INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES

Alina Wróblewska

Polish Dependency Parser Trained on an Automatically Induced Dependency Bank

PhD Dissertation

Supervisor
dr hab. Adam Przepiórkowski, prof. IPI PAN

Warsaw 2014

Streszczenie

W ostatnich latach coraz większą wagę przywiązuje się do parsowania zależnościowego, czyli do automatycznej analizy składniowo-semantycznej zdań. Dzieje się tak dlatego, że parsowanie wydobywa strukturę predykatywno-argumentową zdania, której można użyć do udoskonalenia systemów dialogowych, tłumaczenia maszynowego, czy ekstrakcji informacji. Większość współczesnych systemów parsowania zależnościowego opiera się na metodach statystycznych. Na podstawie danych treningowych parsery uczą się, jak należy analizować zdania w języku naturalnym i generować odpowiednie struktury zależnościowe dla tych zdań. Jak dotychczas najlepsze wyniki osiągają parsery trenowane za pomocą metod z nadzorem. Parsery zależnościowe trenowane na poprawnie zaanotowanych danych są bardzo skuteczne, nawet w odniesieniu do języków ze swobodnym szykiem zdania, takich jak czeski czy bułgarski.

Niemniej jednak metody z nadzorem wymagają dużej liczby poprawnie zaanotowanych struktur zależnościowych, które powstają w wyniku bardzo czasochłonnego i kosztownego procesu anotacji ręcznej. Dla wielu języków nadal nie istnieją żadne banki struktur zależnościowych, dlatego poszukuje się alternatywnych metod trenowania parserów albo pozyskiwania danych treningowych. Ponieważ uczenie bez nadzoru często nie jest najlepszym rozwiązaniem głównie za sprawą małej efektywności oraz bardzo dużej złożoności obliczeniowej, w niniejszej rozprawie doktorskiej rozpatrujemy alternatywne metody pozyskiwania wysokiej jakości struktur zależnościowych oraz szukamy odpowiedzi na następujące pytania badawcze:

1. Czy jest możliwe automatyczne (lub półautomatyczne) pozyskiwanie drzew zależnościowych?

2. Czy można przy pomocy metod z nadzorem wytrenować parser zależnościowy na automatycznie lub półautomatycznie pozyskanych danych?

Dysertacja rozpoczyna się teoretycznym opisem głównych założeń gramatyki zależnościowej oraz parsowania zależnościowego (rozdział drugi). W rozdziale trzecim został przedstawiony schemat anotacji zależnościowej zdań w języku polskim. Schemat ten definiuje zbiór reguł, przy pomocy których można wyznaczyć relacje dominacji (albo zależności) pomiędzy tokenami w zdaniu. Schemat jest dostosowany do specyfiki języka polskiego i bierze pod uwagę główne zjawiska lingwistyczne opisane w literaturze i występujące w losowo wybranych zdaniach. Schemat wyróżnia 28 typów relacji zależnościowych podzielonych na trzy grupy: relacje zawierające podrzędniki pełniące funkcje argumentów, relacje zawierające podrzędniki niepełniące funkcji argumentów oraz relacje przeznaczone do anotowania konstrukcji z koordynacją. Zgodnie z tym schematem anotujemy automatycznie wygenerowane struktury zależnościowe zdań w języku polskim.

W dysertacji zostały zaprezentowane dwie metody automatycznego pozyskiwania struktur zależnościowych. Pierwsza metoda wykorzystuje ideę konwersji drzew składnikowych do postaci drzew zależnościowych (rozdział czwarty). Wykorzystanie metody konwersji jest możliwe, ponieważ dla języka polskiego istnieje bank struktur składnikowych. W związku z tym, że relacje zależnościowe można stosunkowo łatwo wywieść ze struktur składnikowych z wyróżnionymi elementami głównymi, nacisk jest położony przede wszystkim na dostosowanie przekonwertowanych struktur do schematu anotacji drzew zależnościowych oraz na przypisanie etykiet do krawędzi w przekonwertowanych drzewach. W celu dostosowania struktur do schematu anotacji opracowano zbiór reguł modyfikujących relacje pomiędzy tokenami w konstrukcjach strony biernej oraz w konstrukcjach zawierających frazy nieciągłe, zdania podrzędne, frazy z korelatem, czy spójniki inkorporacyjne. Ponieważ współczesne systemy parsowania zależnościowego są dostosowane do uczenia modeli na drzewach, których krawędzie mają przypisane etykiety, istotne znaczenie miało opracowanie zbioru reguł etykietujących krawędzie w przekonwertowanych drzewach funkcjami gramatycznymi podrzędników danych relacji. Ostatecznym wynikiem procesu konwersji jest bank 8227 drzew zależnościowych z etykietami przypisanymi do krawędzi. W celu oceny jakości pozyskanych drzew zależnościowych wykorzystano zewnętrzną metodę ewaluacji (ang. 'extrinsic evaluation'). Metoda ta polega na wytrenowaniu parsera zależnościowego na przekonwertowanych drzewach, a następnie na ocenie wpływu danych treningowych na jakość parsowania. Zgodnie z wynikami, w stosunkowo prostych polskich zdaniach nawet 92,7% tokenów może mieć przypisany poprawny nadrzędnik, a 87,2% tokenów może mieć przypisany poprawny nadrzędnik oraz poprawną funkcję gramatyczną etykietującą relację. W przypadku bardziej skomplikowanych i rozbudowanych zdań wyniki te są zdecydowanie niższe – 76,6% tokenów ma przypisany poprawny nadrzędnik, a 70,1% tokenów ma przypisany poprawny nadrzędnik oraz etykietę relacji.

Drugi sposób automatycznego pozyskiwania drzew zależnościowych jest oparty na nowatorskiej metodzie rzutowania ważonego (rozdział piąty). Główna idea metody rzutowania informacji lingwistycznych polega na odwzorowaniu anotacji lingwistycznych w

zdaniach z części korpusu równoległego w jednym języku na odpowiednie zdania z części korpusu w drugim języku. Informacje lingwistyczne są rzutowane z wykorzystaniem automatycznie wygenerowanych przyporządkowań słownych (ang. 'word alignment'). W przedstawionej w rozprawie i opartej na idei rzutowania procedurze pozyskiwania struktur zależnościowych dla zdań w języku polskim można wyróżnić dwa główne kroki: rzutowanie ważone angielskich relacji zależnościowych na zdanie polskie oraz indukcję ważoną drzew zależnościowych na podstawie zbioru rzutowanych krawędzi. Angielskie relacje zależnościowe są rzutowane na odpowiednie zdania polskie poprzez rozbudowany zbiór przyporządkowań słownych z przypisanymi wagami. W wyniku tego zdaniom polskim zostają przypisane grafy skierowane z wagami na krawędziach. Wagi krawędzi są szacowane na podstawie wag przyporządkowań słownych wykorzystanych w rzutowaniu. Indukcja ważona polega na szukaniu – w grafach skierowanych zawierających rzutowane krawędzie ze zoptymalizowanymi wagami – maksymalnych drzew rozpinających, które spełniają kryteria poprawnego drzewa zależnościowego. Do optymalizacji wag wykorzystano rozkład prawdopodobieństwa krawędzi w k najlepszych drzewach rozpinających znalezionych w rzutowanym grafie skierowanym. Rozkład prawdopodobieństwa krawędzi można obliczyć za pomocą zmodyfikowanej wersji algorytmu EM. Nowatorstwo przedstawionej metody polega na włączeniu czynnika ważenia do procesów rzutowania relacji oraz indukcji struktur zależnościowych. W rzutowanych grafach skierowanych ze zoptymalizowanymi wagami na krawędziach znaleziono prawie 4 miliony maksymalnych drzew rozpinających spełniających kryteria poprawnego drzewa zależnościowego. Następnie krawędziom w drzewach pozyskanych z wykorzystaniem metody rzutowania ważonego zostają przypisane etykiety. Parser wytrenowany na takich drzewach znajduje poprawne nadrzędniki dla 74,6% tokenów, a poprawne nadrzędniki wraz z poprawną etykietą relacji dla 69,4% tokenów. W następstwie zastosowania dodatkowych reguł korygujących oraz filtrujących w odniesieniu do wyindukowanych drzew, parser wytrenowany na ulepszonym zbiorze drzew przypisuje poprawne nadrzędniki do 86,0% tokenów, a poprawne nadrzędniki i poprawne etykiety relacji do 80,5% tokenów. Mimo że te wyniki są istotnie niższe niż wyniki osiągnięte przez parser trenowany na przekonwertowanych drzewach, ewaluacja w oparciu o dłuższe i bardziej skomplikowane struktury pokazuje, że parser trenowany na drzewach pozyskanych metodą ważonej indukcji działa nieznacznie lepiej niż parser trenowany na przekonwertowanych drzewach.

Na podstawie wyników przeprowadzonych eksperymentów możemy udzielić pozytywnych odpowiedzi na zadane pytania badawcze, ponieważ udało nam się pozyskać struktury zależnościowe w sposób automatyczny, wykorzystując metody oparte na konwersji i indukcji, a także wytrenować parser na automatycznie wygenerowanych strukturach zależnościowych.