Uniwersytet im. Adama Mickiewicza
w Poznaniu
Zakład Lingwistyki Informatycznej i Sztucznej Inteligencji
PL-61-614 Poznań, ul. Umultowska 87, tel. +48-61-8295380, fax +48-61-8295315
Kierownik Zakładu: prof. zw. dr hab. Zygmunt Vetulani
http://amu.edu.pl/~zlisi
vetulani@amu.edu.pl

Recenzja rozprawy doktorskiej mgra Piotra Przybyła
pt. "Odpowiadanie na pytania w języku polskim z użyciem głębokiego rozpoznanwania nazw"

Mgr Piotr Przybyła przedłożył do oceny rozprawę, której część merytoryczna obejmuje 133 strony standardowego tekstu (czcionka Times 11, interlinia pojedyncza). Na tekst składa się sześć rozdziałów, bibliobrafia oraz dodatki. Rozprawa sytuuje się w dziedzinie informatyki i jest przyczynkiem do dziedziny przetwarzania informacji przez systemy komputerowe. Problem, który Autor podejmuje, jest zagadnieniem z zakresu Question Answerning (QA)[1], gdzie celem jest automatyczne generowanie odpowiedzi na pytania zadawane w języku polskim. Nad tak postawionym zagadnieniem prowadzone są intensywne prace w wielu ośrodkach dla różnych języków i to od bardzo dawna (wczesne lata 60-te), co zresztą odnotował Autor w rozprawie. Prace te poszły w dwóch kierunkach: w kierunku głębokiej analizy pytań sprowadzającej się do ich rozumienia (kierunek wymagający stosowania technik sztucznej inteligencji), oraz w kierunku powierzchniowego przetwarzania tekstu pytania[2]. Ten drugi kierunek okazał się bardzo atrakcyjny i inspirujący dla badaczy od czasu, gdy okazało się, że za źródło wiedzy uznać można zasoby tekstów np. dostępnych w sieci w postaci zdigitalizowanej, a więc wygodnej do przetwarzania. Wobec sukcesu internetu i dostępności do olbrzymich zasobów wiedzy zgromadzonej w tekstach kierunek ten nabrał praktycznego znaczenia, podczas gdy ten pierwszy długo miał charakter akademicki (ta sytuacja się obecnie zmienia). Dla języka polskiego, poważne wyniki zostały osiągnięte w ramach tego pierwszego podejścia /także przez recenzenta/ – o czym Autor informuje, podczas gdy – kuriozalnie – powstała istotna luka w zakresie technik opartych na stosunkowo płytkim przetwarzaniu tekstu. Autor rozpartuje problem QA w ograniczeniu do przypadku, gdy przez pytanie rozumie się zdanie inicjujące proces wyznaczania prostej odpowiedzi znajdującej się *explicite* w bazie tekstów. Zawężenia problemu idą daleko i wymagają, żeby zadanie było rozwiązywalne bez odwoływania się do kontekstu, a tym bardziej bez konieczności rozumienia na poziomie pragmatycznym (czy wręcz semantycznym). Przy powyższych wyjaśnieniach można zgodzić się z Autorem, że opracowany przez niego system Rafael jest rozwiązaniem pionierskim dla języka polskiego i wypełnia lukę w zakresie poszukiwania odpowiedzi na pytania w tekstach (w tekstowych bazach wiedzy) bez stosowania wyrafinowanych technik sztucznej inteligencji. Zaprojektowany system Rafael należy więc uznać za ważny krok w kierunku nadrobienia „zaległości" w stosunku do istniejącego stanu badań dla języka angielskiego (i niektórych innych), tym bardziej, że Autor

[1] W terminologii polskiej nie wykształcił się odpowiednit terminu Question Answering, który byłby zgodnie zaakceptowany przez środowisko, zatem pozoztaję przy powszechnie używanym w tekstach polskich terminie angielskim.

[2] „Pytanie zdefiniujemy jako pojedyncze, krótkie zdanie, będące poprawnym zdaniem w języku polskim, na które odpowiedź stanowi prosta informacja zawarta w bazie tekstów"

proponuje rozwiązanie nowatorskie (przez zastosowanie technologii wordnetowej), idące w kierunku *pogłębienia* analizy pytania. Ciągle jednak nie możemy mówić o analizie głębokiej, co mógłby sugerować tytuł rozprawy. Reasumując, doktorant podejmuje problem ważny, a jego rozwiązanie wypełnia istotny brak w dotychczasowych badaniach związanych z przetwarzaniem informacji wyrażonych w tekstach redagowanych w języku polskim. Można tym samym uznać, że praca spełnia wymagania oczekiwane od prac doktorskich innowacyjności odnośnie rozpraw doktorskich.

Uwagi szczegółowe.

Jako główne osiągnięcie rozprawy, jak należy wnioskować z treści Rozdziału 1 (Wprowadzenie), Doktorant deklaruje opracowanie systemu RAFAEL (Rapid Factoidal Answer Extracting aLgorithm), którego podstawową funkcjonalnością jest "otwarto-dziedzinowe odpowiadanie na pytania o proste fakty na podstawie bazy tekstowej". Sformułowanie problemu w ten sposób jest bardzo obiecujące i generuje oczekiwania, które nie w pełni są spełnione. Sformułowanie powyższe bowiem nawiązuje do sposobu stawiania problemu QA w latach 80-tych (i wcześniej), lecz już w tym okresie było jasne, że generowanie odpowiedzi na pytania wymaga przede wszystkim jego zrozumienia, a generowanie odpowiedzi wymaga z kolei przetwarzania wiedzy i wykorzystania logiki, a także analizy pragmatycznej. W tym kierunku poszła cześć badań w obszarze QA, w tym także, wspomniane w rozprawie prace dla języka polskiego (w tym także recenzenta). Niewątpliwie Autor świadomie ograniczył się do stosunkowo powierzchniowego przetwarzania tekstu pytań i źródeł informmacji (baza tekstów), wbrew temu co może sugerować tytuł pracy przez użycie określenia „głębokie rozpoznawanie nazw", lecz nie uzasadnia tego ograniczenia. Sposób wyjaśnienia pewnych innych samoograniczeń budzić może pewne zdziwienie. Np. w rozdziale 2.2.3 Autor skrótowo omawia cztery typy udzielania odpowiedzi, wybierając typ trzeci, skrócony. Jednak wytłumaczenie, że NLG jest „odrębnym, nierozwiązanym jeszcze zadaniem" mija się z rzeczywistością, przymnajmniej w odniesieniu do przykładu ze strony 15, gdzie przekształcenie pytania do odpowiedzi typu 4 jest trywialne, pod warunkiem przeprowadzenia analizy semantyczno-syntaktycznej (porównaj rozdział 2.3.2 pt. „Dopasowanie strukturalne").

Niezależnie, od tego, że system RAFAEL, jako ew. system użytkowy ma ograniczenia, o których zresztą Autor pisze w rozprawie, to podstawowa technika, która jest w nim zastosowana, i która kryje się pod nazwą (nieco mylącą!) „głębokiego wykrywania nazw bytów" jest interesującym samo przez się przykładem wykorzystania technologii wordnetowych dla powiązania bytów określonych deskryptywnie z nazwą. W tym celu dla elementów znaczących części deskryptywnej (np. pytania) wyszukuje się określające je synsety, i korzystając z relacji, w które wchodzą, uzyskuje się znaczne ograniczenie przestrzeni poszukiwań kandydata na nazwę. Powiązania, o których mowa powyżej, mają typowo miejsce przy obsługiwaniu pytań *o nazwę (Jak nazywa się ...?)*, jest to więc technika ważna w kontekście QA. **Opisanie tej techniki uważam za główny, najbardziej innowacyjny, wynik pracy**. Opis tej techniki jest wyodrębniony w postaci Rozdziału 4 („Głębokie rozpoznawanie nazw") z bardzo obszernej i szczegółowej deskrypcji systemu RAFAEL (której Autor poświęcił aż 49 stron). Opis przyjmuje postać algorytmu, który Autor nazywa DeepER (Deep Entity Recognition). Istotą tego algorytmu jest wykrywanie nazw dla bytów (abstrakcyjnych lub konkretnych) scharakteryzowanych deskrypcją tekstową, np. bytów, o które pytamy w systemie QA. Wymaga to porównania treści deskrypcji z wiedzą o bytach reprezentowanych przez nazwy, tj. znajomości nazw własnych, jak np. Napoleon Bonaparte, oraz nazw pospolitych, jak np. cesarz. Dopiero wtedy będziemy w stanie

prawidłowo odpowdzieć na pytanie *Kim był wódz, który zwyciężył pod Austerlitz?* na podstawie wiedzy encyklopedycznej. Wiedzę tę można pozyskać np. z:
- opisów (słowniki, leksykony),
- definicji (encyklopedie, teksty naukowe bądź techniczne, bazy terminologiczne).
Z tych prostych obserwacji wychodzi też Doktorant zakładając stworzenie bazy nazw/pojęć zwanej „biblioteką bytów", w której indywidua nazwane oraz oraz pojęcia scharakteryzowane są przez odwołanie się do ontologii leksykalnej typu wordnet. (W zasadzie można by wykorzystać dowolną ontologię służącą jednocześnie do wyrażenia tejże wiedzy o bytach, jak też do otagowania tekstów.) Z dwóch istniejących dla języka polskiego ontologii typu WordNet, tj. PolNet oraz plWordNet, doktorant wykorzystuje tę ostatnią. Niezależnie od takiego czy innego konkretnie wykorzystywanego zasobu leksykalnego typu wordnet, zagadnienie identyfikacji bytów reprezentowanych w tekstach jest jednym z kluczowych zagadnień semantyki, i ogólniej, przyczynkiem do rozumienia tekstu (czy to przez automat, czy też przez człowieka), a udowodnienie przydatności do tego celu technologii wordnetowej pokazuje jej użyteczność wykraczającą daleko poza wąskie rozumienie dziedziny Question Answering (do którego odwołuje się Autor poprzez system RAFAEL). Uwaga Autora, że DeepER jest „uogólnieniem i pogłębieniem" dla technologii NER (Named Entity Reconition) poprzez wyjście poza nazwy własne jest w sposób oczywisty poprawna. Jej istotą jest to, że „pogłębienie i uogólnienie" polega na obserwacji, że nazwy gatunkowe (w podanym przykładzie „rybitwa pospolita") zachowują się składniowo i semantycznie podobnie jak nazwy własne z tym, że nie w odniesieniu do indywiduów lecz różnego rodzaju zbiorowości.

Odrębną sprawą jest to, że algorytm DeepER tylko częściowo realizuje funkcję rozumienia i nie dochodzi do poziomu głębokiej reprezentacji znaczenia, jak o tym świadczą przykłady zamieszczone w tekście rozprawy (np. str. 84-85). Autor zdaje sobie sprawę z ograniczeń swojej metody, czemu daje wyraz np. w rozdziale 6.1.1. pt. „Błędy głębokiego rozpoznania nazw". Stwierdza on tam (str. 108, rozdział 6, Wnioski), że „technika DeepER jest ograniczona głównie przez brak ujednoznacznienia słów, co utrudnia wnioskowanie oparte o WordNet". Całkowicie zgadzam się z tą uwagą, sądzę ponadto, że dopiero rozwiązanie problemu ujednoznacznienia słów – równoważne de facto ich rozumieniu w konkretnym użyciu (tu w kontekście definicji) – usprawiedliwia określenie „głębokie rozpoznawanie nazw". *Jest to jednak dopiero pespektywa „dalszych prac", co zapowiada Doktorant w* rozdziale 6.2. (pt. „Dalsze prace").

Za główną wartość pracy uważam nie tyle pokonanie problemów technicznych, stosunkowo prostych, co pomysłowe i nowatorskie wykorzystanie z jednej strony technologii wordnetowej, z drugiej zaś polskiej Wikipedii (choć Autor zastrzega na str. 65, że wybór Wikipedii jest jedną z możliwych opcji). Nazwy własne, jak to jest pokazane na Rysunku 4.1. na str. 86) wymagają ekstrakcji listy synsetów opisujących byt z definicji encyklopedycznej (podobnie dla nazw gatunkowych). Nietrudno zauważyć, że tworzoną w ten sposób „bibliotekę bytów" systemu RAFAEL można przekształcić do formatu wykorzystywanego w nim systemu wordnetowego upraszczając sam system RAFAEL; z drugiej strony, odpowiednio rozbudowany system wordnetowy w prosty sposób konwertuje się do biblioteki bytów (wręcz ją zastępując). Warto zauważyć, że skuteczność algorytmu DeepER (i systemów typu RAFAEL) będzie zależała od własności wykorzystywanej leksykalnej sieci semantycznej. Autor zdaje sobie z tego sprawę, co *implicite* wynika z Rozdziału 4.1. („Pokrewne rozwiązania").

Praca ma układ bardzo przemyślany i dojrzały. Poza omówioną powyżej zawartością merytoryczną (z akcentem na główny wynik pracy) należy docenić rzetelną dyskusję uzyskanych wyników, czemu poświęcone są Rozdział 5, pt. „Ewaluacja" oraz Rozdział 6. pt.

z Vel

„Wnioski". Warościowe są także sekcje poświęcone ukazaniu badań pokrewnych i źródeł inspiracji choć odnotowałem pewne braki w Rozdziale 2.4. pt. „Rozwiązania dla języków słowiańskich", gdzie stosowne byłoby wzmiankowanie publikowanych wielokrotnie ogólnych wyników Jędrzeja Osińskiego odnośnie obsługi zapytań o czas i przestrzeń (nb. zaimplementowanych w znanym Doktorantowi projekcie POLINT-112-SMS).

Słabe strony pracy są nieliczne i nie przeważają w zdecydowanie pozytywnym bilansie.

Jako słabą stronę rozwiązania zaproponowanego przez Doktoranta uważam to, że algorytm wydobywania synsetów z definicji encyklopedycznych (Wikipedia) zdaje się na numerację znaczeń słów (słowosensów), przypisując jej zgodność z kryterium częstości występowania. Arbitralność tej decyzji wynika stąd, że brak jest przekonywujących argumentów (Autor takowych nie przytacza), że plWordNet w sposób zasadny i konsekwentny stosuje numerację znaczeń wg tej zasady. (Osobiście uważam, że kryterium częstościowe jest nieuprawnione, gdyż trudno jest mówić o jakiejś absolutnej częstości wystąpień zjawiska lingwistycznego abstrahując od przestrzeni obserwacji.) Ponadto, w charakter statystyczny takiego podejącia wpisuje się nieuchronność popełniania błędów. W efekcie, wobec braku istotnie głębokiej analizy semantycznej, generowane mogą być błędy podobne do odnotowanego przez Autora błędnego przypisania Prezydentowi Komorowskiemu w charakterze deskryptora synsetu wskazującego na to, że jest on prezydentem miasta (por. Rys. 4.1. na str. 86). Oczywiście, w tym wypadku błędnie przypisany deskryptor mógłby być skorygowany (ręcznie) przez zastąpienie synsetu <prezydent1, prezydent miasta1> przez synset typu <prezydent2, krezydent kraju1>. Wymagałoby to jednak dużego nakładu (ręcznej?) pracy weryfikacyjnej dla poprawienia biblioteki bytów uwzględniającej analizę semantyczną (z której jednak Autor świadomie rezygnuje).

Oceniając formalną stronę pracy należy docenić fakt, że praca napisana jest bardzo starannie, poprawnym i czytelnym językiem, czym wyróżnia się zdecydowanie korzystnie spośród znanych mi rozpraw doktorskich. W niektórych istotnych przypadkach Autor podejmuje trud uzasadnienia podjętych decyzji, dla obrony przed arbitralnością. Tak jest np. w przypadku polisemi utrudniającej wskazanie właściwego synsetu (str. 89).

Tym niemniej, Autor nie ustrzegł się pewnych usterek formalnych. Istotnym mankamentem, znacznie utrudniającym studiowanie pracy, jest brak indeksu rzeczowego. Autor najwyraźniej nie pomyślał o tym, że może się zdarzyć czytelnik biorący do ręki egzemplarz wydrukowany pracy, coć zadbał o narzędzia o znaczeniu drugorzędnym, jak spis tabel czy spis algorytmów. Dobrym pomysłem jest za to umieszczenie w hasłach bibliografii odniesienia do miejsca zacytowania danej pozycji (strona) (co świadczy o tym, że Autor jednak bierze pod uwagę potrzeby czytelnika wersji papierowej!).

Z błędow terminologicznych odnotowuję błędne (wielokrotne) używanie terminu WordNet jako terminu generycznego. Jest to bowiem nazwa własna odnosząca się do konkretnej, dobrze znanej, leksykalnej bazy danych stworzonej dla języka angielskiego przez zespół prof. Millera (Czasami nazwa ta zastępowana jest przez skrót PWN (Princeton WordNet)). Użycie generyczne wymaga małych liter („wordnet") lub określenia analitycznego („system typu WordNet"). Reasumując, użycie terminu „WordNet" w charakterze generycznym lub zastępczo dla plWordNet jest niepoprawne.

Opracowanie tekstu jest bez zarzutu. Nie stwierdziłem usterek ortograficznych czy istotnych błędów interpunkcyjnych. Literówki typu „Pollitechnika" zamiast „Politechnika" (str. 95) są rzadkością.

Z. Vet

Odrębną kwestią budzącą wątpliwości recenzenta jest sprawa tytułu rozprawy. Po pierwsze definicje pojęć tytułowych winny być umieszczone we wstępie, a nie ukryte w tekście na dalszych stronach. Dotyczy to w pierwszym rzędzie pojęcia „głębokie rozpoznawanie nazw", którego definicja znajduje się dopiero na stronie 81 rozprawy. Dopiero tu Czytelnik zorientuje się, że pojęcie to jest zbyt obiecujące, a cały tytuł na wyrost, jako, że przymiotnik „głęboki" sugeruje dezambiguizację nazwy na poziomie rozumienia (tj. przynajmniej na poziomie semantycznym), podczas gdy z przykładu ze str. 85 i algorytrmu opisanego na str. 86-90 wynika, że opis czynności kryjących się pod tą nazwą ogranicza się do powierzchniowego przypisania synsetów, co nie jest jednoznaczne z rozumieniem

Konkluzja

Biorąc pod uwagę sumę poczynionych obserwacji stwierdzam, że nie mam najmniejszych wątpliwości, iż przedłożona rozprawa spełnia warunki stawiane przez Ustawę o Stopniach i Tytułach Naukowych w stopniu w umożliwiającym przejście do dalszych etapów przewodu doktorskiego. Wnioskuję zatem o przyjęcie rozprawy oraz o dopuszczenie jej Autora do kolejnych etapów tegoż przewodu.

Poznań, dnia 3.04.2015

Prof. zw. dr hab. Zygmunt Vetulani