Recenzja pracy doktorskiej Marcina Pietrasa "Syntaktyczna i semantyczna analiza danych tekstowych z wykorzystaniem modeli Markowa realizowanych sprzętowo"

Krzysztof Jassem

14 czerwca 2018

1 Wstęp

Celem niniejszej recenzji jest zweryfikowanie, czy rozprawa Marcina Pietrasa zatytułowana "Syntaktyczna i semantyczna analiza danych tekstowych z wykorzystaniem modeli Markowa realizowanych sprzętowo" spełnia wymagania stawiane pracom doktorskim.

W pierwszej części recenzji omówiony jest tytuł pracy oraz hipoteza badawcza stawiana przez Autora. W drugiej części oceniona zostaje wartość merytoryczna pracy – z podziałem na poszczególne rozdziały. Recenzję kończy podsumowanie.

2 Ocena tytułu i tezy pracy

Tytuł pracy nie jest w pełni adekwatny do jej treści. Uważam bowiem, że nie zrealizowano części tytułu: "semantyczna analiza". Autor wprowadza pojęcie analizy semantycznej w rozdziale 5, a tytuł podrozdziału 5.3 "Disambiguacja i dopasowywanie wyrażeń" sugeruje, że możemy w nim oczekiwać rozwiązania autorskiego, Nie potrafiłem jednak tego rozwiązania się doszukać (a może zrozumieć). Na podstawie rysunku 5.3. (który nie jest omówiony w rozprawie, a jego analiza, ze względu na mały rozmiar czcionki, możliwa jest wyłącznie w wersji elektronicznej) postawiłbym hipotezę, że celem Autora było ujednoznacznienie semantyczne, czyli wybór aktualnego znaczenia wyrazu spośród kilku możliwości opisanych w sieci WordNet. Hipotezy tej nie wspierają jednak sekcje 5.3.1 i 5.3.2 poświęcone odległości edycyjnej Levenshteina, niemającej związku z semantyką leksykalną. Z rysunku 5.1. można by z kolei wnioskować, że pod pojęciem analizy semantycznej Autor rozumie oznaczanie wyrazów ich funkcjami syntaktycznymi (podmiot, orzeczenie,...). Podsumowując, jeśli celem Autora była rzeczywiście analiza semantyczna danych tekstowych, to oczekiwałbym rozszerzenia rozdziału 5. o jasny opis rozwiązania autorskiego z przykładami danych wejściowych i wyjściowych oraz wyjaśnieniem powiększonych elementów rysunku 5.3.

Teza pracy postawiona jest następująco: stosując realizacje sprzętowe modeli Markowa można analizować teksty na różnych poziomach językowych z wysoką dokładnością. Jako potwierdzenie tej tezy Autor w zakończeniu (str. 127) podaje udowodniony w pracy lemat o bezpieczeństwie numerycznym ukrytych modeli Markowa.

Moim zdaniem, poprzez wspomniany lemat wykazano stabilność proponowanych algorytmów, a nie ich wysoką dokładność (czy też jakość), która w zagadnieniach związanych z przetwarzaniem

języka naturalnego jest z reguły ewaluowana za pomocą miar precyzji i pokrycia (lub ich agregacji w postaci tzw. *miary F*).

Na zakończenie rozdziału 4. podana jest tabela podająca "dokładność rozkładu zdania na części zdania" (w rzeczywistości tabela opisuje poprawność rozpoznania typów zależności między wyrazami). Nie zdefiniowano jednak stosowanego w tabeli pojęcia "dokładność" (czy jest to odpowiednik terminu "accuracy", "precision", czy "recall"), ani nie porównano osiągniętych wyników z rezultatami innych badań.

3 Ocena merytoryczna poszczególnych rozdziałów pracy

3.1 Wprowadzenie

Autor, zgodnie z wszelkimi "kanonami sztuki", stara się omówić zastosowanie terminów, które pojawiają się w temacie pracy: analiza syntaktyczna, analiza składniowa, realizacja sprzętowa algorytmu, modele Markowa. Według mnie bardzo pomocne dla dalszej lektury byłoby przynajmniej nieformalne wprowadzenie definicji tych pojęć – a szczególnie pojęcia *sprzętowej realizacji algorytmu.*

W części 1.1. podane są cele pracy. Cel nr 2 określony jest następująco: "Opracowanie metody rozkładu zdań na części mowy i części zdania". Zgodnie z moją wiedzą zdania nie rozkłada się na części mowy. W przetwarzaniu języka naturalnego stosuje się natomiast następujące terminy:

- *anotacja morfosyntaktyczna (ang. POS-tagging)* – przypisanie wyrazom znaczników morfosyntaktycznych (zwanych również częściami mowy),
- *analiza składniowa* – określenie struktury zdania (na przykład w formie drzewa składników frazowych).

Cel nr 3 sformułowany jest następująco: "Opracowanie metody opisu istotnych informacji w tekście na różnym poziomie semantycznym". Zakładając, że autorowi chodziło o liczbę mnogą ("różne poziomy semantyczne"), to cel ten nie został w pracy zrealizowany. W autorskiej aplikacji *HMM-Toolbox* mamy możliwość skorzystania z zasobów sieci WordNet i na tym analiza semantyczna się kończy – nie dostrzegłem wielorakości poziomów opisu semantycznego.

We Wprowadzeniu można zauważyć pewne problemy w operowaniu pojęciami z lingwistyki komputerowej: pojawiają się nienaturalne zbitki wyrazowe, np. "jakość struktury roli", czy też nienaturalnie sformułowane zdanie przykładowe, np. "Mamy biegły, a za nimi biegły mamy, z zawodu biegły..." (łatwo wymyślić naturalnie brzmiące polskie zdania zawierające homonimy, np. "Piękna gra trwała długo", "Janek woli tonę soli").

Podsumowując, aby móc pozytywnie ocenić Wprowadzenie, oczekiwałbym zdefiniowania (przynajmniej nieformalnego) terminów występujących w temacie pracy, przeformułowania celów badawczych oraz lepszego operowania językiem z dziedziny przetwarzania języka naturalnego.

3.2 Rozdział 2. Ukryte modele Markowa

W rozdziale 2. Autor omawia pojęcie ukrytych modeli Markowa, które stanowią matematyczne fundamenty rozprawy. Treści wprowadzane są w sposób formalnie poprawny, a przy tym uzupełniane są autorskimi przykładami pomagającymi w rozumieniu pojęć.

Chciałbym zwrócić uwagę na kilka drobnych nieścisłości.

Elementy zbioru stanów modelu oznaczane są naprzemiennie dużymi literami:

$$(S_1, S_2, ...)$$

i małymi literami:

$$(q_1, q_2, ...)$$

Podobna niespójność występuje przy oznaczaniu obserwacji (małe i duże litery):

$$o_i, O_i$$

Nazwa algorytmu "Forward-Backward nie może być odmieniana ("metody Forwarda-Backward", str. 14), gdyż nie pochodzi od nazwiska. Zastanawiam się, czy w całej pracy nie można by się pokusić o stosowanie polskiego odpowiednika ("algorytm propagacji wstecznej").

Nie jest dla mnie jasne, jak interpretować wzory podane na rysunku 2.2, np.

$$\lambda_1(O_1) + O_1$$

Rysunek 2.2. jest niezgodny z podanym opisem. Zgodnie z rysunkiem na pierwszym poziomie rozpoznawane są formy graficzne wyrazów a nie, jak się podaje, "części mowy", na drugim poziomie – części mowy, a nie – "części zdania", a na trzecim – części zdania (kategorie syntaktyczne), a nie – "role semantyczne". (Role semantyczne są podane poprawnie na rysunku 1.1; są to np. *predykat* i *argument*).

W rozdziale 2. prezentowane są zrzuty ekranów z aplikacji *HMM-Toolbox* (rysunki 2.3 i 2.4). Rysunki te (jak i inne zrzuty ekranów w pozostałych częściach pracy) są dla mnie nieczytelne: czcionka jest za mała, a liczba elementów interfejsu graficznego znacznie przekracza możliwości mojej percepcji. W moim odczuciu wszystkie zrzuty ekranu należałoby zmontować w taki sposób, aby na rysunkach znalazły się tylko elementy niezbędne dla zilustrowania danego fragmentu pracy.

Podsumowując, pozytywnie oceniam wartość merytoryczną rozdziału 2. Zwracam przy tym uwagę na drobne niedoskonałości natury technicznej.

3.3 Rozdział 3. Ekstrakcja informacji z HTML

Omówienie rozdziału 3. rozpocznę od przedyskutowania jego tytułu. Sądzę, że zgodnie z zasadami pisowni polskiej angielski akronim HTML należałoby poprzedzić polskim rzeczownikiem np. "...for-

matu HTML". Pod względem merytorycznym tytuł nie stoi w pełnej zgodności z treścią. Przez ekstrakcję informacji rozumie się automatyczny proces konwersji danych niestrukturyzowanych w dane strukturyzowane. Tymczasem celem działań opisywanych w rozdziale 3. jest **klasyfikacja** treści zapisanych w formacie HTML w celu wyodrębnienia tzw. "treści właściwej".

W stosunku do samej treści rozdziału mam również kilka zastrzeżeń.

Autor stawia zbyt ogólnikowe lub zbyt śmiałe tezy bez należytego ich poparcia, np. "Z badań wynika, że internet jest szeroko wykorzystywany w celu pozyskania informacji ..." (str. 28), " Jednakże skuteczność pozyskiwania informacji jest ograniczona brakiem narzędzi eksploracji danych..." (str. 28), "...analiza ekspercka wykazała, że większość portali informacyjnych stosuje pewien przyjęty schemat prezentacji artykułu..." (str. 33).

Niekiedy stosowany jest język potoczny: "...treści dodatkowej, którą użytkownik konsumuje...", a niekiedy sztuczny, a przez to niezrozumiały: "Tokeny reprezentują dane w pewien **zdyskretyzowany** sposób" (str. 32)., "model Markowa, który **natywnie** obsługuje..." (str. 35), "sekwencja tokenów poddana zostaje odpowiedniej **agregacji**" (str. 35), "obserwacji **agregowanej wyrażeniowo**" (str. 40).

Stosowane są skróty myślowe, np. "System ten jest całkowicie zależny od umiejętności specjalistów"(str. 31) (prawdopodobnie chodzi o to, że jakość działania systemu uzależniona jest od jakości reguł stworzonych przez specjalistów), "Reprezentacja danych z HTML" (prawdopodobnie chodzi tu o reprezentację danych wyekstrahowanych z formatu HTML).

Występują błędy składniowe: "Dodatkowym celem może być **ogólne** semantyczna kategoryzacja..." (str. 28), "...przeglądarki internetowe **muszę**..." (str. 30).

Zdarzają się powtórzenia, na przykład termin *wrapper* definiowany jest zarówno na stronie 30, jak i 31.

Niektóre terminy są nieprawidłowo definiowane, np. *Resource Description Framework* określony jest jako format reprezentowania informacji **prasowej**.

Zdarzają się błędy w stosowaniu terminologii, np. "interfejsu aplikacji DOM" (str. 32; DOM jest typem reprezentacji danych, a nie aplikacją), "wewnętrzne węzły drzewa odpowiadają znacznikom, które opisują kolejne gałęzie" (str. 33; gałęzie łączą węzły, a nie są przez nie opisywane).

Rysunek 3.1. jest nieczytelny (za mała czcionka) – sugerowałbym zamieszczenie tylko części jednej strony internetowej zamiast pełnych czterech.

Język sekcji 3.1.4 nie jest ścisły - znajdują się tu stwierdzenia nieudokumentowane, np "nadal istnieje kilka trendów...", niepoprawne semantycznie, np. "informacja zrozumiała dla każdej przeglądarki", czy nielogiczne, np. "...kluczowe jest, aby działania wstępnego parsera były identyczne bez względu na treść HTML...".

W niektórych fragmentach brakuje konsekwencji. W sekcji 3.2.1 wprowadza się trzy główne kategorie ciągów znaków, co ilustrowane jest rysunkiem 3.3. na którym występuje co najmniej 6 kolorów odpowiadających kategoriom.

W podrozdziale 3.3 z niejasnych względów stosowany jest czas przyszły ("...będą analizowane",

str. 37).

Podsumowując rozdział 3., stwierdzam, że liczne niedociągnięcia formalne nie pozwalają mi w sposób jednoznaczny ocenić jego wartości merytorycznej.

3.4 Rozdział 4. Syntaktyczna analiza treści

W rozdziale 4. występuje cały szereg błędów merytorycznych z dziedziny przetwarzania języka naturalnego. W niniejszej recenzji przedstawiam błędy wyłącznie z pierwszej strony tego rozdziału (str. 43):

W trzecim wierszu akapitu zamiennie stosowane są terminy "wyraz" i "wyrażenie" (w dalszej części rozdziału stosowany jest w tym samym znaczeniu jeszcze termin "słowo"). W piątym wierszu zawarte jest zdanie "każdy język posiada pewien zestaw reguł zwany gramatyką" (gramatyka jest sposobem opisu języka, a nie jego cechą). W miejscu definicji analizy syntaktycznej podaje się definicję anotacji morfosyntaktycznej (ang. *POS-tagging*). Podana jest nieprawdziwa teza, że "analiza syntaktyczna odnosi się do porównywania znaków".

Stosuje się niewprowadzony nigdzie wcześniej termin "parsowanie syntaktyczne". Dla jednego pojęcia stosuje się różne terminy: "bank drzew", "treebank".

W dalszych fragmentach tego rozdziału błędy z dziedziny przetwarzania języka naturalnego występują równie często. Na przykład opis algorytmu CKY w moim odczuciu wykazuje jego niezrozumienie.

Autor błędnie interpretuje stosowanie analizy składniowej opartej na gramatyce zależnościowej. Celem takiej analizy nie jest bowiem, jak zakłada Autor, podział zdania na składniki (taki cel przyświeca analizie opartej na strukturach frazowych), lecz znalezienie związków pomiędzy poszczególnymi wyrazami zdania. Dlatego też rysunek 4.13 nie przedstawia (jak to podpisano) rozkładu zdania na części mowy i zdania, lecz wyniki anotacji morfosyntaktycznej (w postaci części mowy, np. *verb*, *subst*) oraz zależności pomiędzy wyrazami zdania. Dla przykładu, strzałka etykietowana napisem *subj* oznacza, że wyraz *Sejm* pozostaje w zależności typu *subj* z wyrazem *przyjął*.

Z błędnej interpretacji gramatyki wynikają liczne błędy w sekcji 4.3.1, z których wymienię tutaj tylko te, które znalazły się w jednym podpunkcie ze str. 51:

- Jest: "przyimkowe uzupełnienie regulowane:", powinno być "dopełnienie przyimkowe zależne od (lub determinowane):"'
- Jest: "przymiotnikiem, np. zdolny zrobić", powinno być: "przymiotnikiem, np. zdolny do natychmiastowej reakcji" (podaję tutaj przykładową realizację zależności tego typu)
- Jest: przysłówkiem, np. "właśnie przez", powinno być: "przysłówkiem, np. niezależnie od siebie" (podaję tutaj przykładową realizację zależności tego typu).

Podsumowując rozdział 4, nie jestem w stanie ocenić jego wartości merytorycznej, bo nie potrafię stwierdzić, jaki był cel opisywanych w nim badań. Prawdopodobnie chodziło o zaprezen-

towanie nowego algorytmu wyznaczania zależności zachodzących między wyrazami w zdaniu z wykorzystaniem typów zależności opracowanych w pracy doktorskiej A. Wróblewskiej "Polish Dependency Parser Trained on an Automatically Induced Dependency Bank". Jeśli tak, to kluczowe dla oceny wartości pracy byłoby porównanie jakości działania rozwiązania autorskiego z parserem prezentowanym w pracy Wróblewskiej, a tego w rozprawie nie znalazłem.

3.5 Rozdział 5. Opis informacji na różnych poziomach semantycznych

Tytuł rozdziału 5 sugeruje, że omówiona jest w nim autorska metoda semantycznego opisu danych językowych. Treść rozdziału jednak nie odpowiada tym oczekiwaniom. W podrozdziale 5.1 omówione są modele semantyki, poczynając od reprezentacji logicznej (gramatyka Montague), przez model funkcjonalny oparty na rolach, a skończywszy na modelu leksykalnym.

Na rysunku 5.1. podano "graficzną reprezentację relacji semantycznych w zdaniu". Nie dostrzegam jednak na tym rysunku żadnych ról semantycznych (typu: *predykat*, *argumenty*, czy też: *agent*, *obiekt*, *instrument*), lecz oznaczenie funkcji syntaktycznych (*podmiot*, *orzeczenie*). Nie występują również relacje semantyczne (typu: *X jest w relacji znaczeniowej R z Y*), lecz zależności syntaktyczne, np. związek zgody.

Definicje podane po zdaniu: "Ogólnie w opisie semantycznym zdań wyróżnić należy:" (str. 79) są dla mnie niezrozumiałe. Przykładowo, w definicji: "słowa bazowe – są to związki semantyczne z formą predykat-argument" *definiendum* – "słowo bazowe" – jest bytem innego typu niż *definiens* – "związek semantyczny".

Podrozdział 5.2 składa się tylko z jednego zdania.

Podrozdział 5.3 omawia odległość edycyjną Levenshteina i jego wagową modyfikację – bez podania związku tego pojęcia z omawianym tematem.

Podsumowując, stwierdzam, że treść rozdziału 5 nie realizuje założeń sugerowanych przez tytuł.

3.6 Rozdział 6. Bezpieczeństwo numeryczne ukrytych modeli Markowa o zredukowanej reprezentacji liczbowej

Zaprezentowany w rozdziale 6 lemat o długości sekwencji niebezpiecznej numerycznie, stanowi najprawdopodobniej najbardziej wartościowy rezultat recenzowanej rozprawy. Jest to wynik interesujący zarówno pod względem teoretycznym, jak i praktycznym. W przetwarzaniu języka naturalnego można by go stosować z powodzeniem w szeroko stosowanym modelowaniu języka, gdzie problem rzadkości danych (i w konsekwencji: niestabilności numerycznej) rozwiązywany jest za pomocą metod tzw. *wygładzania*.

3.7 Rozdział 7. Akceleracja algorytmów uczenia maszynowego

Rozdział 7 poświęcony jest metodom przyspieszania algorytmów metodami sprzętowymi. Rozdział ten przekonująco obrazuje zastosowanie lematu udowodnionego w rozdziale 6 w implementacji

komputerowej.

4 Podsumowanie recenzji

Recenzowaną rozprawę należy podzielić na dwie części. Pierwsza z nich (rozdziały od 3 do 5) całkowicie lub częściowo poświęcona jest zagadnieniom przetwarzania języka naturalnego. Część ta wymaga ponownego zredagowania. Druga część (rozdziały 2, 6 i 7) omawia interesujące wyniki zarówno teoretyczne, jak i praktyczne z zakresu bezpieczeństwa numerycznego. Sugeruję ponowne zweryfikowanie recenzowanej rozprawy po korekcie błędów popełnionych w pierwszej z tych części.