Dr hab. Małgorzata Marciniak
Instytut Podstaw Informatyki PAN
ul. Jana Kazimierza 5
01-248 Warszawa

Warszawa, 13.10.2022

RECENZJA
rozprawy doktorskiej mgr. Sebastiana Dziewięckiego pt.
"Polsko-angielska terminologia Wojsk Specjalnych: Metodologia tworzenia branżowych leksykograficznych baz danych wraz z modelowaniem aplikacji sieciowej"

1. Tematyka rozprawy

W pracy opisana jest koncepcja polsko-angielskiego słownika terminologii wojskowej wraz z implementacją. Proponowany słownik jest leksykalną bazą obrazującą powiązania semantyczne (relacje) pomiędzy jednostkami. Terminy opisane są definicjami, zdaniami ilustrującymi ich przykładowe użycie oraz relacjami pomiędzy pojęciami. Słownik został zaimplementowany i opracowana została treść kilkuset haseł wraz z przykładami użycia pochodzącymi z zebranego korpusu tekstów. Działająca aplikacja została poddana testom według opracowanego przez doktoranta schematu badawczego.

Przedstawiona praca ma charakter interdyscyplinarny, łączy wiedzę z dziedziny wojskowości, informatyki i leksykografii. Opracowany słownik ma charakter nowatorski, nie znam innego tego typu polsko-angielskiego słownika dziedzinowego, który zostałby wdrożony do użytku. Doceniam, że w pracy doktorant samodzielnie opracował aplikację wymagającą wykazania się umiejętnościami programistycznymi i, jak rozumiem, doprowadził ją do etapu wykorzystania przez żołnierzy.

2. Układ rozprawy

Praca składa się z dwóch części. Pierwsza zawiera opis teoretycznych podstaw konstruowania słowników terminologicznych. Zostały w niej przedstawione zagadnienia związane z opisem wiedzy dziedzinowej, sposób komunikacji między specjalistami, zasady tworzenia terminologii dziedzinowej i słowników ją zawierających. Jest też rozdział poświęcony wykorzystaniu korpusów dziedzinowych w procesie pozyskiwania terminów.

Druga część poświęcona jest opisowi opracowanego przez autora polsko-angielskiego słownika terminologii stosowanej w wojskach specjalnych. Przedstawiona jest specyfika komunikacji w środowisku międzynarodowych sił specjalnych oraz uzasadniona jest potrzeba opracowania słownika ułatwiającego szkolenie żołnierzy. Omówione są też informatyczne aspekty implementowanego słownika oraz testy użytkowe aplikacji wraz z wnioskami.

3. Uwagi do rozprawy

Przedstawiona praca zawiera pewne niedociągnięcia i nieścisłości. Część z nich wynika prawdopodobnie z faktu, że podjęta tematyka wymaga szerokiej wiedzy z wielu dziedzin, co pociąga za sobą konieczność biegłego posługiwania się terminologią specyficzną dla tych dziedzin. Uwagi zostały podzielone na dwie grupy, pierwsza zawiera nieścisłości o charakterze merytorycznym, druga zaś to uwagi dotyczące edycji tekstu i języka rozprawy. Wszystkie uwagi należy rozpatrywać

w kontekście wskazanych fragmentów pracy. Zostały umieszczone zgodnie z kolejnością wynikającą z lektury tekstu.

3.1. Uwagi merytoryczne

W rozdziale 2.2 autor stara się opisać przynależność do kategorii obiektu z wykorzystaniem aparatu matematycznego. Formalizacja ta zawiera błędy i nie odzwierciedla zamysłu autora. Dodam, że fragment ten wymaga od czytelnika domyślania się intencji autora. Po pierwsze, nie jest dobrą praktyką oznaczanie zarówno obiektu jak i pewnego współczynnika tą samą literą, w tym przypadku „O". Autor użył jej do określenia obiektu wzorcowego („O_W"). Wybrał również literkę „O" z dodatkowym indeksem „w_{Cj}" do oznaczenia współczynnika podobieństwa. Nie jest to błędne a jedynie utrudnia czytanie omawianego fragmentu. Następnie zdefiniował obiekty jako pary <cecha, wartość>, gdzie wartość jest owym współczynnikiem podobieństwa (w definicji zbioru O_x ostatni, błędny fragment „=1" jest prawdopodobnie wynikiem kopiowania definicji O_w). Przy tak zdefiniowanych obiektach, zbiór będący iloczynem $O_w \cap O_x$ zawiera jedynie pary <cecha, wartość>, dla których drugi element pary ma przypisaną wartość 1. Oznacza to, że iloczyn zawiera tylko te pary, dla których cecha istotna dla obiektu wzorcowego jest dokładnie taka sama jak cecha rozważanego obiektu. Przy tak zdefiniowanym podobieństwie i podanych współczynnikach w tabeli 1, obiekty „O_1" i „O_3" powinny mieć miarę podobieństwa D równą 0, bo odpowiednie iloczyny zbiorów par nie zawierają wspólnych elementów. Ponadto podane przykłady cech zawierają zarówno cechy, które mogą być określane na różny sposób (np., „kształt" może być „owalny", „okrągły", „sześcienny"...) jak i cechy o wartościach binarnych jak „kranik". W tabeli 1 na str. 47 obiekt O_1 ma wpisaną wartość 1 dla cechy „kraniki" (czy było to celowe?). Jest to cecha spoza zbioru cech obiektu wzorcowego, nie wiem jak jest liczona i dlaczego ma taką wartość. Czy obiekt O_4 nie ma określonej wartości cechy „materiał" (puste miejsce w tabeli 1)?

W pracy wzmianka o Słowosieci znajduje się na stronie 80. Następnie pojawia się jej szerszy opis w zaskakującym miejscu – na stronach 172-173 – w części pracy, w której autor koncentruje się na opisie implementacji własnego słownika. Autor błędnie wskazuje na różnice pomiędzy organizacją Słowosieci a Princeton Wordnet, twierdząc co następuje: „Podstawową jednostką organizacyjną w polskim serwisie jest wyraz"... „Z kolei rozwiązanie proponowane przez projekt princetoński opiera się właśnie na synsetach." Słowosieć bazuje na jednostkach leksykalnych (lemat + numer znaczenia), które łączone są w *synsety*. Na przykład „bazar" oznaczony etykietą 1 jest synonimem słowa „rynek" oznaczonego etykietą 3, oba te pojęcia należą do tego samego *synsetu*.

Na stronie 98 autor opisując program tokenizujący, błędnie przypisał mu funkcję usuwania znaków interpunkcyjnych. Nie robi się tego z wielu powodów. W zadaniu ekstrakcji terminologii znaki interpunkcyjne mogą być kluczowe przy wskazaniu granic fraz rzeczownikowych, tak jak w poniższych przykładach:

- Jan zobaczył samochód, Piotra i Marię.
- Jan zobaczył samochód Piotra i Marię.

Z pierwszego zdania wynika, że Jan zobaczył trzy obiekty, natomiast z drugiego – tylko dwa i przecinek jest tu istotny.

Na stronie 157 autor opisując korpus podaje informację, że „przy zestawieniu z listą nltk.corpus.words zostało odrzuconych 482663 tokenów, które nie występowały w korpusie". Niestety, opis ten nie umożliwia zrozumienia co autor zrobił i skąd usunął niemal pół miliona tokenów. Dalej na tej samej stronie autor pisze: „zanim kolekcja zostanie poddana oznaczeniu, należy oczyścić go ze znaków nie należących do alfabetu, jak również z cyfr." Dlaczego autor usuwa cyfry i czy usuwa wszystkie? Czy cyfry służące do oznaczania np. modeli sprzętu też są usuwane? Ponadto do czego odnosi się zaimek „go" w powyższym zdaniu?

3.2. Uwagi dotyczące formy i języka

Obecna wersja poprawionej pracy (pierwszą otrzymałam do oceny w lutym 2022) jest napisana w sposób komunikatywny, choć język rozprawy jest daleki od doskonałości, której oczekiwałabym od pracy złożonej na Wydziale Lingwistyki Stosowanej. Poniżej wymieniam błędy językowe, stylistyczne, edytorskie oraz niezręczności:

- str. 12, „wpływa na percepcje sygnałów" -> percepcję;
- str. 14, rysunek 1, w opisie powinno być wyjaśnione czym jest niebieskie koło w centrum rysunku;
- str. 22, „data początku jest mniejsza niż data od końca" nadmiarowe „od";
- str. 22, „przez co informacje mają mały rozmiar" -> dzięki czemu informacje można oszczędnie zakodować;
- str. 22, pisząc o „szybszym dostępie" należy napisać do czego porównujemy;
- str. 22, „czynnikiem standaryzującym je" do czego odnosi się zaimek „je"?
- str. 22, „Jednakże, teksty specjalistyczne są zbudowane w charakterystyczny sposób i z biegiem czasu formy ulegają zmianom. To powoduje, że coraz trudniej znaleźć je w tekstach ogólnych." dlaczego? ;
- str. 24, „i stanowi suma logiczna idiolektów";
- str. 30, „występuje znaczna ilość znaków językowych", wprawdzie RJP dopuściła wymienne stosowanie słów „liczba" i „ilość", jednak dla policzalnych rzeczowników zdecydowanie lepiej brzmi „liczba";
- str. 31, rysunek 4, czym różnią się strzałki narysowane linią przerywaną od tej narysowanej linią podwójną ciągłą;
- str. 33, w zdaniu zaczynającym się od „Na tym etapie..." niezbyt zręczna koordynacja;
- str. 35, „i ,";
- str. 38, „KomisjaTerminologiczna Morską przy" oraz „Terminologicznaprzy";
- str. 43, przypis 13, pojawia się tajemniczy zaimek „nas" choć do tej pory autor stosował liczbę pojedynczą;
- str. 45, „obiekty należące niej";
- str. 48, rysunek 8 wymaga komentarza;
- str. 49, „umieścić w semantyczne strukturze";
- str. 50, „głównie |";
- str. 59, rysunek 9, brak wyjaśnienia czym różnią się strzałki przerywane od ciągłych i podwójnych linii;
- str. 60, „Zaleca się również neutralność terminów oraz stosowania" -> nie stosowanie;
- str. 62, „zaczął nabierać coraz większych rozmiarów" rozwój może mieć tempo ale nie rozmiar;
- str. 62, punkty 1-4 nie są opisem dziedziny;
- str. 66, „algorytm Louvain" – należy podać referencję bibliograficzną;
- str. 68, „W projektach z sieciami semantycznymi" w projektach wykorzystujących lub bazujących na...;
- str. 71 „Dotyczy wymagań związanych z budowaniem haseł terminologicznych w normach.";
- str. 71 „Reprezentacje terminu";
- str. 74 „ od definicji treściwych";
- str. 75 „Implikacje takiego zabiegu są całkiem kłopotliwe i mogą wywoływać trudności w komunikacji." należy wyjaśnić na czym polegają trudności;
- str. 75 „przedstawia w swojej pracy propozycja";
- str. 75 co to jest „jednostka rozumienia"?

- str. 78 „łopatka + saper = saperka" lepiej napisać odwrotnie: saper + łopatka;
- str. 82 „przez co są dla danej grupy zawodowej" brakuje słowa np. „istotne";
- str. 88 „podstawową relacje generyczną";
- str. 93 „Zawartość korpusu ma różne pochodzenie i zależy to głównie od jego przeznaczenia korpusu.";
- 98 str. „poprzez zastosowanie współczynnika C-value" jest to tylko jedna z metod, więc należałoby dodać „na przykład";
- str. 99 „lematyzajcja" oraz „stosowane..";
- str. 125 „wojsko skupiało się na języku angielskim, który jest bardziej dostępny" włada nim dużo osób?
- str. 137 „jeżeli występuje będzie" -> „jeżeli występuje, będzie;
- str. 153 „Źródła papierowe mogą być zeskanowane i do postaci plików graficznych";
- str. 153 „może wpłynąć na zanieczyszczenie różnych wyników" bez „różnych";
- str. 154 „To nie dotyczy MongoDB" czym jest „To";
- str. 155 w zdaniu „ Do testów została wybrana..." następuje zmiana czasu „wydaje się być";
- str. 155 przypis 46: „Ma to wpływ na wszystkie czynności opierające się na korpusie";
- str. 159 „Jak wiadomo, w plikach pdf z rozpoznanym tekstem występują błędy różnej postaci." prawdopodobnie autorowi chodziło o proces pozyskiwania tekstów z plików pdf oraz o to, że jest to proces obarczony błędami;
- str. 161 punkt 5 nie wiadomo czego są to statystyki;
- str. 169 „W bazie danych znajdują się jednak połączone elementy z różnych dziedzin." -> W bazie danych znajdują się połączenia pomiędzy elementami należącymi do różnych dziedzin?
- str. 174 „Wybór jednej z opcji..." błędne uzgodnienie czasownika -> powinny;
- str. 175 „Budowanie definicji" -> opracowanie;
- str. 177 „Kolejnym narzędziem do wyznaczania kierunku jest tabela," kierunku czego?
- str. 180 „przez co można zaoszczędzić czas"-> dzięki czemu;
- str. 187 „co może wpłynąć na ogólną ocenę uczestników"-> ocenę wystawioną przez uczestników;
- str. 198 „Może się okazać, że pewne wyniki są za wysokie i należy zaimplementować inne biblioteki" może się okazać, że czas przetwarzania jest zbyt długi i należy użyć...;
- str. 201 „do którego będę mieli dostęp";

4. Konkluzja

Autor wykazał się znajomością teorii opracowywania słowników dziedzinowych, sprawnością programistyczną oraz umiejętnością oceny osiągniętych wyników. Praca stanowi istotny wkład w opracowywanie interaktywnych polsko-angielskich słowników specjalistycznych, z przeznaczeniem do wykorzystania w specyficznym środowisku szkoleń wojskowych. Wskazane w recenzji uwagi merytoryczne nie są kluczowe dla oceny całej rozprawy. Stwierdzam więc, że rozprawa pod tytułem: *Polsko-angielska terminologia Wojsk Specjalnych: Metodologia tworzenia branżowych leksykograficznych baz danych wraz z modelowaniem aplikacji sieciowej* spełnia wymogi, jakie ustawa o stopniach i o tytule naukowym przewiduje dla rozpraw doktorskich. Na tej podstawie wnoszę o dopuszczenie Autora pracy – mgr. Sebastiana Dziewięckiego – do publicznej obrony.

Małgorzata Marciniak