ZESZYTY NAUKOWE WYDZIAŁU ELEKTRONIKI, TELEKOMUNIKACJI I INFORMATYKI POLITECHNIKI GDAŃSKIEJ

Nr 10 Seria:ICT Young 2012

SEMANTYCZNE ZNACZNIKOWANIE ARTYKUŁÓW WIKIPEDII SYNSETAMI SŁOWNIKA WORDNETA[1]

Łukasz Borek, Julian Szymański

Politechnika Gdańska, Katedra Architektury Systemów Komputerowych

Streszczenie

Artykuł opisuje algorytmy ujednoznacznia artykułów za pomocą synsetów słownika WordNet. Opisano alternatywne podejścia i wskazano ich braki. Przedstawiono nowe, autorskie podejście jak i wykazano jego wyższość nad istniejącymi do tej pory rozwiązaniami.

1. WSTĘP

Dynamiczny rozwój Wikipedii to doskonały przykład na to, jak szybko rozwija się World Wide Web – czyli świat Internetu. Rosnące zasoby stawiają przed naukowcami i specjalistami z dziedziny przetwarzaniu tekstu wiele wyzwań.

Jak większość danych w sieci Web, Wikipedia jest zapisana w języku naturalnym, który w chwili obecnej zrozumiały jest tylko i wyłącznie dla człowieka, a dla maszyny jest on bardzo trudny w samym przetwarzaniu. W pracy zostały przedstawione algorytmy, które pozwolą na powiązanie Wikipedii ze słownikiem WordNet. Istnienie takiego połączenia umożliwia nałożenie na strukturę uporządkowanego słownika semantycznego artykułów, które nie są powiązane ze sobą żadnymi relacjami. Dzięki temu można uniknąć złożonego przetwarzania języka naturalnego, a badane są tylko artykuły powiązane różnymi relacjami ze sobą.

[1]Praca zostala sfinansowana przez Narodowe Centrum Badań i Rozwoju w ramach grantu N N 516 432338

W przypadku posiadania dużego zbioru artykułów i równie dużego zbioru synsetów, mówi się o przypisaniu artykułowi jednego bądź kilku odpowiadających mu synsetów. Takie przypisanie nazywa się **ujednoznacznieniem**.

Proces ten polega na identyfikacji sensu słowa w przypadku, gdy ma ono kilka znaczeń (występuje zjawisko polisemii). Jest to jeden z otwartych problemów przetwarzania języka naturalnego. Obecne algorytmy cechują się, w przypadku bazujących na częstości występowania słów, skutecznością na poziomie od 50% do ok 90%.

2. DANE I PODEJŚCIA DO UJEDNOZNACZNIANIA

2.1. WordNet

WordNet[8] jest leksykalną bazą danych języka angielskiego, która zawiera pogrupowane słowa w tzw. **Synsety**. Pojedynczy synset składa się ze zbioru synonimów oraz definicji. W słowniku można znaleźć różnego typu relacje zachodzące pomiędzy synsetami. Wordnet to połączenie słownika wyrazów bliskoznacznych i zwykłego słownika, zawierającego definicje danego słowa. Połączenie tych dwóch źródeł wiedzy daje ogromne możliwości. Projekt ten znalazł wiele zastosowań, głównie w dziedzinach informatyki związanych z przetwarzaniem tekstu i sztuczną inteligencją.

Wersja 3.0 WordNeta zawiera ponad 155 tysięcy słów pogrupowanych w około 117 tysięcy synsetów [8]. W pracy użyto tylko rzeczowniki, pomimo tego, że WordNet zawiera klasyfikację również dla czasowników, przymiotników i przysłówków. W interesującej nas grupie znajdziemy około 82 tysięcy synsetów.

2.2. Wikipedia

Wikipedia jest to internetowy projekt encyklopedii tworzonej nieprzerwanie od 15 stycznia 2001 roku. W chwili obecnej istnieje ponad 200 edycji językowych tego słownika, ilość wszystkich haseł przekroczyła już liczbę 14 milionów. W Wikipedii angielskiej, która odgrywa główną rolę w tej pracy, można znaleźć około 3,8 mln różnych haseł. Na potrzeby testów została wykorzystana uproszczona wersja Wikipedii - *Simple English Wikipedia*.

Podstawową jednostką Wikipedii jest artykuł (albo strona), która opisuje dany fragment wiedzy. Składa się z określonych części i podlega ustalonej składni zaproponowanej przez Wikipedię.

2.3. YAGO

YAGO [9] to ontologia zbudowana na podstawie Wikipedii i Wordneta, zaprojektowana i wdrożona na Instytucie Informatyki im. Maxa Planca. W czerwcu 2010 roku ontologia zawierała ponad 2 miliony obiektów oraz 20 milionów faktów operujących na danych obiektach [8]. Projekt cały czas jest rozwijany, a ontologia jest wytwarzana

automatycznie za pomocą specjalnie skonstruowanych algorytmów do pozyskiwania informacji z Wikipedii i łączenia ich z WordNetem [1, 2]. Doświadczalnie stwierdzono, że poprawność automatycznego doboru faktów jest na poziomie 95% i - co jest sukcesem tego przedsięwzięcia, biorąc pod uwagę ogrom wiedzy zawartej w projekcie oraz sposób jej pozyskiwania [2].

2.4. Algorytm statystyczny

Algorytm statystyczny jest to algorytm opracowany w 2005 roku przez zespół pod kierownictwem Mari Ruiz-Casado na Autonomicznym Uniwersytecie Madrytu [6]. Publikacja, w której został opisany ten algorytm, jest często cytowana i można znaleźć ją praktycznie w każdej pracy związanej z łączeniem Wikipedii z WordNetem.

Niestety podany algorytm jest możliwy do zastosowania na wąskiej grupie artykułów, które stanowią w przybliżeniu 1% całości[2] (15000 artykułów). Na tym zbiorze uzyskano skuteczność rzędu 93.89%.

2.5. Algorytm bazujący na współwystępowaniu słów

Algorytm ten wykorzystuje tylko i wyłączenie powiązania słów w artykule z definicjami synsetu [2]. W algorytmie zastosowano dwa podejścia – pierwsze j podobne do poprzedniego algorytmu i najbardziej intuicyjne, próba ujednoznacznia artykułów za pomocą synsetów WordNeta. Drugie podejście opiera się o ujednoznacznienie synsetów za pomocą artykułów, czyli jest próbą wykonania zadania od drugiej strony. Dodatkowo algorytm stara się dokonać ujednoznacznienia masowo – czyli dla kilku artykułów jednocześnie.

W stosunku do poprzedniego algorytmu można w nim znaleźć kilka różnic. Przede wszystkim masowe ujednoznacznianie, a także od drugiej strony, próbę zastosowania podejścia przyporządkowania synsetowi artykułu. Wyniki badania poprawności zwróconych rezultatów, jakie otrzymali autorzy w swojej pracy, mieszczą się w przedziale od 39.51% do 49.28% [2].

3. AUTORSKI ALGORYTM UJEDNOZNACZNIAJĄCY

W poniższym rozdziale zostały zaprezentowane autorskie algorytmy ujednoznaczniania. Są to trzy różne podejścia, które w połączeniu ze sobą, jak pokazały badania, dały skuteczność na poziomie 89% przy zachowaniu dużego zbioru wyników.

3.1. Algorytm siłowy

Pierwszy algorytm jest najprostszym podejściem, zainspirowanym algorytmem wykorzystanym w YAGO do mapowania nazwy kategorii na synsety Wordneta. Ogólny sposób postępowania jest następujący: dla każdego tytułu artykułu – T_A:

1. Sprawdź czy istnieje synset zawierający w zbiorze synonimów T_A.
2. Jeśli istnieje tylko jeden synset, to wybierz go.
3. W przypadku wielu wyników wybierz ten, który jest najpopularniejszy [2].

Jak wykazały rezultaty badań jest to podejście bardzo skuteczne. Jego skuteczność wynika z faktu, iż dla bardzo specyficznych i specjalistycznych artykułów nie otrzyma się żadnego lub tylko jeden wynik wyszukiwania. Nie zostanie więc popełniony dla nich błąd. Podejście identyczne zostało zastosowane w algorytmie proponowanym przez zespół M. Ruiz – Casado [6].

W wyniku doświadczenia stwierdzono, że podczas szukania tytułu artykułu w zbiorze synonimów (nazwie synsetu) otrzymano co najmniej jeden synset dla 12 090 tytułów. Co stanowi ok. 20% całej Simple English Wikipedii. Z tego w wyniku wyszukiwania, w pierwszym kroku algorytmu, dla 8 161 tytułów otrzymano tylko jeden synset, co wskazuje na jednoznaczne przypisanie. Przedstawione powyżej podejście może wydawać się bardzo intuicyjne, powód leży w konstrukcji tytułu artykułów Wikipedii. W przypadku, gdy mamy do czynienia z popularnym artykułem, istnieją duże szanse na to, że w WordNecie jest podobny element opisany w taki sam sposób. Jeśli artykuł opisujący określoną tematykę może mieć wiele znaczeń,wtedy najprawdopodobniej w Wordnecie znajdziemy więcej synsetów powiązanych z danym tytułem. W tym wypadku istnieje duże prawdopodobieństwo, że w Wikipedii znajdą się również takie artykuły, ale o innym, lekko zmodyfikowanym tytule.

3.2. Algorytm nawiasowy

Drugi algorytm opiera się na specyficznej konstrukcji tytułu artykułu [3]. Otóż możemy znaleźć grupę artykułów, które w tytule mają już swoje ujednoznacznienie. Na przykład: **Led Zeppelin (album)**. Artykuł ten informuje o albumie zespołu, który nosi taką samą nazwę. Nazwa w nawiasie wskazuje jednoznacznie na tematykę danego artykułu w kontekście pierwszego słowa. Niestety nie jest to aż tak trywialne zadanie, aby wystarczyło wybrać części nawiasu i opisać jej jako synset. Dlatego wypracowane zostały kroki:

Oznaczenia:

- T - tytuł artykułu postaci $T_1(T_2)$, gdzie T_1 to fraza główna, a T_2 to fraza uogólniająca. Dla przykładu *Led Zeppelin (album)*, T_1 = *Led Zeppelin*, T_2 = *album*.
- Synset(T) – lista synsetów zawierajacych T w swoim zbiorze synonimów.

Dla każdego artykułu postaci $T_1(T_2)$:

1. Tworzymy listy synsetów dla ciągów T_1, T_2: $L_1 = \text{Synsety}(T_1)$, $L_2 = \text{Synsety}(T_2)$.

[2]Liczba wystąpień w tekstach, jest statystyką podawaną w synsecie

2. Gdy $|L_1| = 0$ i $|L_2| = 0$ algorytm kończy procedurę i zwraca null.
3. Gdy $|L_1| = 0$ (lista dotycząca głównej frazy jest pusta): wybrana zostaje pierwsza pozycja z listy synsetów L_2.
4. W przeciwnym wypadku:
   (a) Jeśli T_2 należy do definicji któregoś z synsetów, to zwracamy ten synset,
   (b) Jeśli żaden nie zawiera T_2, to wybieramy pierwszy synset z listy L_2.

3.3. Algorytm uogólniający

Trzeci algorytm został nazwany uogólniającym, gdyż dzięki niemu można poznać tematykę artykułu. Zgodnie z podejściem opisanym w algorytmie bazującym na współwystępowaniu słów, do przetwarzania bierze się tylko początek artykułu, a mianowicie pierwsze zdanie [4]. W większości artykułów w tym zdaniu jest zapisana najważniejsza informacja dla przebiegu algorytmu na tym etapie. Aby dobrze zrozumieć jego istotę, należy przyjrzeć się przykładowemu pierwszemu zdaniu z artykułu *Dog*:

The dog (Canis lupus familiaris) is a **mammal** from the family Canidae.

Głównym problemem tego zagadnienia jest wydobycie, z pierwszego zdania artykułu, relacji **is a**, a właściwie wyszukiwania dopełnienia. Poniżej został zaproponowany autorski algorytm dokonujący analizy zdania i wydobywania z niego dopełnienia zdania po relacji **is a**.

Procedura Is_a

Poniżej został przedstawiony proponowany algorytm wyszukujący relację is_a i zwracający dopełnienie z tej relacji:

1. Pozbądź się wszystkich niepotrzebnych znaczników notacji wikii (szczególnie wszelkich „infoboxów" oraz obrazków i plików, które są często zamieszczane na początku artykułu).
2. Wyodrębnij z artykułu pierwsze zdanie.
3. Zastosuj tagery do odzielenia wyrazów i lematyzatory tekstu w celu zamiany wyrazów na ich podstawowe formy (w_1,w_2,...,w_n).
4. Zastosuj do oddzielonych wyrazów narzędzie do oznaczania części mowy.
5. Znajdź w zdaniu pierwsze wystąpienie czasownika „to be" (angielskie być, w jakiejkolwiek formie i czasie).
6. Jeśli nie występuje czasownik „to be", to skończ działanie algorytmu.

7. Dla każdego kolejnego wyrazu w_i:
   (a) Jeśli w_i jest rzeczownikiem i nie istnieje kolejny wyraz w_{i+1} $(i = n)$ to zwróć w_i jako wynik.
   (b) Jeśli w_i jest rzeczownikiem i kolejny wyraz (w_{i+1}) istnieje, i nie jest rzeczownikiem to zwróć w_i jako wynik.
   (c) Jeśli w_i jest nie jest rzeczownikiem lub jeśli kolejny wyraz w_{i+1} jest rzeczownikiem, to wróć do punktu 7.

3.4. Sposób zbierania danych

Wynik zbiorczy, jest to zbiór powstały przez połączenie ze sobą wyników poszczególnych algorytmów. W pierwszej kolejności są brane wyniki otrzymane z działania algorytmu *siłowego*, następnie *nawiasowego*, a na samym końcu *uogólniającego*. Oznacza to, że w przypadku kiedy np. algorytm siłowy ujednoznacznił tytuł do synsetu A, a algorytm uogólniający do synsetu B, to wybierany jest synset A.

4. OCENA I REZULTATY

4.1. Metryki oceny

Wynik działania algorytmu zostały ocenione w trzech obszarach:

Pokrycie zbioru – miara, która wyraża procentową wartość przypisanych artykułów do konkretnych synsetów, bez rozróżniania czy jest to poprawne czy błędne powiązanie.

Poprawność pokrycia – miara ta wyraża procentową wartość poprawnych powiązań, które zwrócił algorytm.

Średni czas przypisania – miara określająca czas działania algorytmu, uzyskana przez podzielenie całkowitego czasu przez liczbę artykułów dla, których podjęto próbę ujednoznacznienia.

F-miara - jest to miara stosowana w dziedzinach przetwarzania informacji, a w szczególności w ocenie działania wyszukiwarek internetowych, łączy ze sobą dwa pojęcia poprawność pokrycia (*precyzja*), pokrycie zbioru (*zwrot*). Wartość F-miary wyliczana jest za pomocą wzoru $F = 2 \cdot \frac{precyzja \cdot zwrot}{precyzja + zwrot}$ [5].

4.2. Rezultaty

W tabeli 1 przedstawiono wyniki dla działania algorytmów, osobno oraz dla wyniku zbiorczego – ostatecznego. W zbiorze testowym (Simple English Wikipedia) znalazło się 60 285 artykułów, a zasada wybierania wyników do oceny została opisana w poprzednim rozdziale. W kolumnach poniższej tabeli:

Tabela 1: Rezultaty zbiorcze dla poszczególnych algorytmów

Nazwa algorytmu	Pokrycie zbioru testowego	Poprawność pokrycia	Wartość F-miary	Czas działania	Średni czas przypisania
Siłowy	20% (12102)	98 ± 1.8%	33 ± 0.5%	1h	0.06
Nawiasowy	2.5% (1480)	86.8 ± 4.2%	4.1 ± 0.2%	3h	3.4
Uogólniający	84.9 % (51182)	85.0 ± 3.6%	84 ± 2%	100h	6
Wynik zbiorczy	89.5% (54199)	89.5± 2.2%	89 ± 1.2%	1min	0.001

Algorytm siłowy, który wykorzystuje bardzo prostą zasadę, daje jednocześnie wysoką skuteczność i pokrywa się on z najczęściej używanymi artykułami Wikipedii. W pierwszym doświadczeniu otrzymano 12102 ujednoznacznień, co stanowi ok. 20% zbioru ujednoznacznianego. W tym przypadku otrzymana bardzo wysoka poprawność jest ważna, gdyż można się spodziewać, że zostały ujednoznacznione artykuły najczęściej wyszukiwane i przeglądane przez czytelników. Używając algorytmu wyszukującego tytuł artykułu w synonimach synsetu, uzyskano ok. 20% ujednoznacznienia Simple English Wikipedia, z czego 8161 ujednoznacznień było jednoznacznych.

Algorytm nawiasowy operuje na bardzo specyficznym rodzaju tytułów artykułów, stąd tak niskie pokrycie zbioru uogólnianego. W samej Simple English Wikipedii jest tylko 3125 takich artykułów, co stanowi około 5.2% jej całości. Są to tytuły, niewystępujące w zbiorze wyników, które daje Algorytm siłowy.

Użycie algorytmów I i II pozwoliło ujednoznacznić około 22% tytułów Simple English Wikipedia, wyjaśnia to fakt, że granulacja wiedzy w Wikipedii i WordNecie jest różna. Sytuacja ta wymusza to uogólnianie niektórych artykułów do szerszych pojęć. Na tej zasadzie działa algorytm uogólniający, potwierdzają to wyniki w tabeli 1 Wyniki eksperymentów. Algorytm ten został uruchomiony na całej Simple English Wikipedii i dał jakikolwiek wynik dla prawie 85% artykułów. Skuteczność przypisania można podnieść przez doskonalenie algorytmu wyodrębniającego relację **is_a**. W pracy zastosowano bardzo proste autorskie podejście, które dało zaskakująco dobrą jakość.

Wynik ostateczny powstał dzięki zebraniu wyników poszczególnych algorytmów. Czas przypisania jest nieporównywalnie krótszy do pozostałych, gdyż bazuje ona na wynikach obliczeń poprzednich algorytmów.

5. PODSUMOWANIE I DYSKUSJA

W wyniku prac powstała procedura uogólniająca składająca się z kilku algorytmów o różnej specyfice. Dzięki zastosowaniu różnych podejść udało się rozwiązać problem zarówno od strony maksymalizowania poprawności wyniku jak i uniwersalności - dla różnego typu artykułów jest możliwe uogólnienie. Takie pojdeście zapewniło ujednoznacznienia 90% zbioru testowego przy zachowaniu skuteczności na poziomie 89.5%.

Istnieje wiele możliwości rozbudowy procedury uogólniania, przede wszystkim nasuwa się pomysł rozbudowy jej przez dokładanie kolejnych algorytmów operujących na specyficznych artykułach w Wikipedii (np.: stron ujednoznaczniających, list tematycznych, itp.).

Jako osobne zagadnienie pozostaje kwestia implementacji metody i jej wydajności tak aby procedura była w stanie działać na zbiorze całej Wikipedii. Jest to niewątpliwie wyzwanie aby sprostać liczba artykułów do ujednoznacznienia.

Bibliografia

[1] Suchanek F., Kasneci G.: *Yago: A large ontology from wikipedia and wordnet*, Web Semantics: Science, Services and Agents on the World Wide Web 2008

[2] Suchanek F., Kasneci G., Weikum G.: *YAGO: A core of Semantic Knowledge Unifying WordNet and Wikipedia*, Proceedings of the 16th international conference on World Wide Web 2007

[3] Mihalcea R.: *Using Wikipedia for Automatic Word Sense Disambiguation*, University of North Texas 2006

[4] Szymański J., Kilanowski D.: *Wikipedia and Wordnet integration based on words co-occurrences*, Proceedings of the 30th International Conference Information Systems, Architecture and Technology 2009

[5] Manning Ch., Raghavan P.: *Introduction to Information Retrieval*, Cambridge University Press. 2008

[6] Ruiz-Casado M., Alfonseca E., Castells P.: *Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets*,Springer LNAI 3528, 2005

[7] `http://en.wikipedia.org`

[8] `http://wordnet.princeton.edu/`

[9] `http://www.mpi-inf.mpg.de/yago-naga/yago/`