Poznań, 14 kwietnia 2014

prof. UAM dr hab. Krzysztof Jassem,

Uniwersytet im. Adama Mickiewicza w Poznaniu,

Wydział Matematyki i Informatyki

Recenzja pracy doktorskiej Aliny Wróblewskiej, pt. „Dependency Parser Trained on Automatically Induced Dependency Bank"

Ocena celu pracy

Celem pracy jest opracowanie nadzorowanej metodologii tworzenia parsera zależnościowego dla języka nieposiadającego odpowiedniego banku drzew. Parser zależnościowy to narzędzie, które dla każdego zdania wejściowego tekstu podaje jego reprezentację w postaci drzewa rozpiętego na wierzchołkach odpowiadających wyrazom zdania. Dla każdej gałęzi drzewa parser wskazuje, który z dwóch wierzchołków łączonych przez krawędź pełni rolę nadrzędną oraz określa typ relacji zachodzącej pomiędzy tymi wierzchołkami.

Parser zależnościowy ma szereg zalet w stosunku do klasycznego parsera gramatyk składnikowych. Choć w podstawowej wersji ma teoretycznie tą samą złożoność pesymistyczną ($O(n^3)$), to w praktycznych implementacjach jest znacząco szybszy. Parsing zależnościowy umożliwia uzyskiwanie drzew projekcyjnych (o niekrzyżujących się gałęziach) dla nieciągłych konstrukcji zdaniowych (np. dla pytania „Która jest godzina?"). Jest to cecha szczególnie pożądana dla języków o luźnym szyku (takich jak język polski). Przy tym moc generatywna gramatyki zależnościowej jest taka sama jak moc gramatyki składnikowej (jeśli w tej drugiej wskazuje się element główny każdego składnika).

Podstawową trudnością w realizacji parsera zależnościowego jest automatyczny wybór najlepszego drzewa zależnościowego spośród wielu hipotez zwracanych przez parser. Na przykład parser zależnościowy UTT autorstwa Tomasza Obrębskiego[1] zwraca 4288 drzew zależnościowych dla zdania:

Rodzice Marty i Piotra byli przekonani, że gdyby przewiezienie Zosi z naszego domu do szpitala wojewódzkiego samochodem brata trwało krócej niż dwadzieścia minut, zapewne zdążyłaby ona jeszcze wynająć swój wymarzony pokój, który był nie tylko ładniejszy, ale i spokojniejszy niż pokój Marii.

Przykład ten wskazuje, że rozpatrywanie wszystkich hipotez zależnościowej struktury zdania jest nieracjonalne. Oczekuje się, że dobrze skonstruowany parser zależnościowy wskaże jedną interpretację zgodną z ludzką intuicją. Przykładowo dla zdania „Idę z bratem do domu" oczekuje się, że parser zależnościowy wskaże, iż bezpośrednim nadrzędnikiem przyimka „do" jest czasownik „idę", a nie – będący w mniejszej odległości w zdaniu – rzeczownik „bratem", nawet jeśli reguły zezwalają na relację nadrzędności zarówno między czasownikiem i przyimkiem, jak i rzeczownikiem i przyimkiem.

---

[1] Tomasz Obrębski, praca dr pt. „Automatyczna analiza składniowa języka polskiego z wykorzystaniem gramatyki zależnościowej"

Wspomniany wyżej parser UTT nie realizuje tego postulatu. Dla zdania wejściowego zwraca zestaw potencjalnych drzew zależnościowych, nie szeregując ich adekwatności do znaczenia zdania.

Autorka recenzowanej pracy postawiła sobie za zadanie opracowanie parsera zależnościowego języka polskiego, który dla danego zdania wyróżnia jedno drzewo zależności. Jest to pierwsze tego typu rozwiązanie dla naszego języka i jako takie, zasługuje na pełne uznanie.

Ocena metodologii

Autorka postanowiła wykorzystać istniejące algorytmy przetwarzania zależnościowego na poziomie „state-of-art", a mianowicie parsery: *MaltParser* i *Mate*, a następnie porównać skuteczność obu narzędzi. Oba parsery trenowane są na banku drzew, czyli zestawie zdań poprawnie anotowanych zgodnie z pewną gramatyką zależnościową. Dla języka polskiego natrafiono na kluczową trudność: brak zweryfikowanego banku drzew zależnościowych.

Autorka zaproponowała dwie metody pokonania tej trudności. Pierwsza z nich polega na wykorzystaniu istniejącego banku drzew składnikowych (o nazwie *Składnica*) po uprzednim automatycznym przekonwertowaniu drzew do reprezentacji zależnościowej.

Druga metoda polega na wykorzystaniu istniejącego banku drzew dla innego języka (w tym wypadku angielskiego) i automatycznym przemapowaniu go – za pomocą autorskiej technologii rzutowania ważonego – na bank drzew języka polskiego.

Obie metody (a szczególnie druga z nich) budzą moje wątpliwości, które szczegółowo omówię, przy ocenie poszczególnych fragmentów pracy doktorskiej.

Ocena rezultatów pracy

Charakteryzując wyniki swoich badań, Autorka wymienia następujące dokonania:

1) Schemat anotacji zależności dla języka polskiego
2) Stworzenie banku drzew zależnościowych poprzez konwersję drzew składnikowych
3) Metoda rzutowania ważonego dla uzyskania banku drzew zależnościowych
4) Eksperymenty w trenowaniu i ewaluacji parserów opartych na bankach drzew zależnościowych
5) Upublicznienie stworzonych banków drzew oraz wytrenowanych modeli zależnościowych

Za główne osiągnięcie swojej pracy autorka uważa metodę rzutowania ważonego.

Moja ocena wyników pracy jest nieco inna. Za najbardziej cenne uważam powstanie parsera zależnościowego, który jest dostępny on-line, o czym autorka w swojej pracy nie wspomina. (Jest o tym natomiast mowa w jednej prac, w której Doktorantka jest współautorką. Być może przyczyną tego stanu rzecz jest fakt, że parser nie jest jeszcze odpowiednio przetestowany i nie zawsze zwraca odpowiedź.)

Wartościowym wynikiem wydaje mi się publicznie dostępny bank drzew zależnościowych. Zakładam, że takowy istnieje i można do niego dotrzeć, pomimo że mnie nie udało się go odnaleźć. (Na stronie zasobów IPI PAN znalazłem bank drzew o nazwie „bushes", ale nie mam do niego dostępu.)

Z naukowego punktu widzenia najwyżej cenię sobie schemat anotacji zależności dla języka polskiego. Mam natomiast zastrzeżenia do autorskich metod pozyskania banku drzew.

Ocena strony formalnej pracy

Praca napisana jest w języku angielskim. Nie dostrzegłem istotnych błędów gramatycznych ani stylistycznych. Zdania budowane są poprawnie, wskazują na dobre opanowanie języka obcego przez Autorkę. Wybór języka angielskiego nie jest jednak dla mnie oczywisty. Uważam, że praca jest interesująca głównie dla badaczy języka polskiego i mogła być napisana w języku polskim.

Układ pracy uważam za dobry. Treści są odpowiednio podzielone.

Zaskakuje mnie nietypowy układ treści pomiędzy wkładem autorskim i pracami innych autorów powiązanymi tematycznie. Otóż w każdym rozdziale podrozdział „Related Work" znajduje się na końcu. Mniemam, że Autorka pragnęła w ten sposób ułatwić lekturę prac powiązanych tematycznie poprzez poprzedzenie ich własnymi przykładami. Pomysł ten uważam za chybiony. Moim zdaniem celem odwoływania się do prac związanych z tematem jest wykazanie na ich tle elementów nowatorskich we własnym rozwiązaniu. Trudno jest dyskutować o nowatorstwie przedstawianych własnych rozwiązań, gdy poprzedzających je wyników jeszcze nie omówiono.

Nie mam zastrzeżeń do redakcyjnej strony pracy. Przestrzegane są dobre zasady stylu naukowego. Praca jest dobrze złożona, robi pozytywne wrażenie estetyczne.

Pewne wątpliwości budzi sposób definiowania pojęć, co wskażę na przykładach.

W wielu miejscach pracy kropka występuje zamiast przecinka.

Podsumowując: stronę formalną pracy uważam za dobrą, choć nie perfekcyjną.

Ocena wartości merytorycznej poszczególnych rozdziałów pracy

Wartość merytoryczną pracy ocenię analizując poszczególne jej rozdziały.

Ocena Wstępu

Wstęp wyjaśnia założenia pracy, wskazuje wkład autorski, omawia strukturę pracy. Oceniam wstęp jako klarowny i dobrze zorganizowany.

Zabrakło mi we wstępie referencji do serwisu webowego z autorskim parserem zależnościowym oraz do utworzonego banku drzew. Moim zdaniem z większym zainteresowaniem czyta się pracę, gdy można od razu dostrzec, że przyniosła ona praktyczne rezultaty.

Ocena Rozdziału 2.

Celem rozdziału 2. jest wprowadzenie pojęć związanych z teorią zależności. Autorka przedstawia krótką historię rozwoju teorii (podrozdział 2.1), następnie przedstawia rozwój teorii zależności w Polsce (podrozdział 2.2), a w dalszej kolejności wprowadza definicje kluczowych pojęć takich jak drzewo zależności.

W tej części widzę pewną niespójność w sposobie definiowania pojęć. Otóż są one wprowadzane na trzy różne sposoby:

1) Definicje nieformalne

W ten sposób wprowadzone jest pojęcie struktury syntaktycznej (Definicja 2.1)

2) Definicje formalne – z użyciem aparatu matematycznego

W ten sposób wprowadzone jest pojęcie drzewa zależnościowego (Definicja 2.2)

3) Definicje wplecione w tekst – niewyróżnione

W ten mniej klarowny sposób wprowadzone jest np. pojęcie relacji zależności (str. 9).

Doktorantka niezbyt pewnie czuje się w definicjach formalnych, czego przykładem jest Definicja 2.2. Punkt 4) definicji wynika bezpośrednio z punktu 1), jest więc redundantny.

W podrozdziałach 2.4. i 2.5. Autorka opisuje współczesne metodologie parsingu zależnościowego z wykorzystaniem uczenia pod nadzorem, a mianowicie *Transition-based parsing* i *Graph-based parsing*. Czyni to w sposób zrozumiały i formalnie poprawny.

Ocena Rozdziału 3.

W Rozdziale 3. Autorka prezentuje autorski schemat anotacji zależności dla języka polskiego.

Zaskoczony jestem kolejnością wprowadzania zależności w tym opisie. Można się było spodziewać, że najpierw zostaną podane zależności klasyczne, często występujące w tekstach polskich (podmiot, dopełnienie), tymczasem opis zaczyna się od zależności rzadszych (pierwszą wymienioną zależnością jest przymiotnikowe dopełnienie czasownika). Nie byłoby zapewne nic w tym złego (praca doktorska nie musi spełniać postulatów dydaktycznych), gdyby nie powodowało to sytuacji, że w przykładach obrazujących poszczególne typy zależności występują relacje wcześniej niewprowadzone.

Nie wszystkie typy zależności są zgodne z moją intuicją. W zdaniach: „Pamiętasz, że to ja stawiam („Przykład 3.9) oraz zapewne w zdaniu „Zapytał, czy to ja stawiam" spójniki („że", „czy") są zależne od czasowników w zdaniu podrzędnym („stawiam"), podczas gdy to czasowniki w zdaniu nadrzędnym („pamiętasz", „zapytał") determinują typ spójnika. W proponowanym schemacie w zdaniach podrzędnych spójnik zależny jest od czasownika, podczas gdy forma czasownika często zależy od spójnika (np. po spójniku „gdyby" czasownik musi wystąpić w czasie przeszłym). (Autorzy nie są odosobnieni w takiej interpretacji zależności – podobnie spójniki traktowane są np. w gramatyce Stanford – jestem jednak ciekawy motywacji dla takiego rozwiązania.)

Podrozdział 3.3. poświęcony jest innym pracom związanym ze schematami opisu zależności. Jak wspomniałem, umieszczenie tych informacji na końcu rozdziału uważam za zaskakujące, a w tym konkretnym przypadku prowadzi do istotnej luki w pracy. Autorka omawia prace, na których wzorowała swój schemat, przyznając, że najbliższym rozwiązaniem jest schemat PDT utworzony dla języka czeskiego w roku 1998. Nie mogę zrozumieć braku odniesienia do polskiego schematu anatocji zależnościowej autorstwa Tomasza Obrębskiego, który powstał później – w latach 2002 – 2003. Gdyby analiza tego schematu znalazła się na początku rozdziału, to Autorka musiałaby się odnieść do tego rozwiązania, wskazując przyczyny, dla których uważa, że warto było stworzyć nowy schemat.

Ocena Rozdziału 4.

Rozdział 4. poświęcony jest stworzeniu banku drzew zależnościowych poprzez konwersję z banku drzew składnikowych o nazwie *Składnica*.

Proces ten odbywa się w dwóch etapach. W pierwszym z nich buduje się drzewo zależności pomiędzy wyrazami zdania na podstawie powiązań składników w odpowiadającym drzewie składnikowym. Etap ten wyjaśniony jest klarownie na czytelnym przykładzie.

W drugim etapie automatycznie oznacza się każdą zależność jej typem. Jedną z trudności w tym etapie jest ustalenie typu zależności między czasownikiem a rzeczownikiem niebędącym podmiotem (w schemacie anotacji wyróżnione są trzy typy zależności, odpowiadające w przybliżeniu dopełnieniu bliższemu, dalszemu lub orzecznikowi). Autorka twierdzi, że w czasie badań nie istniała lista polskich czasowników przechodnich, więc musiała stworzyć ją (zespołowo) we własnym zakresie. Trudno jest zgodzić się z tym stwierdzeniem. Wyczerpującą listę walencji czasowników polskich można odnaleźć w „Słowniku syntaktyczno-generatywnym czasowników polskich" autorstwa Kazimierza Polańskiego wydanym w roku 1980. Ponownie nie sprawdziła się więc metoda analizy prac związanych „po fakcie".

(W Dodatku A. znajduje się pełna lista reguł oznaczania typów zależności. Nie znalazłem jednak w Rozdziale 4. referencji do Dodatku A.)

Inną trudnością w konwersji reprezentacji zdań jest ustalenie jednego elementu głównego (ang: *head*) składnika, gdy w reprezentacji składnikowej takich elementów jest zero lub kilka. Pierwsza sytuacja nie jest w pracy rozwiązana (sądzę, że takie przypadki były odrzucane), natomiast w sytuacji występowania kliku elementów głównych stosowane są zasady wyboru spójne z przyjętym schematem anotacji zależnościowej.

Wydaje się, że podstawowym celem stworzenia gramatyki zależnościowej dla języka polskiego jest możliwość parsowania zdań o nieciągłych składnikach. Przykładem takiego zdania podanym przez autorkę jest zdanie „Wniosków jest raptem kilka". W banku drzew składnikowych powiązanie pomiędzy wyrazami „wniosków" i „kilka" nie istnieje (ze względu na wymóg niekrzyżowania się gałęzi). Aby takie relacje zostały oznaczone w banku drzew zależnościowych, należało ręcznie wykonać reorganizację reprezentacji zależnościowej. W rezultacie reorganizacji otrzymano pewną liczbę konstrukcji nieciągłych, dokładnie 0,15% całości.

Moim, zdaniem skoro i tak zdecydowano się na ręczną ingerencję w proces tworzenia banku drzew, to należało w tej metodologii pójść jeszcze dalej i zapewnić, aby w banku drzew zależności znalazło się więcej przykładów konstrukcji nieciągłych – szczególnie z zaimkami zwrotnymi (np. „On sobie krzywdy nie da zrobić" „On się lubi spóźniać") i pytajnymi zaimkami przymiotnymi („Jaki był powód tej reakcji?"). Wytrenowany na takim korpusie parser opisywałby większy podzbiór polszczyzny. Można by przy tym wykazać wyższość tworzonego parsera zależnościowego wytrenowanego na takim korpusie nad istniejącym parserem składnikowym.

W efekcie konwersji i ręcznej korekty uzyskano 8227 drzew, z których ok. 90% przeznaczono do trenowania parsera, a pozostałe do walidacji wyników. Walidację wyników przeprowadzono dwoma metodami: bardziej restrykcyjną (LAS), która sprawdza poprawne przypisanie elementów nadrzędnych i typ zależności oraz mniej restrykcyjną(UAS), która sprawdza tylko poprawność przypisania elementów nadrzędnych.

Bank drzew stanowił bazę do wytrenowania dwóch typów parserów opisanych w rozdziale 2., a mianowicie *MaltParser* i *Mate*. Nieco lepsze wyniki uzyskano dla parsera *Mate*, który na tekstach walidujących uzyskiwał ponad 90%-procentową skuteczność mierzoną metryką UAS i ponad 80%-skuteczność w metryce LAS. Jak można się było jednak spodziewać, wyniki te były wyraźnie (o ponad 10%) niższe dla próbki zdań niepochodzącej z korpusu drzew konwertowanych ze *Składnicy*.

Nie można tych wyników porównać z innymi rozwiązaniami dla języka polskiego, gdyż analogicznej ewaluacji nie dokonywano wcześniej. Nie podano porównania z wynikami uzyskanymi dla innych języków. Sprawdziłem, że wyniki uzyskane przez Autorkę nie odbiegają znacząco od tych uzyskanych dla innych języków. Można więc uznać eksperyment za zakończony sukcesem.

Ocena Rozdziału 5.

Celem Rozdziału 5. jest zaprezentowanie metody rzutowania wagowego w celu przekonwertowania banku drzew dla języka angielskiego na bank zależnościowy dla języka polskiego. Zastosowana metodyka badań budzi szereg moich wątpliwości.

Na początku rozdziału postawiono tezę, że zdanie w jednym języku oraz jego tłumaczenie mają skorelowane struktury składniowe. Jest to teza w ogólności ryzykowna – szczególnie dla pary języków pochodzących z różnych rodzin, jak język polski i angielski. Moje doświadczenia związane z tłumaczeniem automatycznym pomiędzy tymi dwoma językami wskazują na spore różnice syntaktyczne między nimi (por. Jassem K., 2002, Semantic Classification of Adjectives on the Basis of their Syntactic Features in Polish and English, Machine Translation Vol 17" Springer, str. 19-41).

Myślę, że teza dałaby się obronić na przykład dla dwóch języków słowiańskich. Jest więc dla mnie niezrozumiałe, dlaczego Autorka przeprowadza mapowanie z języka angielskiego, a nie czeskiego, dla którego istnieją zweryfikowane banki drzew zależnościowych.

Nie wiem, dlaczego do rzutowania wybrano bank drzew powstały po konwersji banku drzew składnikowych na bank drzew zależnościowych. W pracy wybór motywuje się faktem, że parser XLE, z którego uzyskano bank drzew, jest jednym z najlepszych dla języka angielskiego, ale tego faktu nie potwierdza się żadnymi porównaniami. W szczególności nie przeanalizowano jakości banku drzew powstałych po konwersji z reprezentacji składnikowej. A przecież eksperyment opisany w Rozdziale 3. wskazuje, że jakość banku zależnościowego po automatycznej konwersji z drzew składnikowych znacząco się obniża.

W Podrozdziale 5.1 opisuje się metodę rzutowania ważonego, której celem jest uzyskanie reprezentacji zależnościowej dla zdania polskiego, w sytuacji gdy znany jest jego odpowiednik w języku angielskim oraz dana jest jego reprezentacja zależnościowa. W pierwszym kroku tej metody buduje się pełny graf dwudzielny, którego węzłami są wyrazy poszczególnych zdań, a krawędzie etykietowane są wagami określającymi prawdopodobieństwo odpowiedniości między wyrazami zdania. Wagi te konstruuje się autorską metodą opierającą się o wyniki automatycznego dopasowania wyrazów zdania. Zakłada się, że dane są wyniki dopasowania dokonane w obu kierunkach oraz wynik symetryzacji (połączenia) tych dopasowań metodą *gdfa (grow-diag-final-and)*. W metodzie autorskiej dla każdej pary wyrazów określa się miarę będąca liczbą całkowitą od 0 do 3, której wartość mówi o tym, w ilu z tych trzech dopasowań para wystąpiła.

Nie potrafię zrozumieć celowości tej strategii. Skoro symetryzacja metodą *gdfa* służy właśnie do określenia najbardziej prawdopodobnego dopasowania, to po co konstruować kolejną miarę, która nie korzysta z nowych parametrów?

Efektem działań opisanych w Podrozdziale 5.1. jest pewien graf zależności. Celem algorytmów opisanych w Podrozdziale 5.2 jest wybranie z grafu drzewa zależności rozpiętego na wierzchołkach – wyrazach zdania polskiego. Stosuje się do tego algorytm oparty na metodzie EM (Expectation –

Maximization). W efekcie otrzymuje się drzewo zależności rozpięte na wyrazach polskiego zdania, ale wciąż oznaczone etykietami zależności w języku angielskim. Aby uzyskać etykiety zgodne z opracowanym schematem określono reguły konwersji do nich tagów angielskich. Wykonano to za pomocą ręcznie utworzonych reguł konwersji i korekty. W stosunku do analogicznego zadania w eksperymencie opisanym w Rozdziale 4. (oznaczanie zależności w drzewach indukowanych ze *Składnicy*) wykorzystano nowe źródło – słownik walencyjny *Walenty*. Ponownie jednak nie skorzystano z klasycznego źródła (słownika Polańskiego).

W Podrozdziale 5.4. opisano przebieg eksperymentu. Polegał on na zebraniu dużych korpusów dwujęzycznych, dopasowaniu ich na poziomie zdań i wyrazów za pomocą otwarto-źródłowych narzędzi oraz uruchomieniu na tych danych algorytmów opisanych w Podrozdziałach 5.1.- 5.3. Podrozdział ten zawiera również opis założeń gramatyki LFG, który nie jest częścią eksperymentu i powinien znaleźć się w innej części pracy.

W Podrozdziale 5.5. opisano przebieg trenowania parsera *MaltParser* oraz *Mate* na danych pozyskanych metodą rzutowania ważonego. Oceniono uzyskane rezultaty. Wyniki są bardzo zbliżone do tych uzyskanych z konwersji *Składnicy*.

Porównując metodologie prezentowane w Rozdziałach 4. i 5. Doktorantka stwierdza, że obie dają podobne, obiecujące rezultaty.

Podsumowanie recenzji

Zgadzam się z tezą przedstawioną w pracy, że wyniki eksperymentów opisanych w pracy doktorskiej są obiecujące. Nie odbiegają one znacząco od rezultatów uzyskanych dla języków o bogatszych zasobach i dłuższej historii istnienia parserów zależnościowych. Wykonując pracę pionierską dla języka polskiego nie ustrzeżono się kilku potknięć. Wykonano jednak postawione zadania: opracowano schemat anotacji zależnościowej i pozytywnie zweryfikowano dwie nowatorskie metody uzyskania parsera dla języka nieposiadającego banku drzew zależnościowych. Wytworzone narzędzie jest dostępne publicznie, co uzupełnia istotną lukę w polskiej lingwistyce komputerowej. Parser umożliwi stworzenie większego banku drzew zależnościowych, co w efekcie powinno przynieść dalszą poprawę skuteczności narzędzi analizy składniowej.

Stwierdzam, że recenzowana praca ma istotne znaczenie dla rozwoju dziedziny przetwarzania języka naturalnego w Polsce. Uważam, że rozprawa mgr Anny Wróblewskiej spełnia wymagania stawiane pracom doktorskim.

Krzysztof Jassem