**Z5** Efektywny parser składniowo-semantyczny

**Z5.2** Rozbudowa kategorialnego parsera składniowo-semantycznego

**M3** Analiza możliwości półautomatycznego rozwijania gramatyki w oparciu o istniejące banki drzew

Okres sprawozdawczy: lipiec – wrzesień 2016

Partner: Instytut Podstaw Informatyki PAN, Warszawa

Typ: raport

30.09.2016, Warszawa

Autor: Wojciech Jaworski, Daniel Oklesiński

# 1. Wprowadzenie

Aktualnie dostępne są następujące zbiory drzew rozbioru składniowego dla języka polskiego:

— Składnica Frazowa: wygenerowana za pomocą parsera Świgra, ręcznie dezambiguowana
— Składnica Zależnościowa: wygenerowana ze Składnicy Frazowej, ręcznie modyfikowana
— pol-składnica-pargram: wygenerowy z użyciem gramatyki POLFIE, ręcznie dezambiguowana
— pol-nkjp1m-pargram-dev: wygenerowy z użyciem gramatyki POLFIE
— Krzaki: wygenerowany ręcznie

Z uwagi na charakterystykę potoku przetwarzania parsera ENIAM, podstawowym celem dla którego istnieje gramatyka kategorialna jest odnajdywanie struktury zależnościowej i w tym właśnie kontekście będzie prowadzona dalsza analiza.

W zadaniu rozwijania gramatyki można wydzielić następujące podzadania:

1. zwiększenie pokrycia, czyli zwiększanie liczby zdań, dla których gramatyka generuje poprawny rozbiór (i być może wiele innych niepoprawych)
2. dezambiguacja drzew rozbioru składniowego
3. walidacja pokrycia gramatyki kategorialnej

# 2. Zwiększanie pokrycia

Wygenerowane automatycznie banki drzew możemy wykorzystać do zwiększania pokrycia gramatyki znajdując w nich drzewa, których parser kategorialny nie potrafi wygenerować, a następnie analizując reguły gramatyczne, prowadzące do ich powstania. Reguły te będzie można przepisać bądź aproksymować w formaliźmie kategorialnym. Maszynowe uczenie zależności językowych na podstawie automatycznie wygenerowanych korpusów wydaje się o tyle bezcelowe, że istnieje już dla nich model, jakim jest gramatyka użyta do ich stworzenia.

Uznajemy jednak, że kluczowe dla rozwoju parsera jest uchwycenie zależności składniowych nie ujętych przez alternatywne gramatyki. Z tego względu jako podstawowe źródła danych do rozwijania pokrycia gramatyki uznajemy korpus Krzaki oraz korpus drzew tworzony w ramach podzadania 5.4 projektu Clarin bis. Oba te korpusy zawierają drzewa zależnościowe w formacie CONLL.

Z powyższy korpusów można automatycznie wydobyć często występujące konstrukcje składniowe stwarzające problemy parserowi kategorialnemu (wymaga to napisania odpowiedniego narzędzia). Następnie parser kategorialny zostanie ręcznie rozszerzony o obsługę tych konstrukcji. Z uwagi na powszechność występowania rozkładu Zipfa we wszelkich zjawiskach lingwistycznych spodziewamy się, że częstotliwość poszczególnych nieobsługiwanych przez parser kategorialny konstrukcji będzie zgodna z tym rozkładem. Oznacza to, że sensowne jest zastosowanie ręcznego rozszerzenia parsera przy częstych konstrukcjach oraz innego, w pełni automatycznego rozwiązania dla sytuacji, gdy występują konstrukcje rzadkie.

W wyniku badań doszliśmy do wniosku, że rozwiązaniem problemu analizy rzadko występujących konstrukcji (jak również konstrukcji nie występujących w korpusie w ogóle) jest scalania fragmentów drzew wygenerowanych przez parser kategorialny za pomocą algorytmu wzorowanego na parserze zależnościowym.

Parser zależnościowy jest uczony w pełni automatycznie na podstawie banku drzew. Analiza parsera zależnościowego (algorytmu MateParser, wybranego przez nas ze względu na to że daje on najlepsze wyniki dla języka polskiego) wykazała, że algorytm składa się z dwóch części:

1. generowanie cech na podstawie lematów i tagów morfosyntaktycznych

2. wybór nadrzędnika za pomocą zmodyfikowanego algorytmu parsera tablicowego, który wybiera go maksymalizując wartości powyższych cech.

Nasza koncepcja polegać będzie na powiązaniu ze sobą częściowych drzew rozbioru generowanych przez parser kategorialny w pełne drzewo rozbioru zgodnie z koncepcją maksymalizacji wartości cech zawartą w algorytmie MateParser.

Realizacja tego rozwiązania wymagać będzie reimplementacji algorytmu MateParser obejmującej istotne zmiany, a mianowicie dostarczanie mu na wejściu struktury generowanej przez parser kategorialny zamiast sekwencji otagowanych tokenów.

Chcemy korzystać z wyuczonych parserem zależnościowym modeli, które są wytwarzane w ramach zadania 5.4 projektu Clarin Bis. Musimy w tym celu nauczyć się konwertować tagset SGJP, z którego korzysta parser kategorialny na tagset NKJP, z którego korzysta parser zależnościowy.

Kolejnym elementem koniecznym do realizacji zadania będzie przetłumaczenie struktury drzew zależnościowych znajdujących się w korpusach oraz generowanych przez parser zależnościowy na strukturę generowaną przez parser kategorialny. Różnice tutaj obejmują między innymi reprezentację koordynacji ze współdzielonym podrzędnikiem oraz reprezentację przyimków niesemantycznych.

## 3. Dezambiguacja

Częściową dezambiguację można uzykać przypisując priorytety poszczególnym regułom gramatycznym (metoda zwana optymality theory). W przypadku gramatyki kategorialnej analogiczną funkcjonalność uzyskuje się dodając wagi do alternatywnych wpisów w leksykonie gramatyki kategorialnej dla poszczególnych leksemów. Wagi te można opracować ręcznie albo wygenerować automatycznie na podstawie zdezambiguowanych banków drzew.

Drugim podejściem jest dezambiguacja niejednoznacznej struktury zależnościowej wykonywana po zakończeniu parsowania. Do tego zadania możemy wykorzystać zmodyfikowany algorytm MateParser: algorytm będzie przechodził skompresowany las niejednoznacznych drzew wybierają nadrzędnik mający lepszą ocenę zgodnie z wartościami cech.

Do dezambiguacji można wykorzystać wszystkie dostępne banki drzew.

## 4. Walidacja

Walidację można wykonać porównując dla kolejnych zdań drzewa wygenerowane przez parser kategorialny z tymi, które są w bankach drzew. Z zadaniem tym wiążą się wzmiankowane już problemy różnicami w tagsecie i założeniach przyjętych przy konstruowaniu drzew, co powoduje, że w przypadku wystąpienia różnic przy porównaniu drzew trzeba będzie ręcznie określać, czy różnica wynika z błędu, czy z odmiennego sposobu reprezentacji danego zjawiska językowego.

Do walidacji można wykorzystać wszystkie dostępne banki drzew.