Halifax, 25 maja 2015

Prof. dr hab. Stanisław Matwin

Instytut Podstaw Informatyki
Polska Akademia Nauk
Warszawa

Faculty of Computer Science
Dalhousie University
Halifax, Kanada

Recenzja rozprawy doktorskiej pana mgr. inż. **Aleksandra Smywińskiego-Pohla** pt. „**Automatyczna ekstrakcja relacji semantycznych z tekstów w języku polskim**”, Akademia Górniczo-Hutnicza, Kraków.

Promotorem pracy jest **Prof. dr hab. Wiesław Lubaszewski**

1. Zawartość rozprawy

Rozprawa składa się z 11 rozdziałów i liczy ponad 180 stron tekstu. Rozprawa zwiera też 15-stronicową bibliografię i sześć dodatków. Zawartość pracy można podzielić na dwie zasadnicze części: w części pierwszej (rozdz. 1-4) Autor wprowadza tematykę, definiuje niezbędne pojęcia i przedstawia stan badań w dziedzinie pracy. W części drugiej (rozdz. 5-11) opisana jest szczegółowo autorska metoda ekstrakcji informacji, w tym zasoby wykorzystywane przez algorytm ekstrakcji, algorytm uczący się wzorców ekstrakcyjnych, metody określania i budowy ograniczeń wzorców i wreszcie zastosowanie całej aparatury skonstruowanej przez autora do ekstrakcji relacji *całość-część* wraz z dyskusją wyników eksperymentalnych. Pracę zamyka podsumowanie zarysowujące kierunki przyszłych badań nad ekstrakcją informacji z tekstów w języku polskim.

2. Uwagi ogólne

Tematyka pracy wpisuje się doskonale w główny nurt współczesnych badań stosowanych w sztucznej inteligencji w ogólności, a w dziedzinie przetwarzania języka naturalnego i analizy danych tekstowych (*text analytics*) w szczególności. Nie ulega wątpliwości, że praca jest cennym wkładem do stosowanej lingwistyki komputerowej, szczególnie w kontekście języka polskiego. Należy przyklasnąć kunsztownemu użyciu całej palety nowoczesnych zasobów lingwistycznych i informatycznych i docenić inżynierski wysiłek i umiejętności niezbędne w realizacji projektu. Warto zauważyć staranność i skrupulatność w opisie przeprowadzonych badań. Praca,

oczywiście dostosowana do kontekstu lokalnego, mogłaby być z powodzeniem przedstawiona i obroniona w wielu wiodących światowych ośrodkach badawczych.

3. Oryginalne osiągnięcia

Prace empiryczne ze sztucznej inteligencji w sposób naturalny z reguły budują na dokonaniach poprzedników, tutaj są nimi R. Girju oraz D. Milne i I. Witten. Doktorant rzetelnie przedstawia, w jaki sposób wykorzystał i, co ważne, istotnie rozszerzył prace tych autorów. Należy zauważyć, że Autor chyba jako pierwszy połączył użycie Cyc i DBpedii w jednym systemie. Pan mgr inż. Pohl-Smywiński zbudował własnoręcznie duży, złożony system, zdolny przetwarzać znacznej wielkości korpusy danych. Autor wykorzystał też w sposób ekspercki bogaty zestaw zasobów lingwistycznych i ontologiczno-semantycznych. Główna teza zaproponowana przez Autora, stanowiąca że podejście hybrydowe łączące użycie metod symbolicznych i statystycznych jest silniejsze niż każda z jego składowych, jest uzasadniona w sposób przekonujący, w oparciu o metodycznie przeprowadzone eksperymenty z użyciem zbudowanego przez Autora oprogramowania. Zbudowane narzędzia, otrzymane z ich użyciem wyniki i ich dyskusja dowodzą informatycznej i inżynierskiej dojrzałości Autora.

Praca zawiera wiele oryginalnych, interesujących rozwiązań. Na pierwszym miejscu wymieniłbym tu metodę wyszukiwania w korpusie zdań zawierających relacje semantyczne (rozdz. 7.1). Wydaje się, że metoda ta, odpowiednio przedstawiona, byłaby warta opublikowania w jednym z czołowych pism z dziedziny stosowanej lingwistyki komputerowej.

4. Uwagi szczegółowe i komentarze

Lektura 180 stron rozprawy skłania mnie do wniosku, że praca mogłaby być lepiej przedstawiona, być może w sposób nieco bardziej hierarchiczny − od ogółu do szczegółu − niż to ma miejsce w manuskrypcie. Praca opisuje dość złożony system, składający się z wielu części i używający szeregu zasobów zewnętrznych; celowe byłoby ułatwić czytelnikowi zrozumienie tej struktury. I tak np. łańcuch operacji opisany w rozdz. 5.2 powinien być opisany w formie pseudokodu, uwzględniając jako wejścia konkretne używane zasoby lingwistyczne i algorytmiczne (jak np. tłumaczenie „haseł" Cyc na polski). Następnie, dla dalszego ułatwienia lektury i zrozumienia szczegółów metody, pożyteczne byłoby przedstawienie całości zbudowanego systemu w postaci schematu, być może w formie hierarchicznej. Część materiału zaczerpniętego z prac innych autorów jest tak szczegółowa, że bez uszczerbku dla zrozumienia całości mogłaby zostać umieszczona w dodatkowym załączniku (mam tu na myśli np. sporą część rozdz. 7., gdzie wymienione algorytmy mogłyby być omówione skrótowo i opisowo, a szczegóły zostałyby przesunięte do dodatku).

Omówienie literatury tematu jest imponujące w swej obszerności i szczegółowości. Autor wykazuje się tu dogłębnym zrozumieniem zagadnienia i widzi je na bogatym i ogólnym tle –

daje znać o sobie niecodzienne, szerokie wykształcenie doktoranta. Dwa drobne przykłady to odwołanie do fundamentalnych prac Tarskiego dotyczących pojęcia prawdy lub wspomnienie interpretacji desygnatorów przez Kripkego. Autor rozumie też bardzo dobrze historyczną ewolucję metod ekstrakcji informacji, od technik symbolicznych do podejść statystycznych, punktując słusznie zalety i wady każdej z nich. Szczególnie interesujące dla czytelnika jest omówienie prac nad ekstrakcją informacji z tekstów polskich, stanowiące kompletne i użyteczne omówienie tematu (być może warte odrębnej publikacji).

W imię kompletności przeglądu literatury należy odnotować brak (poza jedną wzmianką) dyskusji systemu NELL T. Mitchella i jego grupy z Carnegie Mellon University − systemu ekstrakcji, rozwijanego i samouczącego się na przestrzeni kilku już lat, ogólnie znanego i cytowanego. Wydaje się też, że techniki warunkowych pól losowych (*CRF*) zasługują na szersze niż to ma miejsce omówienie i dyskusję, ze względu na ich „skalowalność" i znaczną liczbę konkretnych zastosowań praktycznych w ekstrakcji typu „wypełnianie szablonów".

Niektóre szczegółowe i nieco oboczne stwierdzenia w tekście wymagają być może zastanowienia albo korekty: np. na str. 154 w dyskusji pokrycia dopasowań kategoriami semantycznymi pada stwierdzenie, że ponieważ pokrycie wynosi 80%, dla par powinno ono wynosić 64%. Byłoby to prawdziwe, gdyby elementy par były niezależne od siebie, co oczywiście nie ma miejsca.

Praca jest napisana i zredagowana niezwykle starannie i w zasadzie nie zauważa się żadnych błędów ani literówek wymagających poprawy (z maleńkim wyjątkiem odnośnika 10 na str. 12, powinno być "nazewniczych" zamiast „nazewniczy").

Pewne decyzje podjęte w projektowaniu i realizacji całego systemu nasuwają oczywiste pytania dotyczące ich alternatyw. I tak np.:

- Algorytm C4.5, użyty do syntezy ograniczeń semantycznych z automatycznie zbudowanych przykładów, mógłby być z powodzeniem zastąpiony przez inne metody, być może o lepszej sprawności w tym zadaniu, aby wymienić tylko lasy decyzyjne (*random forest*) czy chociażby *bagging* drzew decyzyjnych.
- Czy użycie odległości cosinusowej, tak powszechnej dla danych tekstowych, nie byłoby użyteczne w miarach definiowanych w rozdz. 8.6?
- Porównawcze wyniki całego projektu zależą najpewniej od wyboru 10% zdań poddanych obróbce ręcznej – oczywiście trudno sobie ze względów praktycznych wyobrazić ewaluację krzyżową lub inną metodę wymagającą ręcznej obróbki znacznie większego podkorpusu, ale należałoby przynajmniej odnotować tę zależność.
- Po lekturze całości pracy nie jest jasne, jak zrealizowane jest mapowanie typowo polskich jednostek referencyjnych na odpowiadające im ogólniejsze pojęcia w Cyc (np. nazwy partii politycznych lub nazwiska celebrytów?). Mam wrażenie, że jest to opisane w pracy, ale trudno mi było wyodrębnić ten aspekt z całego tekstu.

- W rozdz. 10 użyteczne byłoby przedstawienie, choćby orientacyjne, czasu wykonania poszczególnych składowych zbudowanego przez Autora łańcucha ekstrakcji, jak też i wysiłku „ludzkiego” włożonego w ewaluację wyników.

Byłoby interesujące zastanowić się nad nieco szerszymi pytaniami należącymi do przyszłych badań, rozumianych nie tylko jako przyszłe prace autora, ale ogólniej jako zagadnienia dotyczące proponowanego tu podejścia do ekstrakcji informacji z danych tekstowych. I tak np.:

- czego wymagałoby (oprócz spraw oczywistych, jak narzędzie tłumaczące) zastosowanie zaproponowanego podejścia do innego języka? A zastosowanie „wprost”, tj. do tekstów w języku angielskim?
- Czy sensowne i możliwe byłoby zastąpienie użycia Cyc w fazie wyboru zdań Wordnetem?
- Czy dołożenie dodatkowego słownika z lepszym pokryciem jednostek referencyjnych (np. Freebase) mogłoby polepszyć pokrycie?

5. Ocena ogólna i uwagi końcowe

Praca p. Smywińskiego-Pohla stanowi niewątpliwie cenny wkład do sztucznej inteligencji, a w szczególności do lingwistyki komputerowej. Praca zawiera oryginalne pomysły i ich metodologicznie kompetentną, profesjonalną realizację. Praca jest napisana bardzo starannie i z dużą dbałością o formę prezentacji. Uwagi szczegółowe zawarte w recenzji nie pomniejszają zdecydowanie pozytywnej oceny pracy.

Podsumowując stwierdzam, że praca doktorska p. mgr. inż. Aleksandra Smywińskiego-Pohla spełnia wszystkie wymogi odnośnej ustawy i powinna być dopuszczona do publicznej obrony.