Uniwersytet Warszawski
Wydział Filozofii

Kamil Tomaszek
Nr albumu: 432044

Minimalizacja długości zależności w strukturach współrzędnie złożonych: badanie korpusowe na podstawie Polish Dependency Bank

Praca licencjacka
na kierunku KOGNITYWISTYKA

Praca wykonana pod kierunkiem
prof. dr hab. Adama Przepiórkowskiego
Uniwersytet Warszawski

Warszawa, czerwiec 2023

Streszczenie

Praca licencjacka na temat "Minimalizacja długości zależności w strukturach współrzędnie złożonych: badanie korpusowe na podstawie Polish Dependency Bank"jest poświęcona zjawisku minimalizacji długości zależności (DLM) w koordynacji w języku polskim. Celem pracy jest sprawdzenie hipotez na ten temat oraz przedstawienie dodatkowych analiz. Ma ona charakter empiryczny i opiera się na danych pochodzących z Polish Dependency Bank (PDB). Praca składa się z sześciu rozdziałów. W pierwszym rozdziale przedstawiłem motywację, cel i zakres pracy oraz jej strukturę. W drugim rozdziale omówiłem teoretyczne podstawy pracy, tj. reprezentacje koordynacji w języku polskim, teorię zależności składniowej i DLM w koordynacji. W trzecim rozdziale opisałem źródło danych i narzędzia do analizy, tj. PDB i preprocessing danych za pomocą algorytmu napisanego w Pythonie. W czwartym rozdziale zaprezentowałem wyniki analizy statystycznej wykonanej w R oraz testowanie hipotez za pomocą m. in. testu chi-kwadrat. W piątym rozdziale dokonałem dyskusji wyników, interpretacji ich znaczenia i porównania z literaturą naukową. W szóstym rozdziale podsumowałem pracę i wnioski oraz zaproponowałem perspektywy dalszych badań.

Słowa kluczowe

koordynacja, minimalizacja długości zależności, Polish Dependency Bank, drzewo zależnościowe, korpus języka polskiego

Tytuł pracy w języku angielskim

Dependency Length Minimisation in coordinate structures: A corpus study based on Polish Dependency Bank

Spis treści

Rozdział 1

Wstęp

W tym rozdziale przedstawiam motywację i cel pracy licencjackiej na temat "Minimalizacja długości zależności w strukturach współrzędnie złożonych: badanie korpusowe na podstawie Polish Dependency Bank", a także omawiam jej zakres oraz strukturę.

1.1. Motywacja i cel pracy

Praca ta ma na celu analizę zjawiska minimalizacji długości zależności – DLM (z ang. Dependency Length Minimisation), czyli tendencji do umieszczania elementów współrzędnych o różnych długościach w sposób, by zmniejszyć odległość zarówno między nimi samymi, jak i między nimi a innymi elementami zdania, w koordynacjach w języku polskim. Koordynacja to zjawisko, gdy wiele części zdania ma jeden nadrzędnik i każda z nich się z nim koordynuje. Zjawisko to jest istotne dla teorii składniowej i reprezentacji językowych, ponieważ dotyczy zarówno formy jak i znaczenia zdań. W pracy tej sprawdzono dwie hipotezy dotyczące długości członów w koordynacjach w języku polskim: 1. że dłuższy człon koordynacji jest częściej ze strony prawej i 2. że dłuższy człon koordynacji jest częściej dalej od jej nadrzędnika.

(0) *Widziałem* Asię **i** jej śmiesznego, młodszego brata.

Długości członów mierzono na cztery różne sposoby, licząc znaki, sylaby, słowa oraz tokeny. W przykładzie (0) odpowiednie wartości wynosiłyby [4 vs 31, 2 vs 9, 1 vs 4, 1 vs 5]. Szybko pokazano, że jedna z hipotez zachodzi w większości przypadków, więc następnie omówiono wpływ obecności i pozycji nadrzędnika oraz długości różnicy między analizowanymi członami na proporcje danych, w których hipoteza ta jest prawdziwa. Praca ta ma charakter empiryczny, opiera się na danych pochodzących z Polish Dependency Bank (PDB), czyli korpusu języka polskiego zawierającego ponad 22 tysiące drzew zależnościowych oraz na podobnej pracy badającej te same zależności, ale dla języka angielskiego (Anonimowe Zgłoszenie na ACL, nieopublikowane).

1.2. Zakres i struktura pracy

Praca składa się z sześciu rozdziałów. W rozdziale drugim omówiono teoretyczne podstawy pracy, tj. przedstawiono czym są koordynacje – na przykładzie języka polskiego, opisano zarys teorii zależności składniowej, zaprezentowano . W rozdziale trzecim opisano źródło danych, czyli Polish Dependency Bank, jak i ich preprocessing – działanie algorytmu, napisanego w języku Python, wybierającego koordynacje oraz informacje o nich z PDB, a także pokazano format danych po preprocessingu w pliku z rozszerzeniem ".csv". W rozdziale czwartym zaprezentowano hipotezy badawcze, ich testowanie oraz analizy statystyczne w R, między innymi test Wilcoxona, testy chi-kwadrat oraz ogólne modele liniowe (GLM – z ang. Generalised Linear Models). W rozdziale piątym omówiono wyniki badań i ich interpretację w kontekście istniejącej wcześniej literatury naukowej. W rozdziale szóstym podsumowano pracę, wyciągnięto z niej wnioski oraz zaproponowano perspektywy dalszych badań.

Rozdział 2

Teoretyczne podstawy minimalizacji długości zależności w strukturach współrzędnie złożonych

W tym rozdziale omawiam teoretyczne podstawy pracy, tj. reprezentacje koordynacji w języku polskim, teorię zależności składniowej i DLM w koordynacji.

2.1. Koordynacja w języku polskim

Zacznę od przedstawienia pojęcia koordynacji. Koordynacja to zjawisko w językach naturalnych, które zachodzi w struktruach złożonych – zarówno współrzędnie, jak i podrzędnie. Polega na zestawieniu dwóch lub więcej elementów o tej samej funkcji składniowej za pomocą spójników lub interpunkcji i tym samym złączenie ich w jeden, większy element, zachowujący te same funkcje składniowe. Jest ono jednym z podstawowych sposobów łączenia słów, czy zdań. Elementami koordynacji mogą być zarówno pojedyncze słowa (1a, 1b), wyrażenia (1c), jak i całe zdania (1d):

(1)
a. Ania **i** Julia *idą* na spacer.
b. Ania **i** Julia.
c. Wesoła Marysia **oraz** smutny Janek *wybrali się* do parku.
d. *Kuba* zjadł obiad **a** następnie poszedł spać.

Człony koordynacji nazywamy koordynantami, to co je łączy – spójnikiem koordynacji (w przykładach w tej pracy jest on ilustrowany pogrubionym tekstem), a wyraz nadrzędny względem obu członów – głową koordynacji (w przykładach ilustrowany kursywą). Jak widać w (1b) nie zawsze istnie jengłowa koordynacji. W podanych wyżej przykładach koordynantami są: (1a, 1b) Ania, Julia; (1c) Wesoła Marysia, smutny

Janek; (1d) zjadł obiad, poszedł spać.

Ze względów semantycznych zwykle wyróżnia się cztery rodzaje koordynacji: koordynacje koniunkcyjne (2a), koordynacje dysjunkcyjne (2b), koordynacje adwersatywne (2c) oraz koordynacje kauzalne (2d) (Haspelmath, 2007). Każde z nich używają różnych zestawów spójników, które łączą koordynanty. W koordynacjach koniunkcyjnych koordynanty łączą m. in. spójniki [i, oraz, ani, tudzież, również], a w koordynacjach dysjunkcyjnych – [albo, bądź, lub, czy], lecz w obu tych kategoriach wykorzystywana jest także interpunkcja. W koordynacjach adwersatywnych używamy m. in. spójników [ale, lecz, zaś, natomiast, jednak], a w koordynacjach kauzalnych – [bo, ponieważ, dlatego że]. W tej pracy skupię się na pierwszych trzech rodzajach koordynacji, jako że to one występują w strukturach współrzędnie złożonych.

(2)
a. Marta *zjadła* jabłko **i** gruszkę.
b. Ona miała szesnaście **lub** siedemnaście *lat.*
c. *Byli* ładni, **ale** głupi.
d. Nie zrobiłem pracy domowej, **bo** nie chciałem.

2.2. Zarys teorii zależności składniowej

2.3. Minimalizacja długości zależności w koordynacji

Tekst sekcji

2.4. Co poszczególne reprezentacje przewidują

Tekst sekcji

Rozdział 3

Polish Dependency Bank – źródło danych i narzędzia do analizy

Tekst rozdziału

3.1. Krótki opis Polish Dependency Bank

Tekst sekcji

3.2. Preprocessing danych

Tekst sekcji

3.3. Dane po preprocessingu

Tekst sekcji

Rozdział 4

Analiza statystyczna

Tekst rozdziału

4.1. Hipoteza, metody

Tekst sekcji

4.2. Przedstawienie wyników analizy statystycznej w R

Tekst sekcji

4.3. Testowanie hipotez

Tekst sekcji

Rozdział 5

Dyskusja wyników

Tekst rozdziału

5.1. Podsumowanie wyników badań

Tekst sekcji

5.2. Interpretacja wyników

Tekst sekcji

5.3. Przegląd literatury

Tekst sekcji

Rozdział 6

Zakończenie

Tekst rozdziału

6.1. Podsumowanie pracy i wnioski

Tekst sekcji

6.2. Perspektywy dalszych badań

Tekst sekcji

Bibliografia

Anonimowe Zgłoszenie na ACL 2023 (nieopublikowane) Conjunct Lengths in English, Dependency Length Minimization, and Dependency Structure of Coordination

Haspelmath, M. (2007) Coordination, *Language Typology and Syntactic Description, Volume II: Complex constructions*, 1-51

Załączniki