Rozwiązywanie dysambiguacji za pomocą lingwistycznej sieci symantycznych relacji leksykalnych na przykładzie systemu WordNet

Julian Szymański, Kamila Błaszczyk, Artur Chełmecki

18 marca 2007

1 Wstęp

Rozwój informatyki spowodował wkraczanie komputerów i elektronicznego wspomagania działań człowieka w coraz to nowe dziedziny życia. W ramach usprawniania komunikacji i interfejsów między człowiekiem a komputerem coraz bardziej palącą kwestią jest umożliwienie zrozumienia przez tworzone programy sensów zdań. Dotyczy to w głównej mierze translatorów. Do tej pory komputerowe słowniki potrafiły rozpoznać i przetłumaczyć tylko konkretne słowa, nie radziły sobie jednak z ich wieloznacznością. Aby skutecznie wspomagać tłumaczenia konieczne staje się zmuszenie komputera do rozpoznania kontekstu w jakim występuje dane słowo.

Jednak nie tylko tłumaczom niezbędne jest rozpoznawanie znaczeń wyrazów na podstawie kontekstu ich użycia. Problem jest aktualny we wszystkich zagadnieniach związanych z szeroko rozumianą eksploracją danych w nieustrukturalizowanych źródłach tekstowych oraz zasobach sieci Internet. Z pewnością dla każdego internauty takie narzędzie rozumiejące sensy wyrazów i całych zdań miałoby niebagatelne znaczenie. Jak często denerwujemy się na ulubioną wyszukiwarkę internetową na miliony odnośników, które nie mają związku z szukaną frazą. Gdyby wyszukiwarka potrafiła wywnioskować znaczenie wyrazu na podstawie znaczeń pozostałych słów znajdujących się we frazie, otrzymalibyśmy wtedy w wyniku tylko te odnośniki, które dokładnie odzwierciedlałyby nasze oczekiwania.

Problem dysambiguacji nie jest jednak trywialny i pracuje nad nim wiele grup badawczych. Obecnie dzięki projektowi WordNet, który został rozwinięty przez Cognitive Science Laboratory na Uniwersytecie w Princeton, ogólnie dostępny jest słownik języka angielskiego posiadający wszystkie znaczenia słów. Ponadto znaczenia te zostały usystematyzowane, wzajemnie powiązane i pogrupowane. Umożliwia to łatwe i szybkie wyszukiwanie danego słowa i jego wszystkich znaczeń.

Podstawowa struktura WordNetu opiera się na pojęciu synsetu, który reprezentuje konkretne znaczenie konkretnego słowa. Słownik WN jest zbiorem synsetów połączonych relacjami leksykalnymi z wagami - znaczenia które w jakiś

sposób się ze sobą wiążą są połączone. Przykładowe relacje łączące rzeczowniki to: - synonimiczna i antonimiczna - homonimia - określająca różne znaczenia słowa o tej samej pisowni - hiponimia - zwana też relacją „is" mówiąca że jeden wyraz jest bardziej ogólnym opisem innego określenia, np. „tree" jest hiponimem „plant". Odwrotnością hiponimii jest hiperonimia czyli bardziej szczegółowy opis danego określenia. - meronimia - zwana też relacją „has a" definiująca wyraz jako część innego znacznia np. „tire" i „car". Odwrotnością meronimii jest holonimia.

Istnieje również polski odpowiednik WN. Instytut Informatyki Stosowanej na Politechnice Wrocławskiej prowadzi projekt finansowany przez Ministerstwo Nauki i Szkolnictwa Wyższego o nazwie "Automatyczne metody konstrukcji sieci semantycznej leksemów polskich na potrzeby przetwarzania języka naturalnego". Celem tego projektu jest półautomatyczna konstrukcja słowosieci poprzez opracowanie algorytmów automatycznego wykrywania relacji semantycznych języka polskiego, które w połączeniu z pracą lingwistów, dadzą efektywną metodę budowy sieci relacji semantycznych. Oficjalna strona projektu to http://plwordnet.pwr.wroc.pl/main/?lang=pl .

2 Opis rozwiązania

Dla celów niniejszego artykułu powstał program, którego zadaniem jest zasymulowanie wyszukiwania najlepiej pasujących do siebie sensów rzeczowników występujących w zdaniu. W projekcie wykorzystana została istniejąca już biblioteka JWNL(Java WordNet Library), która udostępniana jest na zasadzie licencji GPL jako oprogramowanie Open Source. Przygotowany program pobiera na wejściu zdanie w języku angielskim. Zdanie to może być wielokrotnie złożone i może zawierać wszystkie rodzaje znaków przystankowych.

W pierwszej fazie analizy usuwane są wszystkie zbędne znaki ze zdania i wyodrębniane zostają pojedyncze słowa. Następnie spośród wyodrębnionych słów wybierane są wszystkie rzeczowniki. Dla każdego z rzeczowników wyszukiwana jest jego forma podstawowa, tj. zmieniana jest liczba mnoga wyrazu na pojedynczą oraz rzeczownik doprowadzany jest do formy mianownikowej. Krok ten jest niezbędny dla prawidłowego wyszukania znaczeń danego wyrazu.

Gdy już mamy wszystkie rzeczowniki w formie podstawowej wyszukujemy wszystkie ich znaczenia. Jako wynik tego działania otrzymujemy zbiór znaczeń dla każdego rzeczownika z danego zdania. Teraz chcemy zbadać jak blisko powiązane są ze sobą znaczenia wyrazów. W tym celu musimy wyszukać stopień podobieństwa pomiędzy każdą parą znaczeń dla każdej pary rzeczowników. Oczywiście nie bierzemy pod uwagę par znaczeń pochodzących od tego samego wyrazu.

Stopień podobieństwa dwóch znaczeń wyznaczany jest na dwa sposoby. Pierwszy to sposób porównania zaimplementowany przez bibliotekę JWNL, który opiera swoją ocenę bliskości dwóch znaczeń na podstawie obliczenia odległości pomiędzy nimi na grafie powiązań znaczeń stworzonym przez Wordnet. Im mniej kroków potrzebujemy, aby w grafie dojść z jednego punktu do drugiego

tym bliższe sobie są oba znaczenia. Drugim sposobem jest porównanie oparte na podstawie algorytmu Leska. Polega on na porównywaniu ilości podobnych słów w definicjach danych znaczeń. Im więcej słów mają wspólnych dwie definicje tym bardziej do siebie są podobne oba znaczenia.

Kolejnym krokiem algorytmu jest wybranie dla każdego z wyrazów takiego sensu, aby jego znaczenie było spójne z sensami innych wyrazów w zdaniu. Możemy to osiągnąć dzięki macierzy podobieństwa, która powstała podczas porównywania wszystkich sensów w poprzednim kroku. Macierz ta składa się dwóch wymiarów. Kolejne kolumny i wiersze reprezentują kolejne znaczenia wydobyte z wszystkich rzeczowników ze zdania. Na przecięciu danej kolumny i wiersza znajduje się liczba określająca stopień bliskości obu znaczeń. Im liczba ta jest mniejsza, tym bliższe sobie są opisywane znaczenia. Teraz wystarczy z tej macierzy wybrać najlepszą kombinację sensów, czyli taką która będzie miała najmniejszą sumę liczb w macierzy. Tak wybrana grupa sensów jest wynikiem i rozwiązaniem dysambiguacji.

Inną funkcją programu jest wyszukanie znaczenia danego słowa najbardziej zbliżonego do znaczenia znalezionego w innym słowniku. Wyszukiwanie to opiera się podobnie jak algorytm Leska na porównaniu liczby wspólnych wyrazów w obu definicjach.

Program posiada interfejs graficzny, który umożliwia swobodne wprowadzanie danych oraz obserwowanie wyników jego działania.

3 Przeprowadzone eksperymenty

W celu przetestowania zaimplementowanych rozwiązań wykorzystano 42 pary wyrazów testowych. Wszystkie pary powstawały ze słowa wieloznacznego i drugiego słowa, które uszczegóławiało znaczenie pierwszego słowa. Testy przeprowadzone zostały dla trzech algorytmów. Pierwsze dwa algorytmy to algorytmy zaimplementowane w przygotowanym oprogramowaniu. Pierwszy z nich to algorytm wewnętrzny biblioteki JWNL, drugi to nasza implementacja algorytmu Lesk. Ostatnim z testowanych algorytmów był algorytm udostępniony przez fińską grupę badawczą. Szczegółowe informacje na temat tego algorytmu można uzyskać na stronie www.teamapoint.com. Niestety strona ta napisana jest w języku fińskim i nie byliśmy w stanie doszukać się opisu algorytmu jaki zastosowano do rozwiązywania dysambiguacji. W wyniku przeprowadzonych eksperymentów, lepszym z algorytmów naszego programu okazał się algorytm wewnętrzny biblioteki JWNL. Okazał się on skuteczny w prawie 65% przypadków. Wersja algorytmu Leska okazała się skuteczna zaledwie w połowie przypadków. Zaobserwowaliśmy również bardzo ciekawą sytuację, która nie była błędem działania algorytmu, ale miała miejsce przy analizie tylko dwóch słów.W przypadkach tych algorytmy udzielały sensownych odpowiedzi, ale nie były one zgodne z oczekiwaniami. Przypadek ten został zaobserwowany dla pary słów ”sister” i ”brother”. Przez JWNL słowa te zostały zinterpretowane jako brat i siostra w rozumieniu rodziny, natomiast przez algorytm Lesk jako zakonnik i zakonnica. Bardzo ciekawe wyniki otrzymaliśmy również dla pary słów ”horse” i ”gym”. Okazuje się, że żaden z algorytmów nie potrafił zaklasyfikować tego znaczenia

”konia” jako konia gimnastycznego. Jest to dodatkowo dziwne, że nawet algorytm Lesk badający definicje znaczeń nie potrafił znaleźć wystarczającej liczby wspólnych wyrazów w definicjach obu słów. Ciekawą obserwacją jest fakt, że w ogromnej większości przypadków przynajmniej jeden z algorytmów okazywał się skuteczny. Niewiele przypadków sprawiło trudność obu algorytmom. W poniższej tabelce przedstawione zostały wyrazy, których dysambiguacje były najlepiej wykryte przez algorytmy.

Najlepiej rozpoznane znaczenia wyrazów

JWNL		Lesk	
słowo	procent wykrytych znaczeń	słowo	procent wykrytych znaczeń
club	100%	cap	100%
goal	100%	film	100%
net	100%	net	66%
base	75%	horse	66%
score	66%	score	66%
horse	66%	date	66%

Najgorzej rozpoznane znaczenia wyrazów

JWNL		Lesk	
słowo	procent wykrytych znaczeń	słowo	procent wykrytych znaczeń
pole	33%	space	0%
bow	33%	bow	0%
tree	50%	cap	0%
terminal	50%	pole	33%

Jak widać na podstawie załączonych tabel najtrudniej algorytmom uchwycić dokładny sens wyrazu, gdy znaczenia wyrazu są bardzo do siebie zbliżone. Dobrym przykładem jest tutaj słowo ”bow”. Ponieważ jego znaczenia bardzo blisko związane są z różnymi rodzajami łuków, i słowa te należą do bardzo bliskich kategorii, dlatego poszukiwanie odległości pomiędzy dwoma znaczeniami jest utrudnione. Należy zwrócić uwagę, że badane znaczenia posiadały dosyć spore objaśnienia, niemniej brakowało w nich słów kluczowych, na podstawie których algorytm Lesk mógłby skojarzyć odpowiednie sensy wyrazów.

Jedynym zewnętrznym algorytmem jaki podlegał analizie był wspomniany algorytm grupy ”teemapoint”. Algorytm ten umożliwia analizę znaczenie tylko rzeczowników, ale i innych części mowy. Spodziewaliśmy się, że będzie on skuteczniejszy od naszej implementacji. Okazało się jednak, że dla naszych danych testowych algorytm ten nie uzyskał nawet 50% poprawności wyników. Było to bardzo zaskakujące. Głębsze analizy doprowadziły do wniosków, że algorytm ten znacznie lepiej sobie radzi gdy dysponuje oprócz rzeczowników również czasownikami i przymiotnikami. Gdy nasze grupy rzeczowników zostały rozwinięte do pełnych zdań okazało się, że wyniki zbliżyły się znacznie do oczekiwanych. Jak widać wpływ powiązań pomiędzy czasownikami a rzeczownikami znacznie wpływa na poprawność generowanych wyników przez ten algorytm.

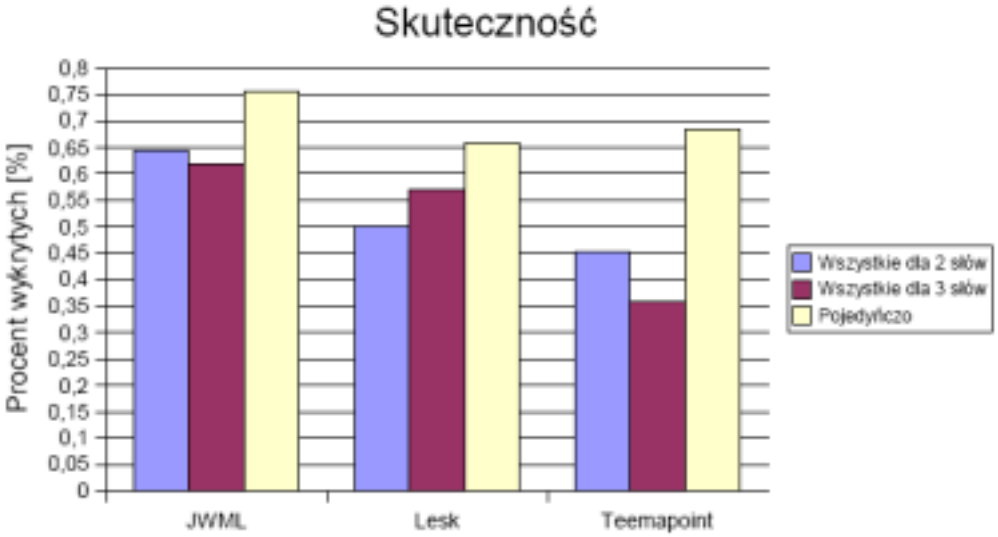
Przeprowadziliśmy również testy dla przykładów składających się większej liczby wyrazów. Pokazały one, że ogólnie wraz ze zrostem liczby wyrazów w

zdaniu poprawia się liczba poprawnie wykrywanych znaczeń wyrazów. W szczególności dotyczy to algorytmu Lesk. Jak widać algorytm ten posiadając większą ilość definicji może w lepszy sposób znaleźć najlepsze znaczenia wyrazów.

W powyższych testach za poprawną odpowiedź algorytmu uznawaliśmy tylko taką, która podawała poprawną odpowiedź dla wszystkich wyrazów. Jednak gdy będziemy rozpatrywać poprawność wykrywania znaczeń dla poszczególnych wyrazów z osobna wyniki działania algorytmów ulegną diametralnej poprawie. Dla tak ocenianej poprawności otrzymaliśmy następujące wyniki:

- JWNL – 75,61%
- Lesk – 65,85%
- Teemapoint – 68,29%

Ciekawostką, jaką udało nam się zaobserwować było, że dla niektórych wyrazów, a konkretnie niektórych znaczeń tych wyrazów, znaczenia badanego wyrazu było poprawnie interpretowane, a wyraz który miał precyzować znaczenie tego słowa był źle interpretowany. Ponadto należy zwrócić uwagę, że dla badanych par wyrazów praktycznie dla żadnej z par nie zaistniała sytuacja, że oba wyrazy zostały źle zinterpretowane. Przeważnie przynajmniej jeden z nich był zawsze dobrze zinterpretowany. Rysunek **??** podsumowuje wyniki uzyskane przez nas podczas badania.



Rysunek 1: Wyniki eksperymentu

Reasumując testy, najlepszym algorytmem okazał się algorytm wyszukiwania bliskości znaczeń zaimplementowany w bibliotece JWNL. Algorytm ten okazał się bezkonkurencyjny w porównaniu do pozostałych. Zwiększenie ilości wyrazów w zdaniu nie zmienił znacząco poprawności wykrywania znaczeń rzeczowników.

Miało to miejsce gdy badaliśmy dwa pierwsze algorytmy. Dla algorytmu grupy Teemapoint dodanie do zdania większej liczby wyrazów znacząco pogorszyło jakość wyników. Jeżeli weźmiemy pod uwagę procent poprawności wykrywania znaczeń dla poszczególnych wyrazów w zdaniu, a nie całych zdań, ogólna ocena poprawności wykrywania znaczeń wzrasta. Dla algorytmów z naszego programu stosunek poprawnie wykrywanych znaczeń nie zmienił się. Okazało się jednak, że gdy bierzemy pod uwagę pojedyncze słowa algorytm Teemapoint dorównał algorytmom zastosowanym w programie.

Przeprowadzone zostały również testy drugiego narzędzia wyszukującego najlepiej pasujące znaczenie wyrazu na podstawie podanego wyrazu i nowej definicji. W trakcie testów narzędzie to działało bezbłędnie wyświetlając oczekiwane definicje. Zgodnie z oczekiwaniami im więcej słów zostało podanych w nowej definicji słowa tym lepiej dopasowywane były stare definicje. Podsumowując narzędzie działało zgodnie z oczekiwaniami.

4 Wnioski

Problem dysambiguacji jest nadal bardzo trudnym i skomplikowanym zagadnieniem. Jak pokazały nasze eksperymenty próba identyfikacji znaczeń wyrazów tylko na podstawie rzeczowników może okazać się niewystarczająca. Osiągnięte wyniki przy analizie tylko rzeczowników dają wyniki poprawności na poziomie 2/3. Naszym zdaniem rozszerzenie analizy znaczeń wyrazów o analizę czasowników i przymiotników powinno w sposób znaczący poprawić skuteczność wykrywania poprawnych znaczeń wyrazów. Dowodem na to jest wynik jaki uzyskała grupa badawcza Teemapoint. Ich narzędzie potrafiące odnajdywać najlepiej pasujące znaczenia w danym kontekście o wiele lepiej radzi sobie dla całych zdań, które posiadają zarówno rzeczowniki jak i przymiotniki. Należy zauważyć, że drugie narzędzie do wyszukiwania najlepiej pasujący znaczeń dla zadanego wyrazu i jego definicji działa bardzo dobrze, osiągając poprawność sięgającą 100%. To drugie z narzędzi może zostać rozwinięte o możliwość rozbudowywania słownika WordNet o nowe znaczenia. Dzięki temu WordNet może zostać uaktualniony o nowe wyrazy i znaczenia pochodzące z innych słowników i encyklopedii. Powstałe narzędzie może stać się dobrą podstawą do rozszerzania jego funkcjonalności w przyszłości, w celu otrzymania dokładniejszych wyników, dodając do niego analizę innych części mowy niż rzeczowniki.