dr hab. Tomasz Jurdziński
Istytut Informatyki
Uniwersytet Wrocławski

Wrocław, 25 lipca 2014

Recenzja rozprawy doktorskiej Aliny Wróblewskiej „Polish Dependency Parser Trained on an Automatically Induced Dependency Bank"

Przedstawiona rozprawa doktorska poświęcona jest tematyce automatycznego pozyskiwania danych na potrzeby uczenia maszynowego systemów parsowania zależnościowego. Trenowane metodą uczenia z nadzorem parsery zależnościowe pozwalają obecnie na skuteczne odkrywanie struktury zależnościowej zdań. Jednak ich zastosowanie dla języka polskiego jest ograniczone z uwagi na brak dostatecznie dużych banków drzew zależnościowych. Niemożliwe jest zatem wytrenowanie dostępnych parserów zależnościowych na potrzeby analizy zależnościowej tekstów polskich. Ponieważ ręczne tworzenie takich struktur jest kosztowne i czasochłonne, autorka rozprawy podjęła temat tworzenia takich struktur w sposób automatyczny lub półautomatyczny.

W rozprawie badano dwie metody tworzenia drzew zależnościowych. Pierwsza metoda opiera się na automatycznej konwersji dostępnych drzew składnikowych (constituency trees) do postaci drzew zależnościowych. W drugiej metodzie oparto się na wykorzystaniu korpusów równoległych: odwzorowaniu drzew zależnościowych zdań w języku angielskim na struktury zależnościowe w odpowiednich zdaniach w języku polskim.

Z uwagi na brak schematu anotacji zależnościowej zdań uwzględniającej wszystkie ważne cechy specyficzne dla języka polskiego, praca nad automatycznym tworzeniem struktur zależnościowych poprzedzona została wprowadzeniem przez autorkę schematu anotacji (rozdział 3). Zadanie to wymagało wiedzy, doświadczenia i wyczucia w zakresie lingwistyki. Stworzony schemat został bardzo skrupulatnie i dokładnie opisany, zapewne będzie dobrym materiałem wyjściowym w dalszych badaniach automatycznego parsowania zależnościowego dla języka polskiego.

Opisana w rozdziale 4 metoda oparta na konwersji drzew składnikowych do postaci drzew zależnościowych wydaje się stosunkowo prosta, przynajmniej jako zadanie o charakterze informatycznym. Zbiór krawędzi tworzących drzewo zależności można wywieść z informacji zawartych w drzewie składnikowym przy pomocy dość prostych reguł. Niemniej, co typowe przy przetwarzaniu języka naturalnego, zastosowanie automatycznych reguł nie pokrywa szeregu sytuacji szczególnych i może dawać w wyniku niepoprawne drzewa zależnościowe. Dlatego konieczne było wprowadzenie adekwatnych automatycznych metod przebudowy powstałych drzew. Dodatkowym niełatwym zadaniem było etykietowanie uzyskanych zależności, w oparciu o dostępne informacje lingwistyczne zawarte w drzewie składnikowym.

System projektowany był na potrzeby konwersji drzew z banku składnikowych drzew składniowych *Składnica frazowa*. Około 90% spośród ponad 8000 utworzonych drzew zależnościowych wykorzystanych zostało jako zbiór treningowy dla parserów zależnościowych. Autorka wykorzystała je w procesie uczenia dwóch parserów zależnościowych: Malt i Mate. Pozostałe 10% drzew wykorzystane zostało do oceny jakości wytrenowanych parserów. Uzyskane zostały bardzo obiecujące wyniki, dające odsetek poprawnych zależności i etykiet porównywalne z wcześniejszymi wynikami dla innych języków, gdzie zbiory treningowe zostały wygenerowane ręcznie. Bardzo dobre wrażenie robi cały proces ewaluacji uzyskanego rozwiązania. Oprócz drzew uzyskanych przez konwersję zdań z banku *Składnica*, przygotowany został (częściowo ręcznie) zestaw 100 bardziej skomplikowanych zdań na których wykonano dodatkowe testy. Zbadano również wpływ na jakość parsera faktu, czy anotacje tokenów generowane są automatycznie czy też ręcznie. Rozważono zasadność doboru różnych parametrów parserów a także wybór progu dla odsetka krzyżujących się krawędzi (non-projective).

Podsumowując część poświęconą metodzie konwersji, należy stwierdzić, że jest ona oparta o raczej znane i naturalne techniki i narzędzia, choć przystosowanie ich do specyfiki języka polskiego wymagało istotnego nakładu twórczej pracy. Ponadto, wykonana została kompleksowa

ocena jakości uzyskanego rozwiązania. Pozwoliła ona potwierdzić jego skuteczność a jednocześnie stanowiła dobry punkt odniesienia dla alternatywnej metody pozyskiwania drzew zależnościowych, przedstawionej w dalszej części rozprawy.

Druga metoda automatycznego pozyskiwania drzew zależnościowych zaprojektowana w rozprawie (opisana w rozdziale 5) oparta jest na technice rzutowania informacji lingwistycznych. W skrócie celem jest „odwzorowanie" drzew zależnościowych dla zdań w języku angielskim na odpowiednie drzewa w języku polskim. Podstawą tego procesu są przyporządkowania słowne odwzorowujące tokeny ze zdań w jednym języku na ich odpowiedniki w drugim z języków. Z uwagi na fakt, że przyporządkowania takie nie są bijekcjami, sposób odzorowania zależności nie jest oczywisty. W rozprawie zaprezentowana została nowa technika odwzorowywania zależności. Najpierw, w oparciu o różne sposoby przyporządkowywania tokenów, tworzony jest ważony graf dwudzielny ilustrujący możliwe przyporządkowania tokenów i ich „wiarygodność". Na tej podstawie tworzony jest skierowany graf ważony ilustrujący możliwe zależności w zdaniach polskich. Wcześniejsze rozwiązania na tym etapie dążyły do utworzenia poprawnego drzewa zależnościowego. Metoda zaproponowana w rozprawie poddaje dalszej analizie uzyskany ważony graf skierowany, wprowadzając iteracyjną metodę modyfikacji wag w oparciu o wybór k najlepszych drzew rozpinających. Na koniec, na grafie ze zmodyfikowanymi wagami wybierane jest największe drzewo rozpinające. W rozprawie przeprowadzono kompleksową ewaluację zaproponowanej metody. Parsery zależnościowe zostały wytrenowane w oparciu o różne dostępne zestawy równoległych tekstów angielskich i polskich (zawierające kilkanaście milionów zdań). Kompletna implementacja zaproponowanej metody i jej ewaluacja wymagały rozwiązania szeregu problemów nie wyspecyfikowanych przez ogólną technikę, m.in. sposobu tworzenia grafów dwudzielnych odwzorowujących przyporządkowania, odfiltrowywania zdań, dla których wynikowe drzewa były niepoprawne, sposobu odwzorowywania zależności angielskich na polskie (w tym etykietowanie zależności). W ramach ewaluacji wytrenowane parsery przetestowane zostały na tych samych zestawach zdań, na których wcześniej oceniana była metoda z rozdziału 4. Uzyskane rezultaty w zakresie odsetka poprawnych zależności i etykiet były gorsze od wyników dla metody opartej na konwersji drzew składnikowych, co nie jest zaskakujące. Niemniej różnica nie była duża, co uzasadnia stosowanie metody opartej na rzutowaniu w sytuacji, gdy w danym języku brak jest banku drzew składnikowych. Co ciekawe, w przypadku 100 testowych bardziej skomplikowanych zdań spoza banku *Składnica*, wyniki parsera wytrenowanego na danych uzyskanych metodą projekcji były porównywalne z rezultatami uzyskanymi metodą konwersji.

Technika zaproponowana w piątym rozdziale rozprawy jest nowa, oryginalna i niezależna od specyfiki konkretnego języka naturalnego. Jej realizacja stanowiła też poważniejsze wyzwanie implementacyjne, z uwagi na duże skomplikowanie i złożoność algorytmów, istotnie większy rozmiar danych do przetworzenia jak i więcej etapów procesu uzyskiwania drzew zależnościowych. Przeprowadzona rzetelna ocena jakości zaproponowanej techniki potwierdziła jej skuteczność. **Sądzę, że zaproponowana metoda stanowić będzie trwały wkład w rozwój narzędzi do automatycznego parsowania zależnościowego.**

Na podkreślenia zasługuje rzetelność w prezentacji wyników badań - w każdej części rozprawy autorka przedstawiła wyczerpująco swoje wyniki w kontekście innych badań o zbliżonej tematyce. Uzupełniła też swoje wyniki przystępnym wprowadzeniem, dzięki czemu czytanie rozprawy nie jest uciążliwe również dla osób nie zajmujących się na co dzień przetwarzaniem języka naturalnego (należy do nich autor niniejszej recenzji). Wysoko też oceniam dociekliwość i staranność w przeprowadzonych wyczerpujących badaniach przedstawionych technik.

W kontekście metod oceny zaproponowanych rozwiązań chciałbym zgłosić dwie drobne uwagi krytyczne. W rozprawie podkreślona została nowatorskość zaproponowanej metody rzutowania ważonego. Sądzę, że jest to uzasadnione. Niemniej, uzupełnienie przeprowadzonej oceny na równoległym korpusie polskim i angielskim o implementację któregoś z wcześniejszych rozwiązań stosującego metodę rzutowania informacji lingwistycznych dałoby ciekawszy materiał do porównań, niż tylko porównanie z wynikami metody z rozdziału 4. Druga kwestia to przyczyny

spadku jakości parsera opartego na metodzie rzutowania po zwiększeniu liczby iteracji procesu uczenia. Sytuacja ta wydaje się nieco zagadkowa. Autorka ograniczyła się do hipotezy, że powodem mógł być szum, którego parser „uczy się" w kolejnych iteracjach. Czy takie zjawisko jest częste w procesie uczenia maszynowego dla danych lingwistycznych? Może warto zaplanować eksperymenty dla zweryfikowania tej hipotezy?

Konkluzja

Rozprawa Pani Aliny Wróblewskiej zawiera oryginalne i ciekawe wyniki dotyczące przetwarzania języka naturalnego. Pozwoliły one potwierdzić hipotezę o możliwości (pół)automatycznego pozyskiwania danych uczących dla parserów zależnościowych i stworzyć odpowiednie narzędzia. W rozprawie zaproponowano nową technikę pozyskiwania danych z korpusów równoległych o potencjalnie rozległych zastosowaniach. Uzyskanie przedstawionych wyników wymagało połączenia wysokich kompetencji z zakresu lingwistyki z zaawansowanymi narzędziami informatyki. Bez wahania można stwierdzić, że w prezentowanej rozprawie Pani Wróblewska wykazała szeroką wiedzę oraz umiejętność samodzielnego prowadzenia pracy naukowej.

Uważam, że rozprawa w pełni spełnia wymogi ustawy o stopniach i tytule naukowym i wnioskuję o dopuszczenie do dalszych etapów przewodu doktorskiego.