskujący (Pellet) nie jest w stanie poprawnie przetworzyć całości ontologii.

Brak pomiarów na rysunku 7.15 dla niektórych wersji ontologii oznacza nieukończenie testu w skończonym czasie. Testy przeprowadzane przez samych autorów (Sirin i in., 2007) wykazały znacznie dalej idącą wydajność, lecz nie wzięto pod uwagę tak złożonej konfiguracji, jaka jest niezbędna do przeprowadzenia niniejszych eksperymentów. Połączenie technologii: Jena, Pellet, reguły SWRL, importów ontologii i ontologii składającej się z 417 klas, 579 instancji oraz do 26304 właściwości obiektowych, powoduje niewydolność mechanizmu wnioskującego.

W związku z problemami wydajnościowymi konieczne było przeprowadzenie operacji dostosowania modelu, która polega na usunięciu z ontologii tych obiektów, które nie mają żadnego wpływu na przeprowadzany test.

Wyniki testu wydajnościowego wskazały na największą wrażliwość liczby obiektów ontologii oraz rodzaju wykonywanych operacji. Ponieważ rodzaj wykonywanych operacji nie zmienia się, optymalizacji dokonano w liczbie obiektów ontologii, głównie liczbie klas oraz instancji.

Pierwsza operacja dostosowania polega na podzieleniu eksperymentów na grupy testów zgodnie z zaproponowanym podziałem na predykaty lingwistyczne. Zgodnie z nim, testy zostały przeprowadzone na trzech oddzielnych grupach predykatów, czyli includesTerm, coOccurInTerm oraz coOccurInDocument.

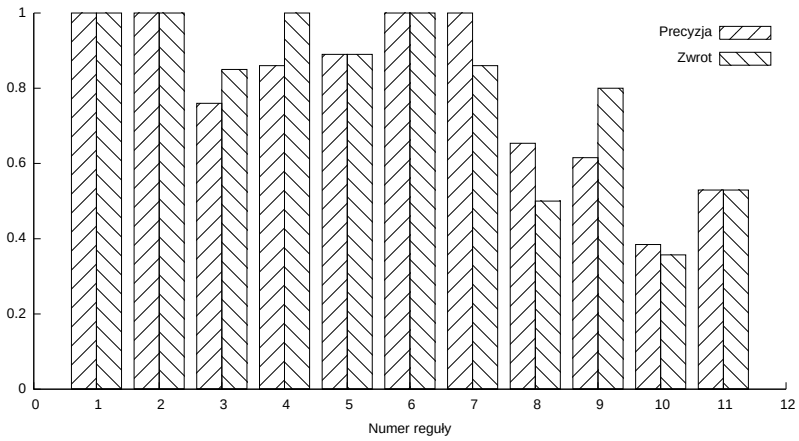
W testach dotyczących predykatów lingwistycznych includesTerm oraz coOccurInTerm usunięte zostały wszystkie instancje, które nie były argumentem predykatów lingwistycznych. Operacja ta nie miała wpływu na wyniki testów, jednocześnie redukując liczbę instancji w teście includesTerm do 82 oraz w teście coOccurInTerm do 28. Różnica pomiędzy liczbą 110 (liczbą właściwości obiektowych relacji includesTerm) a 82 wynika z faktu, że niektóre instancje współwystępują z więcej niż jednym terminem.

W testach dotyczących predykatu lingwistycznego coOccurInDocument dostosowanie ma miejsce na poziomie każdej testowanej reguły. Jest to spowodowane tym, że reguły są na tyle różne od siebie, a jednocześnie razem dość dobrze pokrywają dziedzinę, że wspólne ich dostosowanie nie doprowadziłoby do znaczącej redukcji liczby instancji. Z tego powodu potencjalnie interesujący podzbiór instancji jest specyficzny dla konkretnej reguły. Na przykład reguła *collaborate* potrzebuje wyłącznie instancji typów *Organisation* oraz *University*, co oznacza m.in. usunięcie wszystkich instancji typu *Person* i redukcję liczby instancji o połowę.

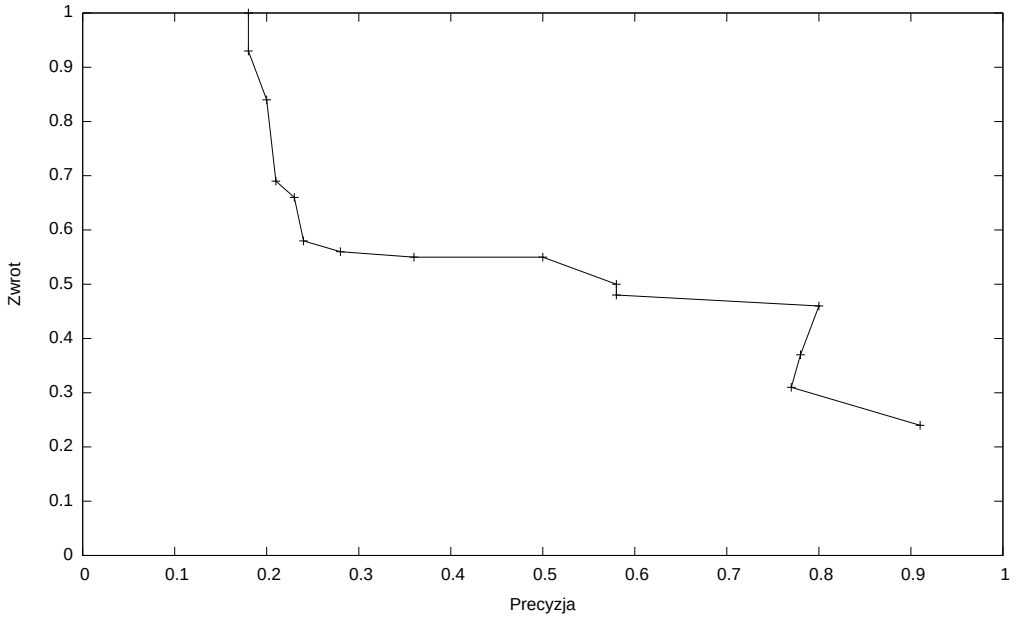
Reguły bez predykatów lingwistycznych nie zostały przetestowane z powodów wydajnościowych. Predykaty wykorzystywane w tych regułach dotyczą znaczącej części wszystkich instancji, a wstępne szacunki wskazały, że konieczne byłoby zastosowanie ontologii z ok. 500. instancjami, co jest niemożliwe. Ponadto reguły te były bezpośrednio związane z wynikami reguł opartych na predykatach lingwistycznych. Wnioskowanie na tych regułach musi zatem obejmować również wcześniej uzyskane wyniki, co dodatkowo zwiększa poziom złożoności modelu i problemy wydajnościowe.

7.3.3 Miary ewaluacji

Wynikiem przeprowadzenia eksperymentów są nowe relacje nietaksonomiczne, które zostają dodane do zbioru NTR (3.8). Wyniki te są porównywane



Rysunek 7.16: Precyzja i zwrot metody ekstrakcji relacji w rozkładzie na reguły. Reguły 1-7 zawierają predykaty lingwistyczne includesTerm oraz coOccurInTerm. Reguły 8-11 zawierają predykat lingwistyczny coOccurInDocument ze zoptymalizowanym progiem.



Rysunek 7.17: Precyzja i zwrot reguł z predykatem lingwistycznym współwystępowalności (8–11) opartych na 26 progach klasyfikacyjnych (0; 0,01, 0,02, . . . 0,1; 0,12; . . . 0,2; 0,25; 0,3; 0;33; 0,4; 0,5; . . . 1)

do zbioru wskazań eksperta opracowanego podczas serii warsztatów, na których dokonano analizy korpusu D oraz wskazano listę aksjomatów dziedzinowych. Do uczestnictwa w warsztatach zaproszono eksperta z instytucji *Knowledge Media Institute* (Open University, Wielka Brytania). Porównania uzyskanych wyników ze wskazaniami eksperta dokonano przy pomocy standardowych miar precyzji i zwrotu.

Wyniki w formie wykresów zaprezentowane są na dwóch rysunkach. Rysunek 7.16 przedstawia miary precyzji i zwrotu dla każdej reguły. Reguły zawierające predykaty lingwistyczne includesTerm oraz coOccurInTerm (reguły 1–7) charakteryzują się wynikiem binarnym i nie są zależne od progu prawdopodobieństwa. Wszystkie wartości precyzji i zwrotu dla tej grupy reguł kształtują się powyżej 75%. Pozostałe reguły (8–11) odnoszą się do relacji z predykatem lingwistycznym coOccurInDocument z progiem klasyfikacji relacji współwystępowalności ustawionym na poziomie dającym najlepszy kompromis pomiędzy precyzją a zwrotem.

Wykres przedstawiony na rysunku 7.17 pokazuje szczegółowe wyniki dla reguł z predykatem lingwistycznym coOccurInDocument dla 26 wartości progu klasyfikacyjnego (0; 0,01, 0,02, ... 0,1; 0,12; ... 0,2; 0,25; 0,3; 0;33; 0,4; 0,5; ... 1).

7.3.4 Wnioski

Główną zaletą przedstawionej metody ekstrakcji relacji jest zwolnienie użytkownika z konieczności ręcznego tworzenie skomplikowanych reguł lingwistycznych. Użytkownik nie jest zobligowany do dostarczenia nawet najmniejszej liczby przykładów w celu uruchomienia mechanizmu ekstrakcji. Przeprowadzona ewaluacja dowodzi, że uzyskane wyniki są satysfakcjonujące (miara F ponad 75%).

Decyzja o wykorzystaniu metody sprzężenia zwrotnego zależna jest jednak od konkretnej aplikacji oraz ograniczeń użytkowników. Jeśli tworzenie zbioru reguł lingwistycznych nie jest problemem, to rozwiązania omówione w rozdziale poświęconym przeglądowi obecnego stanu wiedzy prawdopodobnie przyniosą lepsze wyniki. Wydaje się jednak, że tworzenie reguł lingwistycznych wiąże się w większości przypadków ze zbyt dużym nakładem pracy.

Przeprowadzone eksperymenty obarczone były dużą złożonością obliczeniową. W związku z tym nie udało się powiązać ze sobą wyników uzyskanych w osobnych testach. Znaczną grupę problemów przysporzyły same mechanizmy wnioskujące, które na obecnym etapie rozwoju nie pozwalają na zbyt dużą liczbę obiektów ontologii. Spowodowało to konieczność dostosowywanie ontologii, co znacznie wydłużyło proces dochodzenia do wyników.

Rozdział 8

Zakończenie

Motywacja podjętych badań wynika z omawianych w rozdziale 1. problemów natury biznesowej oraz badawczej. Problem biznesowy zdefiniowany został w dziedzinie handlu elektronicznego i dotyczy braku współdzielonych warstw pojęciowych (por. sekcje 1.1.1 oraz 1.1.2). Problem badawczy został określony przez problemy niedostępności i nieadekwatności ontologii (por. sekcję 1.1.3). Wynikiem motywacji jest zakres pracy wraz z założonymi celami oraz tezami.

W niniejszym rozdziale osiągnięte w toku prac wyniki zostaną porównane z założonymi na wstępie celami oraz tezami pracy. Uzyskane wyniki zostaną następnie sklasyfikowane jako wyniki metodologiczne oraz poznawcze. Klasyfikacja wyników pozwoli również na podsumowanie głównych kierunków przyszłych badań.

8.1 Dowód

Motywacja podjęcia badań w dziedzinie elektronicznego handlu wyznaczyła trzy cele pracy:

1. Opracowanie uogólnionej metody uczenia ontologii z tekstu.
2. Opracowanie metody ekstrakcji terminologii dla uczenia ontologii z tekstu.
3. Opracowanie metody ekstrakcji relewantnych relacji nietaksonomicznych dla uczenia ontologii z tekstu minimalizujących udział eksperta dziedzinowego.

Pierwszy, ogólny cel pracy został osiągnięty i przedstawiony w rozdziale 3, w szczególności poprzez definicję równania 3.1. Wprowadza on ogólną metodę uczenia ontologii z tekstu wyjaśniającą przy pomocy metamodelu elementy

procesu oraz funkcje przejścia pomiędzy nimi. Elementy metamodelu (modele ekstrakcji poszczególnych zadań) zostały rozwinięte w rozdziałach kolejnych (rozdziały 4., 5., 6.).

Drugi cel pracy został osiągnięty i zaprezentowany w rozdziale 5., w którym opisano opracowaną metodę ekstrakcji terminologii z użyciem okna kontekstowego. Przedstawiona metoda jest rozwinięciem modelu ekstrakcji terminologii zgodnie z metamodelem (równanie 3.3).

W rozdziale 6. opracowano metodę ekstrakcji relacji z wykorzystaniem sprzężenia zwrotnego, będącą rozwinięciem modelu ekstrakcji relacji nietaksonomicznych zgodnie z metamodelem (równanie 3.8). Cele drugi i trzeci są celami szczegółowymi i stanowią rozwinięcie celu głównego.

Cele pracy doprowadziły do postawienia następujących tez pracy (por. sekcję 1.2.3):

1. Uogólniona metoda umożliwia przeprowadzenie procesu uczenia ontologii z tekstu w języku angielskim oraz polskim. Opisuje również w sposób abstrakcyjny elementy procesu, zależności pomiędzy elementami oraz wykorzystywane klasy zasobów.

2. Nienadzorowana metoda ekstrakcji terminologii wykorzystująca dynamiczne okno kontekstowe jest bardziej efektywna niż klasyczne metody ekstrakcji terminologii dla uczenia ontologii z tekstu wykorzystujące podejścia lingwistyczno-statystyczne lub klasyczny model n-gram.

3. ykorzystanie sprzężenia zwrotnego pomiędzy aksjomatami dziedzinowymi i informacją lingwistyczną prowadzi do zmniejszenia wymagań przedmiotowych oraz ilościowych w ekstrakcji relacji nietaksonomicznych dla uczenia ontologii z tekstu.

Wszystkie trzy tezy pracy zostały zweryfikowane pozytywnie poprzez:

1. Dowód ogólności opracowanej metody uczenia ontologii z tekstu przedstawiony został w rozdziałach 3. oraz 7. Rozdział o metamodelu wyznacza warstwę teoretyczną, na podstawie której można w sposób powtarzalny przeprowadzać proces uczenia ontologii z tekstu. Dodatkowo w rozdziale poświęconym eksperymentom, wykorzystano dwa znacząco różne korpusy do budowy ontologii zgodnie z metamodelem.

2. Dowód tezy o metodzie ekstrakcji terminologii przedstawiono w rozdziale 7. Na podstawie dwóch znacząco różnych korpusów osiągnięto wyniki lepsze niż dotychczas stosowanych metodach opartych na podejściach lingwistyczno-statystycznych lub klasycznym modelu n-gram. W opisanych eksperymentach uzyskano lepszą efektywność mierzoną połączoną miarą precyzji i zwrotu, co dowodzi tezie.

3. Teza o metodzie ekstrakcji relacji została dowiedziona w wyniku przeprowadzenia eksperymentów opisanych w rozdziale 7. Uruchomienie metody ekstrakcji relacji przy użyciu sprzężenia zwrotnego nie wymagało udziału eksperta, a zatem wymagania przedmiotowe oraz ilościowe zostały zmniejszone w porównaniu z obecnie stosowanymi metodami (rozdział 2). Ponadto osiągnięte miary precyzji i zwrotu przekraczające 75% nie odbiegają od standardów wyznaczonych przez inne metody wymagające znacznego udziału eksperta.

Opracowane metody, sposoby ich weryfikacji oraz uzyskane wyniki zostały docenione przez recenzentów krajowych i międzynarodowych konferencji naukowych. Podsumowanie przeglądu stanu wiedzy oraz metamodel zostały wydane jako rozdział książki. W szczególności opublikowano następujące prace:

- Zarys oraz specyfika metod uczenia ontologii z tekstu dla języka polskiego (Wisniewski, 2006).
- Podsumownie przeglądu obecnego stanu wiedzy oraz metamodel (Abramowicz i Wisniewski, 2008a).
- Metamodel oraz architektura dla języka polskiego (Abramowicz i Wiśniewski, 2008).
- Metoda ekstrakcji terminologii oparta na oknie kontekstowym (Abramowicz i Wisniewski, 2008b).
- Metoda ekstrakcji relacji oparta na sprzężeniu zwrotnym (Abramowicz i in., 2008).

8.2 Wyniki

Podstawowym osiągnięciem niniejszej pracy są wyniki metodologiczne. Jednak wyniki uzyskane w toku przeprowadzonych prac mają charakter nie tylko metodologiczny, lecz również poznawczy. Dążenie do poznawczych wyników na gruncie nauk technicznych jest błędem. Nie jest natomiast błędem wykorzystanie wyników poznawczych w sposób pośredni, w celu uzyskania wyników metodologicznych. W tym sensie zastosowano w pracy metamodel, który jest de facto wynikiem poznawczym (model). Pozwala on jednak dojść do wszystkich pozostałych metodologicznych celów pracy (wszystkie trzy metody).

Do metodologicznych (głównych) wyników pracy należą:

1. Uogólniona metoda uczenia ontologii z tekstu (rozdział 3.).
2. Metoda ekstrakcji terminologii wykazująca na zbiorach testowych lepszą efektywność od obecnie istniejących metod (rozdział 5.).
3. Metoda ekstrakcji relacji wykorzystująca sprzężenie zwrotne (rozdział 6.).
4. Stworzenie metod oraz narzędzi, które pozwalają w sposób powtarzalny na wykonanie zadań ekstrakcji (rozdział 7.).

Główne wyniki poznawcze wraz z powiązanymi wynikami metodologicznymi to:

1. Systematyka metod uczenia ontologii z tekstu — prowadzi do opracowania metamodelu, a więc uogólnionej metody uczenia ontologii z tekstu.
2. Metamodel — prowadzi do opracowania uogólnionej metody uczenia ontologii z tekstu.
3. Zebranie i formalizacja dwóch zróżnicowanych korpusów — prowadzi do przeprowadzenia dowodów dla wszystkich trzech metod (weryfikacja jakości metod).
4. Porównanie efektywności różnych metod ekstrakcji terminologii — prowadzi do przeprowadzenia dowodu tezy związanej z metodą ekstrakcji terminologii.

8.3 Korzyści dla dziedziny handlu elektronicznego

Uzyskane wyniki w postaci metod oraz narzędzi pozwalają na tworzenie ontologii, czyli wspólnych warstw pojęciowych. Rozpatrywane zagadnienia związane z rynkiem handlu elektronicznego otrzymują zatem narzędzia, które umożliwią bardziej dokładną i precyzyjną wymianę informacji. Same ontologie jednak nie wystarczą, ponieważ stanowią tylko jeden z fundamentów realizacji rozpatrywanych scenariuszy biznesowych (por. sekcja 1.1.1 na stronie 3). Wskazują na to badania z zakresu Sieci Semantycznej, których autorzy, oprócz ontologii, podkreślają kluczową rolę agentów oraz infrastruktury Berners-Lee i in. (2001).

Podstawowe korzyści dla dziedziny handlu elektronicznego skupiają się na:

- dokładniejszym wyszukiwaniu informacji,
- efektywniejszym filtrowaniu informacji,
- zwiększeniu przejrzystości rynku, a więc zmniejszeniu skutków asymetrii informacji Akerlof i in. (2001),
- zwiększeniu poziomu rozwoju usług świadczonych przez rynek handlu elektronicznego,
- zwiększeniu stopnia automatyzmu negocjacji warunków świadczenia usług lub parametrów produktów.

Wyszukiwanie konkretnych produktów lub usług staje się mniej czasochłonne, zwłaszcza wtedy, gdy nie korzysta się z serwisów znanych dostawców lub poszukiwany produkt czy usługa posiada zaawansowane warunki brzegowe (np. konkretne wartości cech). Problemy wieloznaczności terminów oraz niejednomianowości pojęć nie istnieją. Wyszukanie konkretnego urządzenia naręcznego spowoduje odnalezienie odpowiednich urządzeń z różnych klasyfikacji, m.in. "urządzenie naręczne", lecz również: "PDA", "handheld" lub "telefon". W tym celu, metody uczenia ontologii z tekstu przedstawione w niniejszej pracy muszą być wykorzystane przez systemy wyszukiwawcze podmiotów z rynku handlu elektronicznego (jeśli stosują własne) lub przez potentatów z rynku wyszukiwarek w Internecie, np. Google.

Zastosowanie ontologii umożliwia efektywniejsze filtrowanie informacji. Mechanizmy budowania profilu użytkownika mogą być niezależne od konkretnych dostawców. Profil użytkownika skonstruowany w jednym sklepie ma zastosowanie w innym miejscu. W tym celu, metody uczenia ontologii z tekstu przedstawione w niniejszej pracy muszą być wykorzystane przez odpowiednio przygotowane agenty, których właścicielami mogą być podmioty obecne na rynku handlu elektronicznego lub właściciele systemów filtrujących. Kluczowe znaczenie ontologii dla potrzeb filtrowania zostało szczegółowo przedstawione w Abramowicz (2008).

Przejrzystość rynku jest znacząco zwiększona. Porównanie produktów odbywa się bez dodatkowych utrudnień spowodowanych niejednomianowością pojęć. Serwisy porównawcze wreszcie prawidłowo klasyfikują produkty, a inteligentne agenty potrafią rozumieć przedstawiane towary. Prowadzi to do pełniejszej informacji o rynku, a indywidualne decyzje konsumentów są w mniejszym stopniu dotknięte skutkami asymetrii informacji Akerlof i in. (2001).

Poziom rozwoju usług świadczonych przez elektroniczny handel osiąga dojrzałość umożliwiającą automatyzację procesów, np. zautomatyzowanie

transakcji kupna produktów z systemem monitorującym stan zapasów. W tym celu, oprócz wykorzystania wyników niniejszej pracy, konieczne jest stworzenie odpowiedniej infrastruktury sprzętowo-programowej.

Ontologie umożliwiają negocjację warunków świadczenia usług (np. umów SLA) lub parametrów produktów. Inteligentne agenty potrafią w sposób półnadzorowany prowadzić rozproszone negocjacje z wykorzystaniem współdzielonego modelu pojęciowego. W tym celu, wyniki niniejszej pracy muszą być wykorzystane przez narzędzia z dziedziny elektronicznych negocjacji.

Przedstawione w niniejszej pracy badania nie stanowią oczywiście rozwiązania wszystkich problemów związanych z rynkiem handlu elektronicznego. Przy pomocy opracowanych metod nie da się stworzyć wspólnej warstwy pojęciowej w każdych warunkach. Wszystkie ewaluowane metody są podatne na jakość każdego z elementu metamodelu (równanie 3.1 na stronie 104). Niska jakość tekstów w korpusie, a więc np. luźne stosowanie reguł gramatycznych dla danego języka naturalnego, powoduje znaczące trudności w uzyskaniu wysokiej efektywności. Błędy popełnione na etapie ekstrakcji pojęć przyczyniają się do znaczącego spadku efektywności metod ekstrakcji relacji. Uzyskana efektywność metod na poziomie 40-50% miary F jest nadal daleka od oczekiwań i potrzeb w celu wykorzystania ich do budowania ontologii w biznesie. Niniejsze badania stanowią jednak kolejny etap w przybliżaniu efektywności metod uczenia ontologii z tekstu do możliwie najwyższej efektywności, zwłaszcza w dziedzinie handlu elektronicznego.

8.4 Przyszłe badania

Przedstawione badania wskazały na wiele możliwych kierunków rozwoju, które wynikają zarówno z głównych (metodologicznych) wyników pracy, jak i z wyników pobocznych (poznawczych).

Przede wszystkim systematyka metod wykazała brak dobrych, uniwersalnych metod do zastosowania bez względu na dziedzinę. Obecnie stosowane uniwersalne metody charakteryzują się niską efektywnością, a poszczególne dziedziny wymagają specyficznych podejść. Ponadto większość metod opracowanych jest dla konkretnego języka naturalnego. Zakres przedmiotowy badań nad uczeniem ontologii z tekstu jest szeroki i obejmuje wiele dziedzin, tj. uczenie maszynowe, przetwarzanie tekstu naturalnego, ekstrakcję informacji i reprezentację wiedzy. Podjęcie badań wymaga zatem umiejętności interdyscyplinarnych, pokazuje również jak trudno jest opracować uniwersalne i dobre metody.

Przyszłe kierunki badań nad przedstawioną w ramach metamodelu metodą ekstrakcji terminologii oscylują wokół stosowania różnych metod wy-

gładzania. Aczkolwiek problem rzadkości przy metodzie okna kontekstowego nie jest znaczący, efektywne metody wygładzania mogłyby nieznacznie poprawić efektywność metody. Ponadto optymalizacja wskazanych parametrów w różnych kontekstach wykorzystania metody przyniosłaby z pewnością lepsze wyniki w poszczególnych dziedzinach.

Wraz z rozwojem efektywności mechanizmów wnioskujących opartych na języku OWL, badania nad metodą ekstrakcji relacji nabrałyby szerszego wymiaru. Niestety, obecnie przeprowadzone eksperymenty pozwalają tylko szacować jak mechanizm sprzężenia zwrotnego zachowywałby się przy innych korpusach o zwiększonej liczbie dokumentów lub tokenów. Zwiększenie efektywności mechanizmów wnioskujących pozwoliłoby zatem na znacznie szerszą ewaluację metody.

Wynikiem badań z dziedziny ekstrakcji informacji dla języka polskiego są zasoby lingwistyczne oraz narzędzia przeznaczone dla języka polskiego (Piasecki i Broda, 2007; Piskorski i in., 2005; Abramowicz i in., 2006). Niestety, ich wykorzystanie dla zadań uczenia ontologii z tekstu jest pracochłonne, ponieważ wymaga dostosowania ich specyficznych formatów do formatów anotacji lingwistycznych wykorzystywanych w dziedzinie uczenia ontologii z tekstu, np. opisywanego formatu dla metamodelu (rozdział 4). Istnieje zatem potrzeba opracowania narzędzi umożliwiających ich łatwiejsze i powtarzalne wykorzystanie. W konsekwencji, ewaluacja metod dla języka polskiego byłaby ułatwiona, a przede wszystkim metodologicznie poprawna.

Uzyskane wyniki prac zostaną wykorzystane do ewolucji ontologii. Większość autorów metod ewolucji ontologii ogranicza się do opracowania modeli ewolucji lub metod, które na wstępie wymagają zbudowanej ontologii (Bloehdorn i in., 2006; Leenheer i Mens, 2008; Flouris, 2006). Niestety, jest to bardzo powierzchowne traktowanie problemu, gdyż większość wyzwań leży po stronie odpowiedniego przygotowania takich ontologii. Dopiero na podstawie odpowiednio przygotowanego modelu uczenia (np. przedstawionego w niniejszej pracy) można budować modele ewolucji ontologii.

Bibliografia

Abramowicz, W. (2008). *Filtrowanie informacji*. Poznan: Wydawnictwo Akademii Ekonomicznej w Poznaniu.

Abramowicz, W., Piskorski, J., Filipowska, A., Wecel, K., Wieloch, K. (2006). Linguistic suite for polish cadastral system. W: *Proceedings of LREC 2006*.

Abramowicz, W., Vargas-Vera, M., Wisniewski, M. (2008). Axiom-based feedback cycle for relation extraction in ontology learning from text. W: *DEXA '08: Proceedings of the 19th International Conference on Database and Expert Systems Applications* 202–206, Los Alamitos, CA, USA. IEEE Computer Society.

Abramowicz, W., Wisniewski, M. (2008a). Metamodel of ontology learning from text. W: Chbeir, R., Hassanien, A.-E., Abraham, A., Badr, Y. (red.), *Emergent Web Intelligence*, Studies in Computational Intelligence. Springer Verlag.

Abramowicz, W., Wisniewski, M. (2008b). Proximity window context method for term extraction in ontology learning from text. W: *DEXA '08: Proceedings of the 19th International Conference on Database and Expert Systems Applications* 215–219, Los Alamitos, CA, USA. IEEE Computer Society.

Abramowicz, W., Wiśniewski, M. (2008). Meta-model dla uczenia ontologii z tekstu. W: *Systemy wspomagania organizacji*, Ustron, Poland. Naukowe Towarzystwo Informatyki Ekonomicznej, Katedra Informatyki Akademii Ekonomicznej w Katowicach.

Adar, E. (2004). Sarad: a simple and robust abbreviation dictionary. *Bioinformatics*, *20*(4), 527–533.

Agichtein, E., Gravano, L. (2000). Snowball: extracting relations from large plain-text collections. W: *DL '00: Proceedings of the fifth ACM conference on Digital libraries* 85–94, New York, NY, USA. ACM.

Agichtein, E., Gravano, L., Pavel, J., Sokolova, V., Voskoboynik, A. (2001). Snowball: a prototype system for extracting relations from large text collections. W: *SIGMOD '01: Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data* 612, New York, NY, USA. ACM.

Agirre, E., Ansa, O., Hovy, E., Martínez, D. (2000). Enriching very large ontologies using the www. W: *Proc. of the Ontology Learning Workshop, ECAI*, Berlin, Germany.

Aho, A. V., Corasick, M. J. (1975). Efficient string matching: an aid to bibliographic search. *Commun. ACM*, *18*(6), 333–340.

A'it-Mokhtar, S., Chanod, J.-P. (1997). Incremental finite-state parsing. W: *ANLP'97*.

Akerlof, G. A., Spence, A. M., Stiglitz, J. E. (2001). Markets with asymmetric information. Nobel prize in economics, Nobel Prize Committee.

Alfonseca, E., Castells, P., Okumura, M., Ruiz-Casado, M. (2006a). A rote extractor with edit distance-based generalisation and multi-corpora precision calculation. W: *Proceedings of the COLING/ACL on Main conference poster sessions* 9–16, Morristown, NJ, USA. Association for Computational Linguistics.

Alfonseca, E., Manandhar, S. (2002a). Extending a lexical ontology by a combination of distributional semantics signatures. W: *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2002)*.

Alfonseca, E., Manandhar, S. (2002b). Improving an ontology refinement method with hyponymy patterns. W: *Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain.

Alfonseca, E., Manandhar, S. (2002c). An unsupervised method for ontology refinement. W: *Poceedings of the First International Conference on General WordNet*, Mysore, India.

Alfonseca, E., Ruiz-Casado, M., Okumura, M., Castells, P. (2006b). Towards large-scale non-taxonomic relation extraction: Estimating the precision of rote extractors. W: *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* 49–56, Sydney, Australia. Association for Computational Linguistics.

Ananiadou, S., Mcnaught, J. (2006). *Text Mining for Biology And Biomedicine*. Norwood, MA, USA: Artech House, Inc.

Appelt, D. E., Israel, D. (1999). Introduction to information extraction technology. W: *IJCAI-99*, Stockholm, Sweden.

Aussenac-Gilles, N., Biebow, B., Szulman, S. (2000a). Corpus analysis for conceptual modelling. W: *Workshop on Ontologies and Text, Knowledge Engineering and Knowledge Management: Methods, Models and Tools, 12th International Conference EKAW 2000*, Juan-les-pins, France. Springer-Verlag.

Aussenac-Gilles, N., Biébow, B., Szulman, S. (2000b). Revisiting ontology design: A methodology based on corpus analysis. W: *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*. Springer-Verlag.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., Patel-Schneider, P. F. (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

Baeza-Yates, R. A., Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Basili, R., Pazienza, M. T., Velardi, P. (1996). An empirical symbolic approach to natural language processing. *Artif. Intell.*, *85*, 59–99.

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

Bloehdorn, S., Haase, P., Sure, Y., Voelker, J. (2006). Ontology evolution. W: Davies, J., Studer, R., Warren, P. (red.), *Semantic Web Technologies*. John Wiley.

Brin, S. (1999). Extracting patterns and relations from the world wide web. W: *WebDB '98: Selected papers from the International Workshop on The World Wide Web and Databases* 172–183, London, UK. Springer-Verlag.

Brody, S., Navigli, R., Lapata, M. (2006). Ensemble methods for unsupervised wsd. W: *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* 97–104, Morristown, NJ, USA. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C., Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–479.

Budanitsky, A., Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, *32*(1), 13–47.

Buitelaar, P. (2003). Ontology learning for semantic web services. Raport techniczny, DFKI GmbH, Saarbrücken.

Buitelaar, P., Cimiano, P. (2006). Ontology learning from text: Tutorial. W: *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Buitelaar, P., Cimiano, P., Magnini, B. (2005a). *Ontology Learning from Text : Methods, Evaluation and Applications*. Frontiers in artificial intelligence and applications. Amsterdam, Washington: IOS Press.

Buitelaar, P., Cimiano, P., Magnini, B. (2005b). Ontology learning from text: An overview. W: Buitelaar, P., Cimiano, P., Magnini, B. (red.), *Ontology Learning from Text : Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence and Applications. Amsterdam ; Washington, DC: IOS Press.

Buitelaar, P., Olejnik, D., Sintek, M. (2004a). A protege plug-in for ontology extraction from text based on linguistic analysis. W: *Proceedings of the 1st European Semantic Web Symposium (ESWS)*.

Buitelaar, P., Sintek, M. (2004). Ontolt version 1.0: Middleware for ontology extraction from text. W: *Proceedings. of the Demo Session at the International Semantic Web Conference (ISWC)*.

Buitelaar, P., Sintek, M., Iqbal, Y. (2004b). *OntoLT Version 1.0: Short User Guide*.

Bunescu, R., Mooney, R. (2007). Learning to extract relations from the web using minimal supervision. W: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* 576–583, Prague, Czech Republic. Association for Computational Linguistics.

Bunescu, R. C., Mooney, R. J. (2005). Subsequence kernels for relation extraction. W: *Proceedings of the Nineteenth Annual Conference on Neural Information Processing Systems*.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. W: *Proceedings of the Conference of the Association for Computational Linguistics*.

Caraballo, S. A. (2001). *Automatic construction of a hypernym-labeled noun hierarchy from text*. Niepublikowana rozprawa doktorska. Providence, RI, USA. Adviser-Eugene Charniak.

Cederberg, S., Widdows, D. (2003). Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. W: *Proceedings of the Conference on Natural Language Learning (CoNNL)*.

Chang, J., Schutze, H. (2006). Abbreviations in biomedical text. W: Ananiadou, S., Mcnaught, J. (red.), *Text mining for biology and biomedicine*. Artech House.

Charniak, E., Berland, M. (1999). Finding parts in very large corpora. W: *Proceedings of the 37th Annual Meeting of the ACL*.

Cimiano, P. (2006). *Ontology Learning from Text*. Niepublikowana rozprawa doktorska. University of Karlsruhe.

Cimiano, P., Hotho, A., Staab, S. (2005a). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, *24*, 305–339.

Cimiano, P., Schmidt-Thieme, L., Pivk, A., Staab, S. (2005b). Learning taxonomic relations from heterogeneous evidence. W: *Ontology Learning from Text: Methods, Applications and Evaluation*. IOS Press.

Cimiano, P., Staab, S. (2005). Learning concept hierarchies from text with a guided agglomerative clustering algorithm. W: *ICML 2005 Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*.

Cimiano, P., Völker, J. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. W: *10th International Conference on Applications of Natural Language to Information Systems (NLDB'2005)*.

Cimiano, P., Wenderoth, J. (2005). Automatically learning qualia structures from the web. W: *Proceedings of the ACL Workshop on Deep Lexical Acquisition*.

Cristani, M., Cuel, R. (2005). A survey on ontology creation methodologies. *Int. J. Semantic Web Inf. Syst.*, *1*(2), 49–69.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. (2002). Gate: A framework and graphical development environment for robust nlp tools and applications. W: *Proceedings of the 40th Annual Meeting of the ACL*.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., Dimitrov, M., Dowman, M., Aswani, N., Roberts, I., Li, Y., Shafirin, A. (2007). *Developing Language Processing Components with GATE Version 4*. Department of Computer Science, University of Sheffield, wyd. 4.0-beta1.

Cunningham, H., Maynard, D., Tablan, V. (2000). Jape: A java annotation patterns engine (second edition). Raport techniczny, Department of Computer Science, University of Sheffield.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. W: Klavans, J., Resnik, P. (red.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, Massachusetts: The MIT Press.

Declerck, T. (2002). A set of tools for integrating linguistic and non-linguistic.

Dellschaft, K., Staab, S. (2006). On how to perform a gold standard based evaluation of ontology learning. W: *Proc. of ISWC-2006 International Semantic Web Conference*, Athens, GA, USA. Springer, LNCS.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R. (2004). Automatic content extraction (ace) program - task definitions and performance measures. W: *LREC 2004: Fourth International Conference on Language Resources and Evaluation*.

Faatz, A., Steinmetz, R. (2002). Ontology enrichment with texts from the www. W: *Semantic Web Mining 2nd Workshop at ECML/PKDD-2002*, Helsinki, Finland.

Fasli, M. (2007). On agent technology for e-commerce: trust, security and legal issues. *Knowl. Eng. Rev.*, *22*(1), 3–35.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fensel, D. (2003). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Fensel, D., van Harmelen, F., Klein, M., Akkermans, H., Broekstra, J., Fluit, C., van der Meer, J., Schnurr, H.-P., Studer, R., Hughes, J., Krohn, U., Davies, J., Engels, R., Bremdal, B., Ygge, F., Lau, T., Novotny, B., Reimer, U., Horrocks, I. (2000). Onto -knowledge: Ontology-based tools for knowledge management. W: *Proceedings of the eBusiness and eWork 2000 (eBeW'00) Conference*, Madrid, Spain.

Finkelstein-Landau, M., Morin, E. (1999). Extracting semantic relationships between terms: Supervised vs. unsupervised methods. W: *Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure*, Dagstuhl Castle, Germany.

Firth, J. (1957). *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis, Philological Society. Longman.

Fisiak, J. (2002). *Collins słownik polsko-angielski*.

Flouris, G. (2006). *On Belief Change and Ontology Evolution*. Niepublikowana rozprawa doktorska. University of Crete.

Fotzo, H., Gallinari, P. (2004). Learning generalization/specialization relations between concepts - application for automatically building thematic document hierarchies. W: *RIAO*.

Francis, W. N., Kucera, H. (1979). *Brown Corpus Manual*. Brown University, Providence, Rhode Island.

Frantzi, K., Ananiadou, S. (2007). C-value for authorship identification. W: *8th Conference on Forensic Linguistics, Language and Law*, University of Washington, Seattle, Washington, USA. International Association of Forensic Linguistics.

Frantzi, K., Ananiadou, S., Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, *V3*(2), 115–130.

Ganter, B., Wille, R. (1999). *Formal Concept Analysis – Mathematical Foundations*. Springer Verlag.

Garfinkel, R., Gopal, R., Tripathi, A., Yin, F. (2006). Design of a shopbot and recommender system for bundle purchases. *Decis. Support Syst.*, *42*(3), 1974–1986.

Garside, R., Smith, N. (1997). A hybrid grammatical tagger: Claws4. W: Garside, R., Leech, G., McEnery, A. (red.), *Corpus annotation: Linguistic information from computer text corpora*. Longman.

Girju, R., Moldovan, D. (2002). Text mining for causal relations. W: *Proceedings of the FLAIRS Conference*.

Greenbaum, S., Svartvik, J. (1990). The london corpus of spoken english: Description and research. W: Svartvik, J. (red.), *Lund Studies in English 82*. Lund University Press.

Grefenstette, G. (1998). *Cross-language information retrieval*. Kluwer international series on information retrieval ; 2. Boston, MA: Kluwer Academic Publishers.

Gruber, T. (2008). Ontology. W: Liu, L., Ozsu, M. T. (red.), *Encyclopedia of Database Systems*. Springer-Verlag.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowl. Acquis.*, *5*(2), 199–220.

Guo, J. (2006). Inter-enterprise business document exchange. W: *ICEC '06: Proceedings of the 8th International Conference on Electronic Commerce* 427–437, New York, NY, USA. ACM.

GuoDong, Z., Jian, S., Jie, Z., Min, Z. (2005). Exploring various knowledge in relation extraction. W: *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* 427–434, Morristown, NJ, USA. Association for Computational Linguistics.

Gómez-Pérez, A., Manzano-Macho, D. (2004). An overview of methods and tools for ontology learning from texts. *The Knowledge Engineering Review*, *19*(3), 187–212.

Haase, P., Sure, Y. (2004). State-of-the-art on ontology evolution. Raport techniczny, Institute AIFB, University of Karlsruhe.

Haase, P., Völker, J. (2005). Ontology learning and reasoning - dealing with uncertainty and inconsistency. W: *Proceedings of the Workshop on Uncertainty Reasoning for the Semantic Web (URSW)*.

Hahn, U., Markó, K. (2001). Joint knowledge capture for grammars and ontologies. W: *Proceedings of the First International Conference on Knowledge Capture K-CAP 2001*, Victoria, Canada.

Hahn, U., Schnattinger, K. (1998). Towards text knowledge engineering. W: *AAAI '98 / IAAI '98 Proceedings of the 15th National Conference on Artificial Intelligence and 10th Conference on Innovative Applications of Artificial Intelligence*, Madison, Wisconsin. AAAI Press / MIT Press.

Hahn, U., Schulz, S. (2000). Towards very large terminological knowledge bases: A case study from medicine. W: *Canadian Conference on AI 2000*.

Hammerton, J., Osborne, M., Armstrong, S., Daelemans, W. (2002). Introduction to special issue on machine learning approaches to shallow parsing. *Journal of Machine Learning Research*, *2002*(2), 8.

Hamp, B., Feldweg, H. (1997). Germanet - a lexical-semantic net for german. W: *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain.

Harris, Z. (1986). *Mathematical Structures of Language*. Wiley.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. W: *14th International Conference on Computational Linguistics*.

Hearst, M. A. (1998). Automated discovery of wordnet relations. W: Fellbaum, C. (red.), *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press.

Hepp, M. (2006). Products and services ontologies: A methodology for deriving owl ontologies from industrial categorization standards. *Int. J. Semantic Web Inf. Syst.*, *2*(1), 72–99.

Hepp, M. (2008). Ontologies: State of the art, business potential, and grand challenges. W: *Ontology Management*.

Hepp, M., Leenheer, P. D., de Moor, A., Sure, Y. (2008). *Ontology Management, Semantic Web, Semantic Web Services, and Business Applications*, tom 7 z *Semantic Web And Beyond Computing for Human Experience*. Springer.

Hepple, M. (2000). Independence and commitment: Assumptions for rapid training and execution of rule-based pos taggers. W: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, Hong Kong.

Holzinger, W., Krupl, B., Herzog, M. (2006). Using ontologies for extracting product features from web pages. W: *International Semantic Web Conference* 286–299.

Hovy, E., Lin, C.-Y. (1999). Automated text summarization in summarist. W: Maybury, M., Mani, I. (red.), *Advances in Automatic Text Summarization*. Cambridge: MIT Press.

Huang, J.-X., Shin, J.-A., Choi, K.-S. (2007). Integrating relations for a domain ontology. W: *Proceedings of the 6th International Semantic Web Conference*, Busan, Korea.

Hwang, C. H. (1999). Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. W: *Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99)*, Linköping, Sweden.

InternetStandard, Sklepy24.pl (2008). *Raport e-commerce 2007*. http://www.sklepy24.pl/.

ISO 1087-1:2000 (2000). *ISO 1087-1:2000 Terminology work – Vocabulary – Part 1: Theory and application*. International Organization for Standardization.

ISO 704:2000 (2000). *ISO 704:2000 Terminology work – Principles and methods*. International Organization for Standardization.

ISO 860:2007 (2007). *ISO 860:2007 Terminology work – Harmonization of concepts and terms*. International Organization for Standardization.

Iwanska, L. (2000). Natural language is a powerful knowledge representation system: the uno model. W: Iwanska, L., Shapiro, S. C. (red.), *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. AAAI Press/MIT Press.

Iwanska, L., Mata, N., Kruger, K. (2000). Fully automatic acquisition of taxonomic knowledge from large corpora of texts: Limited-syntax knowledge representation system based on natural language. W: *International Syposium on Methodologies for Intelligent Systems*.

Jurafsky, D., Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Kaczmarek, T. (2007). *Integracja danych z głębokiego Internetu dla potrzeb analizy otoczenia przedsiębiorstwa*. Niepublikowana rozprawa doktorska. Poznan University of Economics.

Kanzaki, K., Yamamoto, E., Isahara, H., Ma, Q. (2004). Construction of an objective hierarchy of abstract concepts via directional similarity. W: *COLING '04: Proceedings of the 20th International Conference on Computational Linguistics* 1147, Morristown, NJ, USA. Association for Computational Linguistics.

Kasneci, G., Suchanek, F. M., Ramanath, M., Weikum, G. (2007). How naga uncoils: searching with entities and relations. W: *WWW '07: Proceedings of the 16th International Conference on World Wide Web* 1167–1168, New York, NY, USA. ACM.

Kauffman, R. J., Walden, E. A. (2001). Economics and electronic commerce: Survey and directions for research. *Int. J. Electron. Commerce*, *5*(4), 5–116.

Kenter, T., Maynard, D. (2005). *Using GATE as an Annotation Tool.* Department of Computer Science, University of Sheffield.

Khandelwal, S. (2007). Extracting semantic content from webpages. W: *Proceedings of RIAO 2007*, Pittsburgh, Pennsylvania, USA.

Kietz, J., Maedche, A., Volz, R. (2000). A method for semi-automatic ontology acquisition from a corporate intranet. W: *Workshop "Ontologies and text", co-located with EKAW'2000.*

Kipfer, B. A. (2006). *Roget New Millennium Thesaurus, First Edition (v 1.1.1).*

Knublauch, H., Fergerson, R. W., Noy, N. F., Musen, M. A. (2004). The protege owl plugin: An open development environment for semantic web applications. W: *Third International Semantic Web Conference - ISWC 2004*, Hiroshima, Japan.

Kudo, T., Matsumoto, Y. (2003). Fast methods for kernel-based text analysis. W: *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* 24–31, Morristown, NJ, USA. Association for Computational Linguistics.

Kuropka, D. (2005). Uselessness of simple co-occurrence measures for if and ir - a linguistic point of view. W: *Proceedings of the 8th International Conference on Business Information Systems*, Poznan, Poland.

Leenheer, P., Mens, T. (2008). Ontology evolution. W: *Ontology Management.* Springer.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of ACM*, *38*(11), 32–38.

Lin, C.-Y. (1997). *Robust Automated Topic Identification.* Niepublikowana rozprawa doktorska. University of Southern California.

Lin, C.-Y., Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. W: *Proc. of the COLING Conference*, Strasbourg, France.

Lin, D., Pantel, P. (2001). Dirt - discovery of inference rules from text. W: *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

Maedche, A. (2002). *Ontology Learning for the Semantic Web.* Boston: Kluwer Academic Publishers.

Maedche, A., Staab, S. (2000a). Discovering conceptual relations from text. W: *ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence*, Berlin, Germany. IOS Press.

Maedche, A., Staab, S. (2000b). Semi-automatic engineering of ontologies from text. W: *Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering.*

Maedche, A., Staab, S. (2000c). The text-to-onto ontology learning environment. W: *Proceedings of the 12th Internal Conference on Software and Knowledge Engineering*, Chicago, USA.

Maedche, A., Staab, S. (2004). Ontology learning. W: Staab, S., Studer, R. (red.), *Handbook of Ontologies in Information Systems.* Springer Verlag.

Mann, G. S., Yarowsky, D. (2005). Multi-field information extraction and cross-document fusion. W: *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* 483–490, Morristown, NJ, USA. Association for Computational Linguistics.

Manning, C. D., Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: The MIT Press.

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, *19*(2), 313–330.

Markert, K., Nissim, M., Modjeska, N. (2003). Using the web for nominal anaphora resolution. W: *EACL Workshop on the Computational Treatment of Anaphora.*

McBride, B. (2002). Jena: a semantic web toolkit. *Internet Computing, IEEE*, *6*(6), 55–59.

Medjahed, B., Benatallah, B., Bouguettaya, A., Ngu, A. H. H., Elmagarmid, A. K. (2003). Business-to-business interactions: issues and enabling technologies. *The VLDB Journal*, *12*(1), 59–85.

Mikheev, A., Moens, M., Grover, C. (1999). Named entity recognition without gazetteers. W: *Proceedings of EACL*, Bergen, Norway.

Missikoff, M., Navigli, R., Velardi, P. (2002). *Integrated Approach to Web Ontology Learning and Engineering.*

Morin, E., Jacquemin, C. (1999). Projecting corpus-based semantic links on a thesaurus. W: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99).*

Nadeau, D. (2005). Balie - baseline information extraction. multilingual information extraction from text with machine learning and natural language techniques. Raport techniczny, School of Information Technology And Engineering, University of Ottawa, Canada.

Nakagawa, H., Mori, T. (1998). Nested collocation and compound noun for term extraction. W: *Proceedings of the First Workshop on Comutational Terminology (COMPUTERM'98)* 64–70.

Nakagawa, H., Mori, T. (2002). A simple but powerful automatic term extraction method. W: *COLING-02 on COMPUTERM 2002* 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Nakagawa, H., Mori, T. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, *9*(2), 201–219.

Narayanan, S., Petruck, M. R. L., Baker, C. F., Fillmore, C. J. (2003). Putting framenet data into the iso linguistic annotation framework. W: *Proceedings of the ACL 2003 workshop on Linguistic annotation* 22–29, Morristown, NJ, USA. Association for Computational Linguistics.

Navigli, R. (2006a). Consistent validation of manual and automatic sense annotations with the aid of semantic graphs. *Comput. Linguist.*, *32*(2), 273–281.

Navigli, R. (2006b). Meaningful clustering of senses helps boost word sense disambiguation performance. W: *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* 105–112, Morristown, NJ, USA. Association for Computational Linguistics.

Navigli, R. (2006c). Valido: a visual tool for validating sense annotations. W: *Proceedings of the COLING/ACL on Interactive presentation sessions* 13–16, Morristown, NJ, USA. Association for Computational Linguistics.

Navigli, R., Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Comput. Linguist.*, *30*(2), 151–179.

Navigli, R., Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, *27*(7), 1075–1086.

Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F. (2004). Quantitative and qualitative evaluation of the ontolearn ontology learning system. W: *COLING '04: Proceedings of the 20th International Conference on Computational Linguistics* 1043, Morristown, NJ, USA. Association for Computational Linguistics.

Nenadic, G., Ananiadou, S. (2006). Mining semantically related terms from biomedical literature. *ACM Transactions on Asian Language Information Processing (TALIP)*, *5*(1), 22–43.

Nenadic, G., Ananiadou, S., McNaught, J. (2004). Enhancing automatic term recognition through recognition of variation. W: *COLING '04: Proceedings of the 20th International Conference on Computational Linguistics* 604, Morristown, NJ, USA. Association for Computational Linguistics.

Neumann, G., Backofen, R., Baur, J., Becker, M., Braun, C. (1997). An information extraction core system for real world german text processing. W: *ANLP 97 – Proceedings of the Conference on Applied Natural Language Processing*, Washington, USA.

Niles, I., Pease, A. (2001). Towards a standard upper ontology. W: *FOIS '01: Proceedings of the International Conference on Formal Ontology in Information Systems* 2–9, New York, NY, USA. ACM.

Ogden, C. K., Richards, I. A. (1923). *The meaning of meaning : a study of the influence of language upon thought and of the science of symbolism.* International library of psychology, philosophy, and scientific method. New York: Harcourt, Brace.

Okazaki, N., Ananiadou, S. (2006). A term recognition approach to acronym recognition. W: *Proceedings of the COLING/ACL on Main conference poster sessions* 643–650, Morristown, NJ, USA. Association for Computational Linguistics.

Ozsu, M. T., Snodgrass, R. T. (2001). *ACM SIGMOD Anthology.*

Palopoli, L., Rosaci, D., Ursino, D. (2006). Agents' roles in b2c e-commerce. *AI Commun.*, *19*(2), 95–126.

Pereira, F., Tishby, N., Lee, L. (1993). Distributional clustering of english words. W: *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, Columbus, Ohio. Association for Computational Linguistics.

Piasecki, M., Broda, B. (2007). Semantic similarity measure of polish nouns based on linguistic features. W: *Proceedings of 10th International Conference on Business Information Systems*, Lecture Notes in Computer Science 381–390, Poznan, Poland. Springer.

Pinto, H. S., Martins, J. P. (2004). Ontologies: How can they be built? *Knowledge and Information Systems*, *6*(4), 441–464.

Piskorski, J. (2002). *Shallow Text Processor based on Finite-State Technology for Information Extraction.* Niepublikowana rozprawa doktorska. Instytut Podstaw Informatyki Polskiej Akademii Nauk.

Piskorski, J., Drozdzynski, W., Krieger, H.-U., Schafer, U. (2005). Sprout - a general-purpose nlp framework integrating finite-state and unification-based grammar formalisms. W: *Proceedings of the 5th International Workshop on Finite-State Methods and Natural Language Processing*, Helsinki, Finland. Springer - Lecture Notes in Artificial Intelligence.

Piskorski, J., Neumann, G. (2000). An intelligent text extraction and navigation system. W: *6th International Conference on Computer-Assisted Information Retrieval (RIAO-2000)*, Paris, France.

PN-ISO 1087-1:2004 (2004). *PN-ISO 1087-1:2004 Działalność terminologiczna. Terminologia. Część 1: Teoria i zastosowanie.* Polski Komitet Normalizacyjny.

Poesio, M., Almuhareb, A. (2005). Identifying concept attributes using a classifier. W: *Proceedings of the ACL Workshop on Deep Lexical Acquisition.*

Poesio, M., Ishikawa, T., im Walde, S. S., Vieira, R. (2002). Acquiring lexical knowledge for anaphora resolution. W: *Proceedings of the 3rd Conference on Language Resources and Evaluation.*

Pytkowski, W. (1985). *Organizacja badań i ocena prac naukowych*, tom 0239-6114 z *Skrypty Uczelniane Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie*. Warszawa.

Rajman, M., Bonnet, A. (1992). New tools for text analysis: Corpora-based linguistics. W: *The First Conference of the Association for Global Strategic Information*, Bad Kreuznach, Germany.

Ravichandran, D., Hovy, E. (2001). Learning surface text patterns for a question answering system. W: *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* 41–47, Morristown, NJ, USA. Association for Computational Linguistics.

Riloff, E., Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. W: *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Somerset, New Jersey.

Rinaldi, F., Yuste, E. (2005). Exploiting technical terminology for knowledge management. W: Buitelaar, P., Cimiano, P., Magnini, B. (red.), *Ontology Learning from Text : Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence and Applications. Amsterdam ; Washington, DC: IOS Press.

Roark, B., Charniak, E. (1998). Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. W: *COLING-ACL*.

Robinson, J. A., Voronkov, A. (2001). *Handbook of Automated Reasoning (in 2 volumes)*. Elsevier and MIT Press.

Roux, C., Proux, D., Rechenmann, F., Julliard, L. (2000). An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. W: *Proceedings of the ECAI2000 Workshop on Ontology Learning(OL2000)*, Berlin, Germany.

Ruiz-Casado, M., Alfonseca, E., Castells, P. (2007). Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data Knowl. Eng.*, *61*(3), 484–499.

Ryu, P.-M., Choi, K.-S. (2005). An information theoretic approach to taxonomy extraction for ontology learning. W: Buitelaar, P., Cimiano, P., Magnini, B. (red.), *Ontology Learning from Text: Methods, Evaluation and Applications*, tom 123 z *Frontiers in Artificial Intelligence and Applications*. Amsterdam, The Netherlands: IOS Press.

Ryu, P.-M., Choi, K.-S. (2006). Taxonomy learning using term specificity and similarity. W: *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* 41–48, Sydney, Australia. Association for Computational Linguistics.

Sabou, M., Wroe, C., Goble, C., Stuckenschmidt, H. (2005). Learning domain ontologies for semantic web service descriptions. *Journal of Web Semantics*, *3*(4).

Salton, G. (1991). Developments in automatic text retrieval. *Science*, *253*, 974–997.

Sampson, G. (1995). *English for the Computer: The SUSANNE Corpus and analytic scheme.*

Sanderson, M., Croft, B. (1999). Deriving concept hierarchies from text. W: *SIGIR '99*. ACM.

Sawaki, M., Hagita, N., Ishii, K. (1997). Robust character recognition of gray-scaled images with graphical designs and noise. W: *ICDAR '97: Proceedings of the 4th International Conference on Document Analysis and Recognition* 491–494, Washington, DC, USA. IEEE Computer Society.

Schwartz, A., Hearst, M. (2003). A simple algorithm for identifying abbreviation definitions in biomedical texts. W: *Proceedings of the Pacific Symposium on Biocomputing PSB 2003*.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Simperl, E. P. B., Mochol, M. (2006). Ontology engineering cost estimation with ontocom. Tr-b-06-01, Freie Universitat Berlin, Berlin, Germany.

Simperl, E. P. B., Sure, Y., Tempich, C. (2006). Ontocom: A cost estimation model for ontology engineering. W: *Proceedings of the 5th International Semantic Web Conference*, Athens, Georgia.

Simperl, E. P. B., Tempich, C., Mochol, M. (2007). Cost estimation for ontology development: applying the ontocom model. W: Abramowicz, W., Mayr, H. C. (red.), *Technologies for Business Information Systems*. Springer Verlag.

Singh, R., Iyer, L. S., Salam, A. F. (2005). Semantic ebusiness. *Int. J. Semantic Web Inf. Syst.*, *1*(1), 19–35.

Sintek, M., Buitelaar, P., Olejnik, D. (2004). A formalization of ontology learning from text. W: *International Semantic Web Conference*, Hiroshima, Japan.

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., Katz, Y. (2007). Pellet: A practical owl-dl reasoner. *Journal of Web Semantics*, *5*(2).

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computional Linguistics*, *19*(1), 143–177.

Snow, R., Jurafsky, D., Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. W: *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* 801–808, Morristown, NJ, USA. Association for Computational Linguistics.

Sowa, J. F. (1984). *Conceptual structures: information processing in mind and machine.* Reading, Massachusets: Addison-Wesley.

Sowa, J. F. (2000a). *Knowledge representation : logical, philosophical, and computational foundations.* Pacific Grove: Brooks/Cole.

Sowa, J. F. (2000b). Ontology, metadata, and semiotics. W: *Proceedings of the Linguistic on Conceptual Structures: Logical Linguistic, and Computational Issues*. Springer-Verlag.

Spasic, I., Nenadic, G., Ananiadou, S. (2002). Tuning context features with genetic algorithms. W: *Proceedings of 3rd International Conference on Language, Resources and Evaluation* 2048–2054, Las Palmas, Spain.

Spasic, I., Nenadic, G., Ananiadou, S. (2003). Using domain-specific verbs for term classification. W: *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine* 17–24, Morristown, NJ, USA. Association for Computational Linguistics.

Srikant, R., Agrawal, R. (1995). Mining generalized association rules. W: *Proceedings of VLDB 95*.

Suchanek, F. M., Ifrim, G., Weikum, G. (2006a). Combining linguistic and statistical analysis to extract relations from web documents. W: *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 712–717, New York, NY, USA. ACM.

Suchanek, F. M., Ifrim, G., Weikum, G. (2006b). Leila: Learning to extract information by linguistic analysis. W: *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* 18–25, Sydney, Australia. Association for Computational Linguistics.

Suchanek, F. M., Kasneci, G., Weikum, G. (2007). Yago: a core of semantic knowledge. W: *WWW '07: Proceedings of the 16th International Conference on World Wide Web* 697–706, New York, NY, USA. ACM.

Sundblad, H. (2003). Automatic acquisition of hyponyms and meronyms from question corpora. W: *Proceedings of the Workshop on Natural Language Processing and Machine Learning for Ontology Engineering at ECAI'2002*, Lyon, France.

Tablan, V., Maynard, D., Bontcheva, K., Cunningham, H. (2004). *GATE – An Application Developers Guide*. Department of Computer Science, University of Sheffield.

Torii, M., Liu, H., Hu, Z., Wu, C. (2006). A comparison study of biomedical short form definition detection algorithms. W: *TMBIO '06: Proceedings of the 1st international workshop on Text mining in bioinformatics* 52–59, New York, NY, USA. ACM.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. W: *Advances in Informatics - 10th Panhellenic Conference on Informatics*, LNCS 3746 382–392.

Uschold, M., Gruninger, M. (2004). Ontologies and semantics for seamless connectivity. *SIGMOD Rec.*, *33*(4), 58–64.

Velardi, P., Fabriani, P., Missikoff, M. (2001a). Using text processing techniques to automatically enrich a domain ontology. W: *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS)*.

Velardi, P., Missikoff, M., Basili, R. (2001b). Identification of relevant terms to support the construction of domain ontologies. W: *ACL-EACL Workshop on Human Language Technologies*, Toulouse, France.

Vieira, R., Poesio, M. (2000). An empirically based system for processing definite descriptions. *Comput. Linguist.*, *26*(4), 539–593.

Vintar, S., Buitelaar, P., Sacaleanu, B., Raileanu, D., Prescher, D., Ripplinger, B., Brown, R., Bay, J., Weiser, O., Gaussier, E., Dejean, H., Widdows, D. (2001). Muchmore annotation format. Raport techniczny, DFKI.

Voelkel, M. (2005). Versioning rdf and ontologies. Raport techniczny, University of Karlsruhe.

Vossen, P. (1998). *Introduction to EuroWordNet*, tom 32.

Vossen, P. (2001). Extending, trimming and fusing wordnet for technical documents. W: *NAACL-2001 workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*.

Wermter, J., Hahn, U. (2005a). Finding new terminology in very large corpora. W: *K-CAP '05: Proceedings of the 3rd International Conference on Knowledge Capture* 137–144, New York, NY, USA. ACM.

Wermter, J., Hahn, U. (2005b). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. W: *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* 843–850, Morristown, NJ, USA. Association for Computational Linguistics.

Wermter, J., Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction. W: *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* 785–792, Morristown, NJ, USA. Association for Computational Linguistics.

Widdows, D. (2003). Unsupervised method for developing taxonomies by combining syntactic and statistical information. W: *Proceedings of HLT/NAACL*.

Wisniewski, M. (2006). Uczenie ontologii z tekstu. W: *Współczesne trendy w informatyce ekonomicznej*, tom 16 z *Roczniki Kolegium Analiz Ekonomicznych*. Warszawa: Szkoła Główna Handlowa w Warszawie.

Witschel, H. (2005). Using decision trees and text mining techniques for extending taxonomies. W: *Proceedings of Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at ICML-05*.

Woliński, M. (2006). *Analizator morfologiczny Morfeusz*.

Wroe, C., Goble, C., Greenwood, M., Lord, P., Miles, S., Papay, J., Payne, T., Moreau, L. (2004). Automating experiments using semantic data on a bioinformatics grid. *IEEE Intelligent Systems*, *19*(1), 48–55.

Xu, F., Kurz, D., Piskorski, J., Schmeier, S. (2002). A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping. W: *Proceedings of the 3rd International Conference on Language Resources an Evaluation (LREC'02)*, Las Palmas, Canary Islands, Spain.

Yamada, I., Baldwin, T. (2004). Automatic discovery of telic and agentive roles from corpus data. W: *Proceedings of the The 18th Pacific Asia Conference on Language, Information and Computation (PACLIC 18)*.

Yamamoto, E., Kanzaki, K., Isahara, H. (2005). Extraction of hierarchies based on inclusion of co-occurring words with frequency information. W: *IJCAI* 1166–1174.

Yang, X., Su, J., Tan, C. L. (2006). Kernel-based pronoun resolution with structured syntactic knowledge. W: *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* 41–48, Morristown, NJ, USA. Association for Computational Linguistics.

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. W: *Proceedings of COLING-92*, Nantes, France.

Zavitsanos, E., Paliouras, G., Vouros, G. (2006). Ontology learning and evaluation: A survey. Raport techniczny, NCSR Demokritos.

Zhang, M., Zhang, J., Su, J. (2006a). Exploring syntactic features for relation extraction using a convolution tree kernel. W: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* 288–295, Morristown, NJ, USA. Association for Computational Linguistics.

Zhang, M., Zhang, J., Su, J., Zhou, G. (2006b). A composite kernel to extract relations between entities with both flat and structured features. W: *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* 825–832, Morristown, NJ, USA. Association for Computational Linguistics.

Zhao, S., Grishman, R. (2005). Extracting relations with integrated information using kernel methods. W: *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* 419–426, Morristown, NJ, USA. Association for Computational Linguistics.

Zhou, G., Zhang, M. (2007). Extracting relation information from text documents by exploring various types of knowledge. *Inf. Process. Manage.*, *43*(4), 969–982.