dr hab. Adam Przepiórkowski, prof. IPI PAN
Instytut Podstaw Informatyki PAN
ul. Jana Kazimierza 5
01-248 Warszawa

Warszawa, 16 maja 2014

Recenzja rozprawy doktorskiej
p. mgr. Pawła Skórzewskiego

Przedmiotem recenzji jest rozprawa doktorska mgr. Pawła Skórzewskiego przedłożona Radzie Wydziału Matematyki i Informatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu.

1 Krótka charakterystyka pracy

Opiniowana praca nosi tytuł *Wydajne algorytmy parsowania dla języków o szyku swobodnym*, liczy 111 stron i składa się ze spisu treści itp., 7 rozdziałów, 2 dodatków i bibliografii. Rozdziały 1 i 2 zawierają ogólny wstęp do pracy i wyjaśnienie podstawowych pojęć, rozdziały 3 i 5 mają charakter przeglądowy, rozdziały 4 i 6 stanowią autorską część pracy, zaś rozdział 7 to krótkie podsumowanie pracy.

Rozdział 3 zawiera przegląd wybranych formalizmów opisu języków o swobodnym szyku wyrazów, w tym formalizmów FROG i TgBG opisanych w pracy doktorskiej Promotora pomocniczego pracy (Graliński, 2007). Autorski rozdział 4 przedstawia PTgBG, czyli probabilistyczne rozszerzenie drugiego z tych dwóch formalizmów. Następnie rozdział 5 stanowi przegląd algorytmów parsowania gramatyk probabilistycznych (i, ogólniej, gramatyk z wagami), zaś autorski rozdział 6 prezentuje wykorzystanie i optymalizację parsera Gobio w systemie PSI-Toolkit (stworzonym w ramach grantu badawczego realizowanego pod kierunkiem Promotora pracy), a także metodę przypisywania prawdopodobieństw regułom tego parsera.

2 Ocena pracy

Niniejszy punkt recenzji zawiera ogólną ocenę pracy pod względem formalnym, w tym redakcyjnym (§2.1), uwagi do poszczególnych rozdziałów (§2.2), oraz podsumowującą ocenę merytoryczną (§2.3).

2.1 Aspekty formalne

W skali makro praca jest spójna do rozdziału 5 włącznie, niespójność wprowadza jednak rozdział 6. Po zaproponowaniu probabilistycznego rozszerzenia pewnego typu gramatyk w rozdziale 4 i przeglądzie wydajnych algorytmów parsowania gramatyk probabilistycznych w rozdziale 5, w kolejnym rozdziale należało się spodziewać propozycji wydajnego algorytmu parsowania omawianego typu gramatyk probabilistycznych. Zamiast tego w rozdziale 6 znajduje się opis eksperymentu polegającego na zastosowaniu prostej i znanej metody przypisywania prawdopodobieństw regułom gramatyki (zob. poniżej) oraz – przede wszystkim – opis niskopoziomowych optymalizacji komponentów systemu PSI-Toolkit. Rozdział ten ma charakter w zasadzie wyłącznie inżynierski.

W skali mikro praca jest jasno napisana i porządnie złożona – stosunkowo drobne problemy językowe i typograficzne wymieniam w dodatku do niniejszej recenzji.

2.2 Uwagi do rozdziałów

2.2.1 Rozdział 2

Rozdział 2 zawiera definicje podstawowych pojęć z zakresu teorii języków i gramatyk. Pewne niespójności pojawiają się w punkcie 2.3 dotyczącym drzew składniowych. Definicja 2.23 drzewa nie zabrania cykli i nie wymusza skończoności struktury. Niekonsekwentnie są też wprowadzone 2 różne sposoby zapisu drzew: definicja 2.23 wprowadza notację z nawiasami kwadratowymi i dwukropkami, zaś następny akapit *zakłada* notację z nawiasami okrągłymi i ponownie wprowadza notację z nawiasami kwadratowymi jako uproszczenie tej z okrągłymi. Definicja 2.24 stosuje w istotny sposób niewyjaśniony nigdzie symbol $\hat{T}$ (zapewne chodzi o przestrzeń drzew nad zbiorem węzłów i etykiet). W tej samej definicji etykiety liści są raz definiowane z indeksami, a raz bez.

2.2.2 Rozdział 3

Przeglądowy rozdział 3 zaczyna się od wymienienia rodzajów nieciągłości w języku polskim, przy czym nie jest jasne, czy lista na s.22 miała być wyczerpująca (na pewno nie jest), ani dlaczego akurat takie a nie inne rodzaje nieciągłości zostały tu wymienione (nie ma na przykład ilustracji oddzielenia zdania podrzędnego od rządzącego nim czasownika i wielu innych rodzajów nieciągłości).

W przeglądzie formalizmów opartych na gramatykach bezkontekstowych (punkt 3.1) przydałaby się ocena przydatności tych formalizmów do opisu języka polskiego. Tytuł kolejnego punktu, 3.2 „Rozszerzenia gramatyk bezkontekstowych..." jest trochę

niefortunny, gdyż punkt ten omawia rozszerzenia notacji w bankach drzew, a nie rozszerzenia gramatyk. Kolejny punkt, opisujący „Grammatical Framework", jest zbyt skrótowy – osoba nie znająca wcześniej tego formalizmu nie wyciągnie wiele z tego opisu. Nie jest zresztą jasne, czemu ten punkt się tu znalazł, skoro nie ma w nim mowy o nieciągłościach, a i dalej w pracy nie ma żadnych nawiązań do tego formalizmu. Zabrakło natomiast w tym rozdziale omówienia podejścia do nieciągłości w Head-driven Phrase Structure Grammar (Pollard i Sag, 1994) opisanego np. w monografii Kathol 2000 i zastosowanego do języków słowiańskich np. w pracach Penn 1999a,b.

Zaskakujące jest stwierdzenie na stronie 35, że gramatyki zależnościowe nie spełniają postulatu przydatności w tłumaczeniu maszynowym, poparte bez dyskusji pracą z 1984 roku, czyli sprzed czasów nowoczesnego tłumaczenia maszynowego opartego na metodach probabilistycznych. Na tej samej stronie znajduje się stwierdzenie, że gramatyki typu TgBG są formalizmem średniego poziomu, bez wyjaśnienia o jaką skalę chodzi (czy po prostu o hierarchię Chomskiego?).

2.2.3 Rozdział 4

Rozdział 4 jest głównym rozdziałem autorskim pracy (por. §2.2.5 poniżej). Przedstawia probabilistyczną wersję gramatyk TgBG, nazwaną PTgBG, i pokazuje pewne zależności między PTgBG i PCFG (czyli probabilistycznymi gramatykami bezkontekstowymi).

Samo rozszerzenie TgBG do PTgBG jest natychmiastowe i w pełni analogiczne do rozszerzenia CFG do PCFG. Niektóre z rozpatrywanych zależności pomiędzy PTgBG i PCFG (np. te oparte na kontrprzykładzie 4.3 i Twierdzenie 4.4[1]) wynikają wprost z odpowiednich zależności pomiędzy TgBG i CFG wykazanych w pracy Graliński 2007, inne są naturalnym i w miarę prostym do udowodnienia (choć wymagającym rozpatrzenia różnych przypadków w indukcji strukturalnej) rozszerzeniem zależności tam dyskutowanych (np. Twierdzenie 4.1 i natychmiast wynikające z niego Twierdzenie 4.3).

Nowym wynikiem, stanowiącym istotny wkład Autora w tematykę, jest natomiast Twierdzenie 4.5 mówiące, że dla pewnego podzbioru gramatyk PTgBG (bez operacji wstawiania) można skonstruować równoważne – w sensie prawdopodobieństwa przypisywanego zdaniom – gramatyki PCFG. W tym kontekście zabrakło jednak twierdzenia, że to ograniczenie jest istotne, tj. że takiej konstrukcji nie można przeprowadzić w ogólności dla gramatyk PTgBG.

Inne narzucające się pytanie, które w niniejszej pracy pozostało bez odpowiedzi,

[1] Przy twierdzeniu tym zabrakło zresztą dowodu czy choćby odwołania do dowodu analogicznego twierdzenia przedstawionego w pracy Gralińskiego (2007).