Kraków, dnia 25 maja 2015 roku

dr hab. inż. Krzysztof Boryczko
Akademia Górniczo-Hutnicza
Wydział Informatyki, Elektroniki i Telekomunikacji
Katedra Informatyki
al. A. Mickiewicza 30
30-059 Kraków

Recenzja
Rozprawy Doktorskiej
mgr inż. Aleksandra Smywińskiego-Pohl
pt. „Automatyczna ekstrakcja relacji semantycznych z tekstów w języku polskim"
Promotor: Prof. dr hab. Wiesław Lubaszewski
Akademia Górniczo-Hutnicza w Krakowie
Wydział Informatyki, Elektroniki i Telekomunikacji
Katedra Informatyki

Niniejsza recenzja została opracowana na zlecenie Prodziekan Wydziału Informatyki, Elektroniki i Telekomunikacji dr hab. inż. Katarzyny Zakrzewskiej, prof. n. AGH (pismo WIET/115/2015 z dnia 23 lutego 2015 roku), działającej w oparciu o Uchwałę Rady Wydziału z dnia 19 lutego 2015 roku.

1. Ocena wyboru tematu i tez rozprawy

W pracy postawiono i starano się wykazać prawdziwość następującej tezy:

„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 wytrenowany na ręcznie oznakowanym zbiorze uczącym."*

Autor Rozprawy doprecyzował poszczególne elementy tezy. Pod pojęciem *algorytmu hybrydowego* należy rozumieć algorytm bazujący na dwóch paradygmatach przetwarzania języka naturalnego. Chodzi tu o paradygmat statystyczny, wykorzystujący duże zbiory danych oraz algorytmy statystyczne opierające się na prawdopodobieństwie warunkowym. Drugi to paradygmat symboliczny wykorzystujący bazy wiedzy opisujące zjawiska językowe w sposób symboliczny.

Pierwszy punkt tezy mówi, że wyniki otrzymane przez algorytm hybrydowy powinny być lepsze niż otrzymane przez algorytm statystyczny. Pojawia się tu jednak pewna nieścisłość. Autor mówi o wynikach „*bardziej precyzyjnych*", co sugeruje wykorzystanie jednej miary

jakości. Faktycznie Autor używa trzech miar. Należy zatem sądzić, iż sformułowanie „*wyniki bardziej precyzyjne*" ma znaczenie potoczne, a nie odnosi się do jednej, konkretnej miary jakości wyników.
Drugi punkt tezy jest oczywisty i zakłada możliwość działania algorytmu dla tekstów z różnych, wąskich dziedzin wiedzy. To olbrzymia zaleta algorytmu.
Ostatni postulat wymaga minimalizacji nakładu pracy ręcznej dla osiągnięcia celu. Dowiedzenie tego punktu tezy wymaga porównania wyników uzyskanych dzięki algorytmowi opierającemu się na danych pozyskanych ręcznie oraz wyników dostarczonych przez algorytm który pracował na danych otrzymanych w sposób automatyczny.

Dla wykazania prawdziwości tezy pracy zrealizowano kilka celów szczegółowych, których opis stanowi treść kolejnych rozdziałów dysertacji.

Tematyka recenzowanej Rozprawy dobrze lokuje się w obszarze zainteresowań współczesnej informatyki, w zakresie ekstrakcji informacji z tekstów, w tym przypadku pisanych w języku polskim. Skupia się ona głównie na rozpoznawaniu wybranych (jak to zostało określone w tezie) relacji semantycznych w oparciu o automatycznie konstruowane wzorce semantyczne. **Wybór problemu, tezy rozprawy oraz jej celów oceniam pozytywnie.**

2. Ocena zawartości rozprawy

Przedstawiona do recenzji praca liczy 214 stron maszynopisu i składa się z jedenastu numerowanych rozdziałów, spisu pozycji literaturowych zawierającego 164 pozycje uporządkowane alfabetycznie wg nazwiska pierwszego autora, spisu rysunków, spisu tablic oraz dodatków (A. Kompletna lista par symboli połączonych predykatem **#$anatomicalParts**; B. Lista par symboli połączonych predykatem **#$anatomicalParts** przetłumaczonych na język polski; C. Część taksonomii ontologii Cyc zakorzenionej pojęciu **#$Birds** i przetłumaczonej na język polski; D. Wzorce relacji *całość-część* o CDp>=2; E. Lista predykatów DBpedii odpowiadających relacji *całość-część*).

Układ i zawartość rozdziałów merytorycznych są zasadniczo poprawne. Dbałość edytorska ułatwia śledzenie wywodów Autora i dobrze świadczy o jego trosce, nie tylko o poziom naukowy, ale także o komunikatywność przekazu.

Rozdział numer 2 recenzowanej dysertacji to wprowadzenie do zagadnienia ekstrakcji informacji. Autor wyjaśnia w nim pojęcie ekstrakcji informacji przytaczając definicje i lematy kilku badaczy. Pierwszy podrozdział zamyka definicja wykorzystywana w dalszych rozdziałach. Następnie dokumentuje konieczność prowadzenia prac w tym zakresie podając konkretne przykłady. Ostatni podrozdział zawiera określenie zadań, które pojawiają się w ramach ekstrakcji informacji.

W rozdziale 3 „Reprezentacja wiedzy – relacje i sieci semantyczne" wprowadzono definicje podstawowych pojęć z dziedziny. Zostało określone pojęcie symbolu językowego i zasad jego funkcjonowania, jak również relacji semantycznych i sieci semantycznych jako elementów połączonych sieciami z precyzyjnym opisaniem ich rozdziałów.

Rozdział 4 „Historia i stan badań nad ekstrakcją informacji" został podzielony na dwa podrozdziały dotyczące ekstrakcji w języku angielskim i w języku polskim. W obu przypadkach, po przedstawieniu rysu historycznego prac naukowych podrozdziały kończą się paragrafami opisującymi stosowane, specyficzne dla języka metody ekstrakcji relacji semantycznych. W mojej ocenie rozdział ten zawiera nieco niepotrzebnych z punktu widzenia tematyki recenzowanej Rozprawy informacji. Na ich miejscu mogłyby pojawić się „aspekty inżynierskie", na co wskazuje w dalszej części recenzji.

W rozdziale 5 „Szkic algorytmu ekstrakcji relacji semantycznych" Autor precyzyjnie definiuje cel algorytmu. Następnie nakreśla jego strukturę oraz określa wymagania dla

algorytmów pomocniczych. Ponieważ proponowany algorytm wymaga dostępu do źródeł wiedzy, zostają one określone i krótko scharakteryzowane. To jeden z najistotniejszych rozdziałów recenzowanej Rozprawy.

Rozdział 6 „Zasoby wykorzystywane przez algorytm" zawiera charakterystykę źródeł i baz danych niezbędnych do działania algorytmu (w założeniu hybrydowego). Opisane zostały korpusy tekstów, słowniki fleksyjne oraz słownik semantyczny. Osobny podrozdział został poświęcony ontologii w tym organizacji pojęć i predykatom. Rozdział zamykają opisy semantycznej bazy wiedzy oraz symbolu językowego.

Z kolei rozdział 7 „Algorytmy pomocnicze" zawiera specyfikację wymagań oraz określenie jakości wyników dostarczanych przez dodatkowe algorytmy konieczne do działania autorskiego algorytmu. Stąd szeroki opis sposobu wyboru zdań zawierających relacje semantyczne i jego aspektów, opis metod semantycznej klasyfikacji symboli językowych, sposobów ujednoznaczniania sensu wyrażeń w tekście oraz algorytmu określania ograniczeń semantycznych. Za istotne uważam określenie miar i cech na których działają algorytmy.

Rozdział 8 „Algorytm tworzenia wzorców ekstrakcyjnych" zawiera precyzyjny opis istotnych fragmentów działania algorytmu autorskiego. Stąd opis sposobu wyboru przedmiotowej relacji, opis sposobu określania symboli połączonych relacją i wyszukiwania par symboli w korpusie. W dalszej kolejności ekstrakcja wzorców formalnych i działania na nich. Rozdział ten jest bardzo istotny w kontekście implementacji algorytmu.

Rozdział numer 9 zatytułowany skupia się na zagadnieniach związanych z konstruowaniem wzorców relacji *całość-część*. To podstawowa relacja dla której dokonano weryfikacji działania proponowanego algorytmu.

Wyniki ekstrakcji relacji całość-część przedstawiono i opisano w rozdziale 10. Istotne w rozdziale jest określenie metod oraz miar oceny wyników. Autor opisuje uzyskane wyniki dla relacji całość-część, analizuje źródła błędów oraz ogólnie próbuje określać wyniki działania algorytmu dla innych relacji semantycznych.

Rozprawę zamyka rozdział 11 „Podsumowanie". Autor odwołuje się w nim do tezy sformułowanej we Wstępie. Uzasadnia, iż przedstawione w pracy, a będące jego indywidualnymi osiągnięciami, wyniki dowodzą ją. Oprócz podsumowania indywidualnych osiągnięć rozdział zawiera również wskazanie dalszych kierunków badań, co uważam za istotne z merytorycznego punktu widzenia.

3. Oryginalne osiągnięcia Autora rozprawy

W recenzowanej dysertacji zawarto kilka wartościowych i oryginalnych koncepcji oraz rozwiązań (częściowo przedstawionych powyżej), dokonano ich implementacji oraz uzyskano wyniki wzbogacające naszą wiedzę w tym zakresie. Do najważniejszych osiągnięć Autora rozprawy zaliczyć należy:

1. Opracowanie, analizę, weryfikację oraz implementację przedstawionego w rozdziale 7 algorytmu wyboru zdań zawierających wystąpienie wybranej relacji semantycznej.
2. Opracowanie systemu do automatycznej ekstrakcji relacji semantycznych wykorzystującego ontologie Cyc oraz BDpedię.
3. Określenie oraz weryfikację statystycznych cech wzorców formalnych decydujących o ich poprawności.
4. Praktyczne porównanie metod określania ograniczeń semantycznych.
5. Wszechstronną analizę błędów ekstrakcji.

Wymienione osiągnięcia są oryginalne i znaczące. Na nich zatem przede wszystkim opieram ogólnie pozytywną ocenę Rozprawy. Udowadniają one postawioną tezę rozprawy oraz stanowią istotny wkład w rozwój algorytmów ekstrakcji informacji z tekstów w języku polskim, a w szczególności w rozpoznawanie relacji semantycznych poprzez automatycznie konstruowane wzorce ekstrakcyjne.

Recenzowana rozprawa zawiera także inne wartościowe i oryginalne wyniki naukowe, jednak te przedstawione powyżej uważam za szczególnie ważne i godne podkreślenia w recenzji.

4. Uwagi dyskusyjne i krytyczne

Jak już wspomniałem, przedłożona do recenzji rozprawa pod względem merytorycznym i redakcyjnym napisana jest poprawnie. W mojej ocenie Autor nie ustrzegł się jednak pewnych braków i nieścisłości. Należy do nich zaliczyć:

1. W rozdziale „Podsumowanie", na stronie 180 Autor stwierdza, iż: „... prezentowany system został skonstruowany niemal w całości od podstaw przez autora. Jedynie algorytm ujednoznaczniania morfosyntaktycznego realizowany był przez zewnętrzny program *Concraft* [158]. Wszystkie pozostałe moduły zostały zaimplementowane przez autora". W tym kontekście za istotny brak uważam niezamieszczenie w Rozprawie analizy autorskiego algorytmu pod względem złożoności obliczeniowej oraz pamięciowej. Brakuje również informacji o czasach realizacji algorytmu dla różnych danych wejściowych. Wraz z informacją o złożoności pozwoliłoby to na szacowanie czasu wykonania dla innych zbiorów danych. W pracy o charakterze inżynierskim powinny się również pojawić, w mojej ocenie, pewne wskazówki dotyczące implementacji. Mam tu na myśli określenie odpowiednich dla proponowanego algorytmu architektur procesorowych, komputerowych oraz wynikających z ich zastosowania języków i metod implementacji.
   Skrócenie treści rozdziału numer 4 pozwoliłoby na umieszczenie nowego rozdziału bez konieczności zwiększania liczby stron. Dodatkowo urozmaiciłby on Dysertację.
2. W podrozdziale 8.8 „Określenie ograniczeń semantycznych" (strona 134) Autor stwierdza, iż: „Przyjmujemy jednak, że w prezentowanym algorytmie przypadki tego rodzaju są nierozwiązywalne, tzn. nie będziemy używać dodatkowych metod (np. wnioskowania wykraczającego poza relację hiperonimii), gdyż doprowadziłoby to do istotnej komplikacji algorytmu". W świetle poprzedniej uwagi sformułowanie „istotna komplikacja algorytmu" uważam za nieprecyzyjne. Czy chodzi w konsekwencji o wzrost złożoności obliczeniowej, czy nakładów programistycznych?
3. Pewne wątpliwości budzi w mojej ocenie metodyka weryfikacji jakości wyników. Otóż zakłada się, że weryfikacji ręcznej zostanie poddanych 10% wyników dopasowania. W pracy brak jest dostatecznego uzasadnienia tej wartości poza stwierdzeniem, iż nie jest to wysokim wymaganiem, gdyż sprowadza się do weryfikacji kilkuset zdań. Rodzi się jednak pytanie, czy liczba ta wpływa na jakość wyników? Może okazać się, że analiza jedynie 5% umożliwi uzyskanie zadowalających wyników, a nakład pracy ręcznej będzie dwa razy mniejszy lub podniesienie liczby analizowanych ręcznie zdań do 15% opłaci się, gdyż nakład pracy wzrośnie o połowę, ale da to znacznie lepsze wyniki. Stąd wydaje mi się, iż porównanie wyników dla różnej liczby zdań byłoby interesujące.
4. Weryfikacja działania autorskiego algorytmu została przeprowadzona przede wszystkim dla relacji całość-część oraz, w znacznie mniejszym stopniu, relacji posesywnej i lokalizacji (stąd słowo „wybranych" w tezie Rozprawy). Czy istnieją jednak przesłanki pozwalające potwierdzać równie wysoką, względną efektywność

algorytmu dla innych typów relacji? Chodzi o bardziej precyzyjne dowiedzenie niż w podrozdziale 10.3. Można również zapytać, dla jakich relacji semantycznych zaproponowany algorytm nie da zadowalających wyników?

Praca zawiera również kilka błędów interpunkcyjnych czy niezręczności językowych. Przykładowo:

- strona 49, linia 15 od dołu: „... ,lecz proces ten był zbyto powolny, ...",
- strona 99, linia 21 od dołu: „Metoda oparto o kategorie ...",
- strona 99, linia 20/19 od dołu: „...., powoduje, że uzyskane kategorie semantyczne są bardzo specyficzny, ..."
- strona 126, linia 2 od dołu: „Krok ten może zostać przeprowadzony dopiero do zdubowaniu odpowiednich zbiorów, ...",
- strona 131, linia 3/2 od dołu: „Pozwala to na wybranie najbardziego opisu morfo syntaktycznego ...",
- strona 136, linia 6 od dołu: „Pierwszy kwestia, która ...",
- strona 172, linia 3/2 os dołu: „... , a drugim argumentem występowała relacj typ-okaz, ...",
- strona 179, linia 10 od dołu: „Oba te wynik są bardzo ważne, ...".

Biorąc pod uwagę rozmiar pracy należy stwierdzić, iż są one nieliczne.

W tym zakresie moje uwagi budzi również brak jednolitej struktury rozdziałów. Przykładowo, rozdział numer 3 po tytule zawiera krótkie wprowadzenie, po którym pojawia pierwszy podrozdział. Rozdział kończy podrozdział „Podsumowanie". Z kolei rozdział numer cztery rozpoczyna tytuł, po nim pojawia się bezpośrednio podrozdział 4.1, a po nim punkt 4.1.1. Rozdział posiada podsumowania dla prac związanych z zagadnieniami ekstrakcji cech w języku angielskim i polskim osobno. Rozdział numer 5 posiada wprowadzenie, ale nie ma podsumowania. Rozdział numer 6 rozpoczyna się jak rozdział numer 4, ale brak w nim podsumowania. Rozdział numer 9 rozpoczyna się nienumerowanym paragrafem „Wstęp", ale brak w nim podsumowania. Uważam, iż struktura ta powinna być jednolita, zgodna ze strukturą rozdziału numer 3. Ułatwia to lekturę.

Należy zaznaczyć, że podane uwagi nie wpływają w sposób istotny na poznawcze oraz utylitarne wartości zaproponowanego rozwiązania i metodologii testowania. Ich uwzględnienie może okazać się korzystne w dalszej działalności, dotyczącej przedmiotowego zagadnienia, jak również przy okazji ewentualnego publikowania materiału zawartego w Rozprawie.

5. Podsumowanie

Przytoczone powyżej uwagi polemiczne nie umniejszają wartości merytorycznej pracy, która stanowi oryginalny wkład Autora w zagadnienia związane z konstruowaniem algorytmów dla automatycznej, zaawansowanej ekstrakcji relacji semantycznych z tekstów w języku polskim.

Podsumowując recenzję stwierdzam, iż moja generalna opinia o Rozprawie doktorskiej: „Automatyczna ekstrakcja relacji semantycznych z tekstów w języku polskim" mgr inż. Aleksandra Smywińskiego-Pohl jest pozytywna. Uważam, że przedstawiona do recenzji Rozprawa zawiera samodzielne, zaproponowane przez Autora rozwiązanie trudnego i ważnego problemu. Postawiona, trzypunktowa teza rozprawy została dowiedziona, a podstawowe cele i zadania pracy zrealizowane. W pełni odpowiada to wymaganiom stawianym rozprawom doktorskim przez odnośną ustawę o Tytule Naukowym i Stopniach Naukowych. **Na tej podstawie wnioskuję o dopuszczenie Rozprawy do publicznej obrony w celu uzyskania przez jej Autora stopnia doktora nauk technicznych z dyscypliny Informatyka.**