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).

dotyczy języka drzew. Kontrprzykład 4.3 pokazuje, że istnieje gramatyka (P)TgBG generująca zbiór drzew, którego nie generuje żadna gramatyka (P)CFG. Na podstawie tego wyciągnięty został słuszny wniosek, że nie dla każdej gramatyki PTgBG istnieje gramatyka PCFG przypisująca drzewom te same prawdopodobieństwa. Nasuwa się jednak następujące pytanie: *jeżeli* dla danej gramatyki PTgBG istnieje gramatyka PCFG o tym samym języku drzew, to czy istnieje dla niej też gramatyka PCFG o tych samych prawdopodobieństwach drzew?

2.2.4 Rozdział 5

Przeglądowy rozdział 5 omawia wybrane algorytmy parsowania gramatyk z wagami (przede wszystkim probabilistycznych). Dużo uwagi (9 stron na 17 stron tego rozdziału) poświęcono algorytmowi A*. Rozdział ten jest trochę niedopracowany – oprócz algorytmu CYK i A*, które zilustrowano przykładami, pozostałe podejścia omówione są zbyt skrótowo, by ich porównanie było znaczące. (Przykład 5.2 używa też symboli, które nie są zdefiniowane – β, a, b). Poważniejszym zarzutem jest jednak to, że rozdział ten wydaje się zupełnie niepotrzebny, skoro kolejny – autorski – rozdział nie prezentuje żadnego nowego algorytmu parsowania gramatyk z wagami.

2.2.5 Rozdział 6

Rozdział 6 wyraźnie odstaje od reszty pracy. Wydaje się bardziej niedopracowany od reszty pracy, o czym świadczą na przykład nieprzemyślane stwierdzenia typu „Trudności rosną zwłaszcza wtedy, gdy wszystkie wagi są dodatnie, ponieważ wówczas nie mogą być traktowane jako logarytmy prawdopodobieństwa, a zatem nie mogą być w prosty sposób przekształcone na wagi probabilistyczne" (s.75). Nie jest też jasno zaznaczony wkład autora w opisywane w tym rozdziale treści. Na przykład ze wstępu na s.76 nie wynika, w jakim stopniu parser Gobio, któremu duża część tego rozdziału została poświęcona, jest dziełem Autora pracy – następujące zdanie, zawierające odnośnik do pracy [23] firmowanej wyłącznie nazwiskiem Promotora, sugeruje, że w niewielkim: „Gobio pierwotnie opracowano jako głęboki parser dla języka niemieckiego w systemie tłumaczenia automatycznego Translatica [23]".

Związek tego rozdziału z resztą pracy jest wątpliwy. Na górze s.76 przedstawiony on został następująco:

- Gobio jest parserem gramatyk typu TgBG z wagami (ręcznie przypisanymi niektórym regułom),
- można regułom takiego parsera przypisać wagi interpretowane jako prawdopodobieństwa, czyniąc z niego parser typu PTgBG – to usprawiedliwia zajęcie się

tym parserem w tej pracy;

- dodatkowo pokazano wcześniej (ograniczoną) równoważność PTgBG i PCFG, co usprawiedliwia opis parserów PCFG w rozdziale 5.

Większość rozdziału poświęcona jest omówieniu optymalizacji różnych komponentów systemu PSI-Toolkit, w tym parsera Gobio. Optymalizacje te stosują znane techniki, mają charakter wyłącznie inżynierski, a nie naukowy, dlatego pomijam je w niniejszej recenzji. Z tematem pracy związany jest bezpośrednio jedynie punkt 6.2 omawiający metodę przypisania wag interpretowalnych jako prawdopodobieństwa regułom gramatyki TgBG z (ręcznie przydzielonymi) wagami. Metoda jest prosta i polega na sparsowaniu pewnego korpusu z wykorzystaniem gramatyki TgBG a następnie policzeniu, jak często były w wybranych drzewach stosowane poszczególne reguły i przypisaniu im odpowiednich prawdopodobieństw maksymalizujących prawdopodobieństwo sparsowanego korpusu (bez wygładzania itp.). Ocena tak skonstruowanej gramatyki PTgBG polega na sparsowaniu obydwoma parserami – a także parserem ze wszystkimi wagami równymi zero i parserem z losowymi wagami – zbioru 108 zdań.

Wyniki tej ewaluacji przedstawione są w tabeli na stronie 80 i sugerują, że tak wytrenowany parser spisuje się gorzej jako komponent systemu tłumaczenia automatycznego nie tylko od parsera z wagami przypisanymi ręcznie, ale też od parsera bez żadnych wag. Przy czym wynik ten nie jest pewny: nie została policzona statystyczna istotność wyników, nie jest też jasne, jak trudne jest zadanie oceniania, które z dwóch tłumaczeń jest lepsze (nie ma danych typu Inter-Rater Agreement). Niewiele zatem z tego punktu wynika i trudno się oprzeć wrażeniu, że został on w pracy umieszczony wyłącznie w celu usprawiedliwienia w niej obecności rozdziałów 5 i 6 (bez nich treść pracy kończyłaby się na stronie 57).

2.3 Podsumowanie

Autorski wkład pracy w naukę ogranicza się moim zdaniem do rozdziału 4 – naukowy poziom pracy byłby wyższy bez drugiego autorskiego rozdziału 6. Wkład ten polega na zdefiniowaniu probabilistycznego rozszerzenia gramatyk TgBG i udowodnieniu kilku twierdzeń o ich (ograniczonej) równoważności gramatykom PCFG – w większości analogicznych do twierdzeń w pracy Graliński 2007 dotyczących nieprobabilistycznych wersji tych gramatyk. Jedno z twierdzeń jest istotnie nowe, brakuje jednak odpowiedzi na parę narzucających się pytań dotyczących relacji pomiędzy PTgBG i PCFG.

Rozdział 6 ma natomiast istotne znaczenie inżynierskie i pokazuje metody opty-

malizacji komponentów modułowego systemu PSI-Toolkit po ich „wycięciu" z monolitycznego systemu Translatica.

3 Wnioski końcowe

Godne odnotowania są aspekty inżynierskie pracy wynikające z udziału Autora w ważnym dla polskiego środowiska lingwistyki informatycznej projekcie kierowanym przez Promotora niniejszej pracy. Z drugiej strony, z punktu widzenia wymogów odpowiedniej ustawy, która mówi o „oryginalnym rozwiązaniu problemu naukowego", praca niebezpiecznie balansuje na granicy spełnialności tych wymogów. Ponieważ jednak naukowy wkład rozprawy w dziedzinę nie jest zupełnie zaniedbywalny, a ustawa nie precyzuje znaczenia terminu „oryginalne rozwiązanie problemu naukowego", wnoszę o dopuszczenie mgr. Pawła Skórzewskiego do dalszych etapów przewodu doktorskiego.

Adam Przepiórkowski

Literatura

Graliński F., 2007, Formalizacja nieciągłości zdań przy zastosowaniu rozszerzonej gramatyki bezkontekstowej, Rozprawa doktorska, Uniwersytet im. Adama Mickiewicza, Poznań.

Kathol A., 2000, Linear Syntax, Oxford University Press, Oxford.

Penn G., 1999a, Linearization and *WH*-extraction in HPSG: Evidence from Serbo-Croatian, [w:] Slavic in Head-Driven Phrase Structure Grammar, red. R. D. Borsley A. Przepiórkowski, CSLI Publications, Stanford, CA, s. 149–182.

Penn G., 1999b, An RSRL formalization of Serbo-Croatian clitic placement, [w:] Tübingen Studies in Head-Driven Phrase Structure Grammar, red. V. Kordoni, Arbeitspapiere des Sonderforschungsbereichs 340, Bericht Nr. 132, Universität Tübingen, Tybinga, s. 177–197.

Pollard C., Sag I. A., 1994, Head-driven Phrase Structure Grammar, Chicago University Press / CSLI Publications, Chicago, IL.

Korekta

- s.22: którego określa: → który określa:
- s.26: języka pisanego.języka Format → języka pisanego. Format
- s.33, def.3.5: w ostatnich trzech podpunktach ostatnim argumentem delty powinna być strzałka, a nie 0
- s.34, def.3.6: składa się z kolejnych liczb naturalnych → składa się ze skończonego ciągu kolejnych liczb naturalnych
- s.33, def.3.7: o ile się nie mylę, symbol σ użyty w tej definicji nie został nigdzie zdefiniowany (chodzi o bazę?)
- od s.36: rozszerzenie jest czasami oznaczone jako *ext*, a czasami (np. na s.41) jako *xt*
- s.40, przykład 3.6: w zbiorze Q brakuje *lvp*
- s.43, def.4.2: then → to
- s.45, def.4.6: $q \in Q \rightarrow q \in Q_s$ (bo $P_s(q)$ jest zdefiniowane tylko dla $q \in Q_s$)
- s.46: co znaczy „=>="?
- cała praca: w przypisach często brakuje kończących kropek
- cała praca: frazy wprowadzone przez imiesłowy przysłówkowe często nie są oddzielone przecinkami od reszty zdania (np. na s.38, *ten rzeczownik używając operacji*)