Streszczenie rozprawy doktorskiej „Automatyczna ekstrakcja relacji semantycznych z tekstów w języku polskim”

Aleksander Smywiński-Pohl

13 września 2015

1 Wstęp

Praca dotyczy ekstrakcji informacji z polskich tekstów. Zasadniczym jej tematem jest rozpoznawanie relacji semantycznych w oparciu o automatycznie konstruowane wzorce ekstrakcyjne. Przedstawiono w niej również algorytm selekcji zdań, na podstawie których tworzony jest model ekstrakcji oraz algorytmy ujednoznaczniania i semantycznej klasyfikacji wyrażeń języka polskiego.

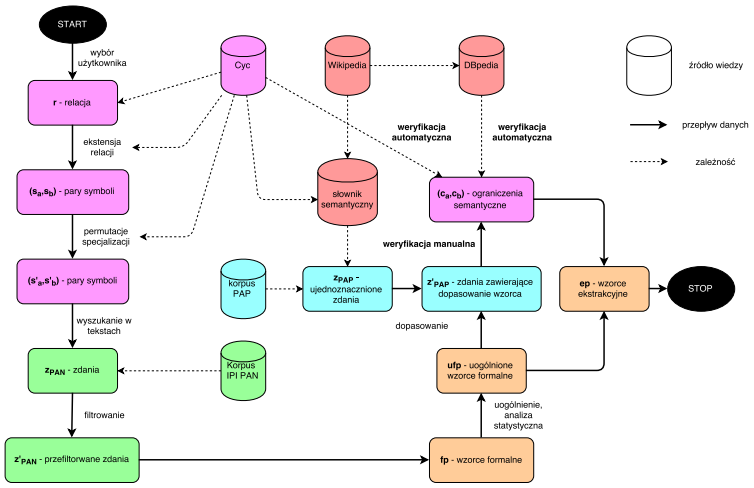
Wzorce ekstrakcyjne są konstruowane na podstawie przykładowych zdań zawierających wyrażenia połączone relacjami oraz wyposażane są w ograniczenia semantyczne zdefiniowane z wykorzystaniem pojęć ontologii Cyc [6, 5]. Ograniczenia określane są na podstawie trzech metod: ręcznej oceny zdań, predykatów ontologii Cyc oraz danych znajdujących się w DBpedii [1].

2 Teza

Teza pracy jest następująca: **możliwe jest skonstruowanie hybrydowego algorytmu ekstrakcji wybranych relacji semantycznych z tekstów w języku polskim, który:**

1. **dawałby wyniki bardziej precyzyjne niż te, otrzymywane za pomocą algorytmów statystycznych,**
2. **nie byłby ograniczony do pojedynczej dziedziny wiedzy,**
3. **wymagałby mniejszego nakładu pracy ręcznej, niż algorytm wytrenowane na ręcznie oznakowanym zbiorze uczącym.**

Tak postawiony cel algorytmu ekstrakcji osiągany jest za pomocą automatycznej konstrukcji wzorców ekstrakcyjnych. Szablon ekstrakcyjny jest traktowany jako zbiór cech morfologicznych, syntaktycznych oraz semantycznych, które muszą zostać spełnione, aby można było uznać, że poszukiwana relacja występuje w analizowanym tekście. Dopasowanie wzorca ekstrakcyjnego interpretowane jest jako wystąpienie relacji semantycznej, przez co, zgodnie z postawionym celem, możliwe jest określenie wystąpienia poszukiwanej relacji dla każdego zdania z osobna, podobnie jak ma to miejsce w przypadku ręcznie konstruowanych wzorców morfosyntaktycznych opisywane przez Piaseckiego i współpracowników [11].



Rysunek 1: Schemat procesu, w wyniku którego powstaje wzorzec ekstrakcyjny.

3 Algorytmy konstrukcji wzorców oraz ekstrakcji relacji

Algorytm konstrukcji wzorców ekstrakcyjnych jest następujący (rysunek 1):

1. Z ontologii Cyc wybierana jest relacja r, której wystąpienia mają być rozpoznawane w tekście.

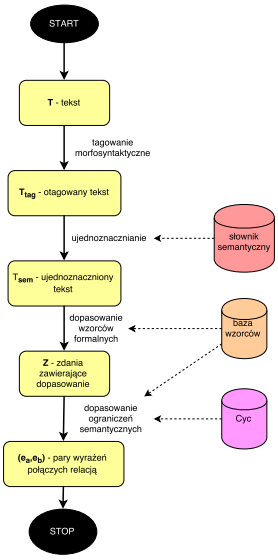
2. Na podstawie ekstensji relacji r generowane są pary symboli: (s_a, s_b), połączone tą relacją.

3. Dla każdej pary symboli (s_a, s_b) w korpusie tekstów udostępnionym przez Instytut Podstaw Informatyki Polskiej Akademii Nauk wyszukiwane są zdania z_{PAN}, w których współwystępują napisy powiązane z tymi symbolami.

4. Powstały zbiór zdań jest filtrowany i usuwane są zdania niespełniające kryteriów poprawności – powstaje zbiór z'_{PAN}.

5. Zdania należące do wynikowego zbioru poddawana są ograniczonej analizie morfosyntaktycznej, na podstawie której tworzone są wzorce ekstrakcyjne posiadające wyłącznie cechy morfologiczne i syntaktyczne, tzw. formalne wzorce relacji: fp.

6. Identyczne wzorce są utożsamiane dzięki czemu powstają uogólnione wzorce formalne ufp.



Rysunek 2: Schemat procesu dopasowywanie wzorców ekstrakcyjnych do tekstu.

7. Korpus tekstów zawierający notatki Polskiej Agencji Prasowej, z którego są ekstrahowane relacje semantyczne, jest ujednoznaczniany względem słownika semantycznego powstałego na bazie Wikipedii oraz ontologii Cyc. W wyniku powstaje zbiór zdań z_{PAP}.

8. Dla każdego unikalnego wzorca formalnego ufp w korpusie z punktu 7 wyszukiwane są zdania z'_{PAP} pasujące do tego wzorca, przy założeniu, że oba dopasowane argumenty zostały ujednoznacznione, a odpowiadające im symbole mają przypisane kategorie semantyczne.

9. Ze zbioru z'_{PAP} losowana jest próbka zdań, które są ręcznie oznaczana jako zawierające, bądź niezawierające wystąpienie relacji r. Na tej podstawie określane są ręczne ograniczenia semantyczne wzorca ekstrakcyjnego (mc_a, mc_b).

9'. Określenie ograniczeń semantycznych realizowane jest automatycznie, w oparciu o analizę relacji wstępujących w ontologii Cyc lub w DBpedii, w wyniku czego powstają alternatywne ograniczenia semantyczne (ac_a, ac_b).

10. Wynikowe wzorce ekstrakcyjne ep powstają w wyniku połączenia wzorców formalnych ufp z ograniczeniami semantycznymi (mc_a, mc_b) (wariant półautomatyczny) bądź (ac_a, ac_b) (wariant automatyczny).

Uzyskane w ten sposób wzorce ekstrakcyjne wykorzystywane są do rozpoznawania wystąpień relacji r w następujący sposób (rysunek 2):

1. Tekst T, w którym mają być rozpoznane relacje semantyczne jest tagowany morfosyntaktycznie z wykorzystaniem taggera Concraft [17]. W wyniku powstaje tekst T_{tag}.

2. Ten sam tekst ujednoznaczniany jest względem słownika semantycznego powstałego na podstawie Wikipedii i ontologii Cyc. W wyniku powstaje tekst T_{sem}, który zawiera tagowanie morfosyntaktyczne oraz semantyczne.

3. Do tekstu T_{sem} dopasowywane są wzorce formalne ufp. Wzorzec pasuje do tekstu jeśli wszystkie ograniczenia formalne są spełnione oraz oba argumenty relacji zostały ujednoznacznione względem słownika oraz posiadają przypisane kategorie semantyczne. W wyniku powstaje zbiór Z zawierających dopasowania wzorca formalnego.

4. Uznaje się, że dana para wyrażeń (e_a, e_b) jest połączona określoną relacją semantyczną, jeśli spełnione są ograniczenia semantyczne zdefiniowane we wzorcu ekstrakcyjnym. Spełnienie ograniczeń może być bezpośrednie (wtedy kategorie określone we wzorcu i przypisane wyrażeniom są identyczne), bądź pośrednio (kategorie przypisane wyrażeniom mogą być specjalizacjami ograniczeń określonych we wzorcu).

4 Algorytmy pomocnicze

Poza głównym algorytmem, którego celem jest budowa wzorców ekstrakcyjnych, autor opracował kilka algorytmów pomocniczych, niezbędnych do zbudowania kompletnego systemu ekstrakcji relacji.

4.1 Wybór przykładowych zdań

Punkt 2 algorytmu konstrukcji wzorców zakłada dostępność zbioru przykładowych par symboli, o których wiadomo, że pomiędzy nimi występuje zadana relacja semantyczna. Algorytmy ekstrakcji informacji posiłkujące się danymi zarodkowymi, zwykle wykorzystują bardzo niewielką liczbę takich przykładowych danych, aby w sposób iteracyjny odkrywać nowe wzorce ekstrakcyjne oraz nowe pary symboli połączone odpowiednią relacją. Tym niemniej wykonywanie wielu iteracji prowadzi zwykle do zjawiska nazywanego dryfem semantycznym [8], polegającego na tym, że jakość ekstrahowanych wzorców oraz par symboli istotnie się pogarsza z każdą iteracją.

Chcąc uniknąć tego problemu można zastosować różne metody statystyczne oraz semantyczne [8], pozwalające na rozpoznawanie par wyrażeń, które wprowadzają szum w danych. W prezentowanym algorytmie zastosowano nieco inne podejście – podstawowy pomysł polega na automatycznym rozszerzeniu pierwotnej listy par symboli, zanim algorytm zacznie poszukiwać odpowiadające im wzorce ekstrakcyjne. Możliwe jest to dzięki wykorzystaniu rozbudowanej taksonomii pojęć zbudowanej w oparciu o ontologię Cyc oraz słownika semantycznego zawierającego dużą liczbę nazw własnych wziętych z Wikipedii. Szczegółowy opis tego algorytmu znajduje się w publikacji [12].

4.2 Ujednoznacznianie sensu wyrażeń

W punkcie 7 algorytmu konstrukcji wzorców oraz w punkcie 2 algorytmu ekstrakcji następuje ujednoznacznienie wyrażeń językowych względem słownika semantycznego. Ujednoznacznianie sensu jest zagadnieniem dość złożony i jego skuteczność w dużej mierze uzależniona jest od wykorzystywanego słownika

semantycznego. Jeden z popularnych algorytmów wykorzystywany do ujednoznaczniania sensu względem angielskiego WordNetu – algorytm Leska [7, 2] – zakłada, że każdy symbol językowy wyposażony jest w tradycyjną definicję. Założenie to spełnione jest dla angielskiego WordNetu, ale nie jest spełnione dla polskiej Słowosieci [11], gdyż opisy symboli językowych w polskim WordNecie mają w przeważającej mierze charakter relacyjny[1].

Chcąc rozwiązać problem ujednoznaczniania sensu w sposób zadowalający, konieczne było wykorzystanie innego zasobu oraz algorytmu posługującego się innym słownikiem. Autor oparł się na wynikach Milnego oraz Wittena [10, 9], którzy wykorzystali Wikipedię jako podstawowy zasób leksykalny. Opracowali oni również algorytm zdolny do ujednoznaczniania wyrażeń względem artykułów Wikipedii. Choć został on opracowany dla języka angielskiego, pozwalał na łatwą adaptację dla innych języków – w tym języka polskiego.

Algorytm opracowany przez Milnego i Wittena został jednak poddany istotnym udoskonaleniom. W szczególności autor wykorzystał inną miarę pokrewieństwa semantycznego oraz zastosował dodatkowe cechy służące do budowy klasyfikatora, co istotnie przyczyniło się do poprawy rezultatów ujednoznaczniania. Szczegółowy opis tego algorytmu znajduje się w publikacji [14].

4.3 Określenie kategorii semantycznych symboli językowych

W punkcie 8 algorytmu konstrukcji wzorców występuje założenie, że dla obu argumentów relacji określone są ich kategorie semantyczne zdefiniowane w wykorzystywanym słowniku semantycznym. Ponieważ słownik ten jest konstruowany automatycznie, jednym z ważniejszych problemów, które musiały zostać rozwiązane było automatyczne określenie kategorii semantycznych dla symboli zgromadzonych w słowniku.

Problem ten został rozwiązywany za pomocą algorytmu klasyfikacji symboli językowych. Był ona realizowana na podstawie kilku źródeł wiedzy, dzięki czemu możliwe było uzyskanie wysokiego pokrycia, przy zachowaniu wysokiej precyzji wyników. Punktem odniesienia klasyfikacji była ontologia Cyc, a elementami podlegającymi klasyfikacji były artykuły Wikipedii. W celu zaklasyfikowania artykułów, algorytm posiłkował się informacjami o typie *infoboksów* występujących w artykułach, pierwszymi zdaniami traktowanymi jak definicje opisywanych obiektów (porównaj [3, 4]), systemem kategorii Wikipedii (porównaj [16]) oraz bezpośrednim mapowaniem pomiędzy artykułami Wikipedii a ontologią Cyc (porównaj [15]). Klasyfikacje uzyskiwane tymi metodami były następnie uzgadniane z wykorzystaniem wewnętrznego mechanizmu wykrywania sprzeczności ontologii Cyc. Szczegółowy opis tego algorytmu przedstawiony jest w publikacji [13].

4.4 Automatyczne określanie ograniczeń semantycznych relacji

Punkt 9 algorytmu konstrukcji wzorców zakłada, że wyrażenia dopasowane do wzorca formalnego zostaną przeanalizowane pod kątem występowania w nich

[1] W ostatnim czasie część symboli w polskiej Słowosieci również została zaopatrzona w tradycyjne definicje.

zadanej relacji r oraz oznaczone jako zawierające, bądź niezawierające tę relację. W punkcie 9' zaproponowano alternatywny sposób określania ograniczeń semantycznych. W pierwszym rzędzie można pozyskać je z ontologii Cyc – kompletna ontologia powinna zawierać informacje o ograniczeniach semantycznych predykatów stosowanych do reprezentowania faktów. Zazwyczaj ontologie wykorzystują ograniczenia tego rodzaju, gdyż pozwalają one na stosunkowo łatwą kontrolę poprawności wprowadzanych danych, a także pozwalają na bardziej efektywne wnioskowanie na temat zgromadzonych faktów. Z drugiej jednak strony, ograniczenia te mogą być dość ogólne, ponieważ najczęściej nie są one projektowane pod kątem odróżniania poszczególnych relacji zdefiniowanych w danej ontologii.

Dlatego algorytm zakłada też inny sposób pozyskania tych ograniczeń – korzystając z bazy wiedzy zawierającej znaczną ilość faktów można podjąć próbę automatycznego określenia szczegółowych ograniczeń semantycznych, dzięki czemu będą one bardziej dokładne, co powinno skutkować wyższą precyzją ekstrakcji. Co więcej – ponieważ baza wiedzy zawiera zwykle przykłady wielu różnych relacji, możliwe jest połączenie wzorców formalnych z różnymi ograniczeniami semantycznymi, a tym samym opracowanie wzorców ekstrakcyjnych dla wielu relacji jednocześnie. W tej roli wykorzystywana jest semantyczna baza wiedzy DBpedia [1].

5 Wyniki

5.1 Dopasowanie wzorców formalnych

W celu weryfikacji tezy pracy głoszącej, że hybrydowy algorytm ekstrakcji relacji daje wyniki bardziej precyzyjne niż algorytm oparty wyłącznie o statystyczna analizę formalnych cech relacji, wyniki dopasowania wzorców formalnych poddane zostały analizie ilościowej oraz jakościowej. W pierwszej kolejności zbadano precyzję tak uzyskanych wyników oraz zidentyfikowano podstawowe źródła błędów ekstrakcji. Następnie przeprowadzono analizę jakościową (została ona pominięta w tym streszczeniu), której podstawowym celem było zidentyfikowanie problemów natury semantycznej, mogących mieć istotny wpływ na skuteczność algorytmu hybrydowego.

Ilościowa analiza dopasowań polegała na porównaniu wyników algorytmu dopasowującego wzorce formalne z odpowiedziami ludzi, którzy zostali poproszeni o ocenienie, czy wybrana relacja semantyczna – *całość-część* – występuje we fragmencie tekstu, który został dopasowany do wzorca formalnego. Gdyby okazało się, że wzorce formalne są wystarczająco skuteczne, zdecydowana większość dopasowań powinna zawierać wystąpienie relacji. Ponieważ jednak ocena algorytmu rozpoznawania relacji zależała od algorytmu ujednoznaczniania pojęć, istotnym elementem oceny przykładów było również stwierdzenie, czy pojęcia wykryte w tekście zostały poprawnie ujednoznacznione.

Zbiór przykładów poddany ewaluacji zawierał ponad 1000 elementów, a jego charakterystyka ilościowa przedstawiona jest w tabeli 1. Największą grupę przykładów stanowiły te, oznaczone jako zawierające niepoprawne dopasowanie argumentów relacji, tzn. przynajmniej jeden z argumentów nie został właściwie ujednoznaczniony. Tak duża liczba błędnych rozpoznań może budzić obawy o poprawność działania algorytmu rozstrzygania wieloznaczności. Wynika one jed-

Tablica 1: Charakterystyka zdań dopasowanych do wzorców formalnych relacji *całość-część*.

Typ zdania	**Liczba**	**Udział %**
zawierające relację	110	10,2
niezawierające relacji	416	38,6
niepoprawne dopasowanie	447	41,5
przykład problematyczny	105	9,7
w sumie	1078	100,0

nak przede wszystkim z tego, że nie ustalono minimalnego progu pewności dopasowania. Dzięki temu pewna liczba poprawnych rozpoznań, które zostałyby odrzucona, ze względu na niski współczynnik prawdopodobieństwa ich poprawności, została uwzględniona zwiększając liczbę przykładów, poddanych analizie.

Drugą istotną grupę stanowią przykłady, w których relacja *całość-część* nie występuje. Te wyniki pokazują wyraźnie, że opieranie się jedynie na wiedzy zawartej we wzorcach formalnych prowadzi wprost do wyników niskiej jakości, gdyż stosunek liczby przykładów zawierających relację, do liczby przykładów niezawierających relacji jest w przybliżeniu 1:4.

Ostatnią interesującą grupą jest zbiór zawierający przykłady problematyczne. Zostały one oznaczone w ten sposób, ponieważ były to przykłady, w których odpowiedzi osób określających występowanie relacji różniły się miedzy sobą lub przynajmniej jedna z osób oznaczyły określony przykład jako problematyczny. Widać wyraźnie, że liczba przykładów tego rodzaju była niemal taka sama jak liczba przykładów zawierających wystąpienie relacji. Ten wynik wskazuje, że jednoznaczne rozstrzygnięcie, czy w tekście występuje relacja *całość-część* stanowi wyzwanie nie tylko dla komputera, ale również dla ludzi.

5.2 Dopasowanie wzorców ekstrakcyjnych

W celu zweryfikowania tezy pracy, głoszącej, że możliwa jest automatyczna ekstrakcja relacji semantycznych, przeprowadzono szereg eksperymentów porównujących wyniki ekstrakcji relacji *całość-część* z wykorzystaniem ograniczeń semantycznych pozyskanych ręcznie, na podstawie Cyc oraz DBpedii.

Dla każdej metody przeprowadzono 4 warianty eksperymentu – biorąc pod uwagę kombinacje dwóch parametrów:

- użycia relacji *generalizacji* (**Gen.**) do wykrywania zgodności kategorii semantycznych wyrażeń z ograniczeniami semantycznymi zdefiniowanymi dla relacji,
  - – oznacza, że relacja ta nie była użyta, zatem wyrażenia musiały posiadać kategorie dokładnie pasujące do ograniczeń semantycznych,
  - \+ oznacza, że kategorie semantyczne wyrażeń mogły być również specjalizacjami ograniczeń semantycznych,
- wykluczenia wystąpienia relacji dla *identycznych kategorii* (**Id.**) obu argumentów,

+ oznacza, że jeśli dwa wyrażenia, dla których sprawdzano wystąpienie, posiadały przynajmniej jedną parę identycznych kategorii semantycznych, to przyjmowano, że relacja nie występuje,

– oznacza, że warunek ten nie był weryfikowany.

Tablica 2: Wyniki dopasowania wzorców relacji *całość-część* wyposażonych w ograniczenia semantyczne.

Źródło ograniczeń	**Gen.**	**Id.**	Pr **[%]**	Rc_{rel} **[%]**	F_1 **[%]**
Weryfikacja ręczna	–	–	89,2	56,9	69,5
	–	+	92,8	55,1	69,2
	+	–	77,7	64,8	70,7
	+	+	84,3	61,8	71,3
Cyc	–	–	82,4	9,4	17,0
	–	+	81,9	8,7	15,7
	+	–	76,6	41,2	53,5
	+	+	87,6	37,5	52,5
DBpedia	–	–	81,0	20,6	33,5
	–	+	84,4	16,4	27,5
	+	–	70,6	80,3	75,2
	+	+	76,4	74,0	75,2

Wyniki ekstrakcji relacji dla trzech wariantów określania ograniczeń semantycznych przedstawione są w tabeli 2, gdzie Pr – to precyzja wyników (ang. *precision*), Rc_{rel} – to względne pokrycie wyników (ang. *relative recall*) a F_1 – średnia harmoniczna precyzji i względnego pokrycia (ang. F_1 *score*).

Najważniejszym wnioskiem z wyników tych eksperymentów jest potwierdzenie tezy pracy, tzn. **możliwe jest skonstruowanie algorytmów automatycznej ekstrakcji relacji z tekstów w języku polskim, których wyniki byłyby lepsze od algorytmów opierających się na ręcznej weryfikacji zdań, przyjmując, że liczba zdań ewaluowanych w celu określenia ograniczeń semantycznych wynosi 10% całkowitej liczby zdań poddawanych analizie**. Najlepszy wynik (w sensie miary F_1) uzyskany został dla ograniczeń pozyskanych z DBpedii, w wariancie wykorzystującym relację generalizacji oraz wykluczającym identyczne kategorie semantyczne argumentów.

Teza rozprawy składa się z trzech części. Część pierwsza dotyczy precyzji wyników uzyskiwanych przez hybrydowy algorytm ekstrakcji informacji. Na rzecz tej tezy świadczy różnica w precyzji wyników ekstrakcji uzyskanych na podstawie wzorców formalnych oraz wyników uzyskanych na podstawie wzorców wyposażonych w ograniczenia semantyczne zdefiniowane z wykorzystaniem pojęć ontologii Cyc. Jak pokazane zostało w punkcie 5.1, wzorce formalne zbudowane na podstawie analizy statystycznej, dają wyniki ekstrakcji o precyzji w przedziale 20%-40%. Natomiast algorytm hybrydowy uzyskuje wyniki charakteryzujące się wyższą precyzją – najgorszy wariant bazujący w całości na ontologii Cyc ma precyzję wynoszącą 41%, natomiast wariant najlepszy, oparty o ograniczenia semantyczne wyekstrahowane z ręcznie ocenionych zdań, posiada precyzję wynoszącą 92%. Te wyniki pokazują, że **zastosowanie algorytmu hybrydowego istotnie przyczynia się do poprawy precyzji uzyskiwanych**

wyników, co dowodzi słuszności pierwszej części tezy.

Druga część tezy dotyczy obszaru zastosowania algorytmu i zakłada, że nie ma być on ograniczony do pojedynczej dziedziny wiedzy. Ta część tezy potwierdzona została na kilka sposobów. Po pierwsze, na żadnym etapie konstrukcji wzorców ekstrakcyjnych nie była wykorzystywana wiedza dziedzinowa. Co prawda, jako przykładów użyto par pojęć z dziedziny anatomii, ale uzyskane wzorce okazały się skuteczne w ekstrakcji informacji w innych dziedzinach.

Po drugie, wynikowe wzorce ekstrakcyjne wykorzystywane były do analizy notatek PAP. Zakres tematów poruszanych w notatkach nie jest ograniczony do jednej dziedziny wiedzy, choć dominują informacje związane z polityką międzynarodową. W efekcie, hybrydowy algorytm ekstrakcji relacji semantycznych rozpoznał wystąpienia relacji *całość-część* w następujących zdaniach[2]:

- Prof. Edward Borowski, szef gdańskiego oddziału tej organizacji, szacuje, że ok. 30 procent mieszkańców trójmiejskiej aglomeracji stanowią osoby, które pochodzą ze *stolicy* ***Litwy*** i okolic,
- Zdaniem posłów koalicji „obowiązkiem Krajowej Rady Radiofonii i Telewizji oraz *Rady Nadzorczej* ***TVP*** jest przerwanie tych destruktywnych działań",
- Koszykarze Portland Trail Blazers i Los Angeles Lakers zagrają w finale *Konferencji Zachodniej* ***ligi NBA***,
- *Posłanka* ***Unii*** dodała, że sama jest za jeszcze dalej idącym rozwiązaniem, które zakłada, że rady nadzorcze nie miałyby wpływu na skład zarządów mediów publicznych.

W każdym z nich mamy do czynienia z inną dziedziną wiedzy. W pierwszym wiedza dotyczy zależności geopolitycznych – *stolica* jest częścią *państwa*, w drugim wiedza dotyczy organizacji spółek handlowych – częścią *spółki* jest jej *rada nadzorcza*, w trzecim zdaniu rozpoznane są zależności w obszarze sportu – *liga NBA* podzielona jest na dwie *konferencje*, natomiast w ostatnim zdaniu rozpoznane zostały zależności w organizacji politycznej – *posłanka* jest częścią *partii politycznej*. Widać zatem, że algorytm zdolny jest do ekstrakcja relacji w wielu dziedzinach wiedzy.

Po trzecie zaś – wszystkie źródła wiedzy wykorzystywane w algorytmie mają charakter uniwersalny. Dotyczy to słownika fleksyjnego, Wikipedii, ontologii Cyc oraz semantycznej bazy wiedzy jaką jest DBpedia. Wszystkie te fakty świadczą na rzecz tezy, że **hybrydowy algorytm ekstrakcji relacji semantycznych jest uniwersalny**, co dowodzi słuszności drugiej części tezy.

Ostatnia część tezy rozprawy dotyczy nakładu pracy ręcznej, jaka musi zostać wykonana, aby można było zastosować prezentowany algorytm do ekstrakcji nowych relacji semantycznych. W założeniu spełnienie warunku o mniejszym nakładzie pracy ręcznej, niezbędnej do wykorzystania algorytmu, ma umożliwić jego praktyczne zastosowanie. Ta część tezy została potwierdzona poprzez wyższą wartość miary F_1 uzyskaną przez wariant algorytmu oparty o DBpedię, w stosunku do wariantu opartego o ręczną ocenę zdań. Czas który trzeba poświęcić na zidentyfikowanie w DBpedii predykatów reprezentujących interesującą nas relację oraz określenie kolejności argumentów w tych predykatach jest

[2]Przykłady te pochodzą z korpusu PAP.

znacznie krótszy, niż czas potrzebny na ręczną weryfikację zbioru zdań, który zostałby użyty w algorytmie o porównywalnej skuteczności. W konsekwencji **algorytm hybrydowy w wariancie wykorzystującym ograniczenia semantyczne określone na podstawie DBpedii, wymaga mniejszego nakładu pracy ręcznej, niż analogiczny algorytm oparty o zbiór danych oznaczonych ręcznie, oferując wyższą jakość uzyskiwanych wyników**, co dowodzi słuszności trzeciej części tezy.

Ponieważ wszystkie części tezy rozprawy zostały udowodnione należy uznać, że teza głosząca, że **możliwe jest skonstruowanie hybrydowego algorytmu ekstrakcji relacji semantycznych** została obroniona.

Literatura

[1] Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z.: „Dbpedia: A nucleus for a web of open data". Aberer K., Choi K.-S., Noy N., Allemang D., Lee K.-I., Nixon L., Golbeck J., Mika P., Maynard D., Mizoguchi R., Schreiber G., Cudré-Mauroux P., redaktorzy, *The Semantic Web*, wolumen 4825, strony 722–735. Springer, Berlin, Heidelberg, 2007.

[2] Banerjee S., Pedersen T.: „An adapted lesk algorithm for word sense disambiguation using wordnet". Gelbukh A., redaktor, *Computational Linguistics and Intelligent Text Processing*, strony 136–145. Springer, Berlin, Heidelberg, 2002.

[3] Chrząszcz P.: „Automatyczne rozpoznawanie i klasyfikacja nazw wielosegmentowych na podstawie analizy haseł encyklopedycznych", 2009.

[4] Gangemi A., Nuzzolese A. G., Presutti V., Draicchio F., Musetti A., Ciancarini P.: „Automatic typing of DBpedia entities". Cudré-Mauroux P., Heflin J., Sirin E., Tudorache T., Euzenat J., Hauswirth M., Parreira J., Hendler J., Schreiber G., Bernstein A., Blomqvist E., redaktorzy, *The Semantic Web–ISWC 2012*, strony 65–81. Springer, Berlin, Heidelberg, 2012.

[5] Lenat D. B.: „CYC: A large-scale investment in knowledge infrastructure". *Communications of the ACM*, 38(11):33–38, 1995.

[6] Lenat D. B., Guha R. V.: *Building Large Knowledge-Based Systems*. Addison Wesley, Boston, 1990.

[7] Lesk M.: „Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone". DeBuys V., redaktor, *Proceedings of the 5th annual international conference on Systems documentation*. ACM, 1986, strony 24–26.

[8] Li Z., Li H., Wang H., Yang Y., Zhang X., Zhou X.: „Overcoming semantic drift in information extraction.". Christophides V., redaktor, *Processing of the 17th International Conference on Extending Database Technolog*, 2014, strony 169–180.

[9] Milne D.: „An open-source toolkit for mining Wikipedia". Blagojevic R., redaktor, *Proceedings of 7th New Zealand Computer Science Research Student Conference*, wolumen 9, 2009, strony 222–239.

[10] Milne D., Witten I.: „Learning to link with Wikipedia". Shanahan J. G., Amer-Yahia S., Manolescu I., Zhang Y., Evans D. A., Kolcz A., Choi K.-S., Chowdury A., redaktorzy, *Proceeding of the 17th ACM conference on Information and knowledge management*. ACM, 2008, strony 509–518.

[11] Piasecki M., Szpakowicz S., Broda B.: *A WordNet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 2009.

[12] Pohl A.: „An Ontology-based Method for an Efficient Acquisition of Relation Extraction Training and Testing Examples". Bouvry P., Klopotek M., Leprevost F., Marciniak M., Mykowiecka A., Rybiński H., redaktorzy, *Security and Intelligent Information Systems*, 2012, strony 318–331.

[13] Pohl A.: „Classifying the Wikipedia Articles into the OpenCyc Taxonomy". Rizzo G., Mendes P., Charton E., Hellmann S., Kalyanpur A., redaktorzy, *Proceedings of the Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference*, 2012, strony 5–16.

[14] Pohl A.: „Improving the Wikipedia Miner Word Sense Disambiguation Algorithm". Ganzha M., Paprzycki M., redaktorzy, *Proceedings of Federated Conference on Computer Science and Information Systems 2012*. IEEE, 2012, strony 241–248.

[15] Sarjant S., Legg C., Robinson M., Medelyan O.: „All you can eat ontology-building: Feeding wikipedia to cyc". Yates R. B., Berendt B., Bertino E., Peng L. E., redaktorzy, *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2009, strony 341–348.

[16] Suchanek F., Kasneci G., Weikum G.: „YAGO: a core of semantic knowledge". Williamson C., Zurko M. E., Patel-Schneider P., Shenoy P., redaktorzy, *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, strony 697–706.

[17] Waszczuk J.: „Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language". Kay M., Boitet C., redaktorzy, *Proceedings of COLING*, 2012, strony 2789–2804.