posłowie do wydania polskiego

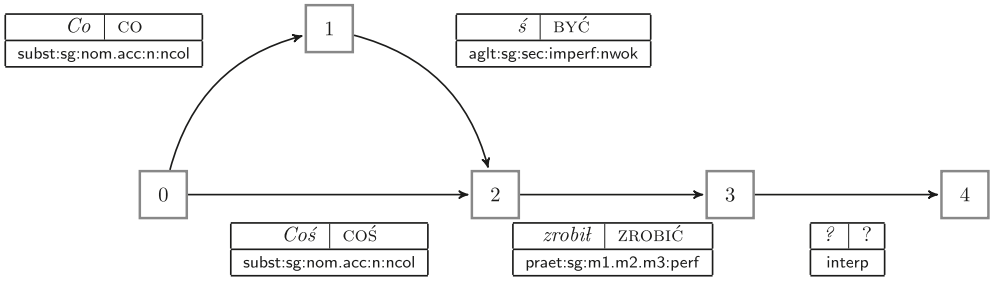
dr Łukasz Kobyliński, Ryszard Tuora

P.1. Wprowadzenie

Dotychczas przetwarzanie języka naturalnego było omawiane na przykładzie języka angielskiego. Ogromna większość przedstawionych koncepcji, problemów i sposobów ich rozwiązania ma zastosowanie również w przypadku przetwarzania innych języków. Istnieje jednak wiele wyzwań w NLP, które są specyficzne dla danego języka. Jest to naturalne: w oczywisty sposób nieco inaczej podeszlibyśmy do przetwarzania języków zapisywanych za pomocą alfabetu, a inaczej tych, których zapis wykonywany jest piktograficznie (pismem obrazkowym). Na szczęście w przypadku języka polskiego nie mamy do czynienia z aż tak wielkimi różnicami w stosunku do prezentowanych w książce podejść i przykładów dotyczących języka angielskiego. Te różnice, które są istotne dla automatycznego przetwarzania języka naturalnego, dotyczą zjawisk na kilku poziomach opisu języka, do których odniesiemy się w tym posłowiu, aby wskazać możliwe problemy i sposoby ich rozwiązywania.

Jedną z najistotniejszych dla zastosowania metod NLP różnic jest kwestia budowy słów w języku polskim. Język ten jest przedstawicielem grupy języków zachodniosłowiańskich, które charakteryzują się bogatą fleksją, a zatem słowa powstają z pewnej podstawy morfologicznej i końcówki fleksyjnej, co skutkuje dużym zróżnicowaniem form ortograficznych poszczególnych leksemów. Na przykład: rzeczownik *brat* odmienia się oczywiście przez przypadki i liczby i przyjmuje wówczas takie formy jak: *brata*, *bratem*, ale też *bracia*, *braci*, *braćmi*. Zachodzi zatem tzw. wymiana w temacie i formy odmienione nie są już łatwo porównywalne ortograficznie z formą podstawową. Ten poziom analizy językowej reprezentowany jest w przypadku NLP przez metody analizy i znakowania morfostyntaktycznego, a także stemmingu czy lematyzacji i metody te dosyć znacznie różnią się od metod, które można zastosować w przypadku języka angielskiego.

Na poziomie opisu morfologicznego słów mamy też do czynienia z problemem segmentacji słów. Dla języka polskiego istnieje wiele przypadków niejednoznaczności podziału ciągu znaków na poszczególne tokeny, które będą podlegać dalszemu przetwarzaniu. Niejednoznaczności te muszą być rozwiązywane na podstawie analizy kontekstu (często szerszego niż zdanie). Przykładem takiej niejednoznaczności jest ciąg znaków *Coś* w zdaniu *Coś zrobił?* Zdanie to ma dwie interpretacje zależne od sposobu przeprowadzenia segmentacji, co jest pokazane na rysunku P.1.



Rysunek P.1. Przykład niejednoznaczności segmentacyjnej dotyczącej ciągu znaków *Coś* w zdaniu *Coś zrobił?*
Źródło: na podstawie instrukcji Morfeusza:- http://download.sgjp.pl/morfeusz/Morfeusz2.pdf.

W pierwszej interpretacji słowo *coś* traktujemy jako całość, a więc rzeczownik, który nadaje zdaniu sens pytania o to, czy osoba, o którą pytamy, „coś zrobiła", czy wykonała jakąś rzecz/czynność (por. *Czy [on] coś zrobił?*). W drugiej interpretacji słowo *coś* jest zbitkiem dwóch leksemów: leksemu *co* oraz leksemu *być* i wówczas zdanie nabiera znaczenia zapytania naszego rozmówcy o to, co ten właśnie rozmówca zrobił (por. Co *zrobiłeś?*)

Inną, istotną dla NLP różnicą między językiem polskim a językiem angielskim jest swobodny szyk zdań. W przypadku języka angielskiego szyk ten jest ustalony i realizowany według schematu SVO, a więc *subject-verb-object* (podmiot-orzeczenie-dopełnienie). W języku polskim taki szyk również jest najczęstszy, natomiast nie ma przeciwskazań, aby zdania tworzyć w zupełnie innej kolejności elementów (por. *John loves Mary* oraz *Jan kocha Marię*, ale też *Jan Marię kocha*, czy *Marię kocha Jan*).

Łącznie wspomniane wyżej trudności w analizie języka polskiego składają się na wiele poziomów **niejednoznaczności**, które są głównym problemem w zastosowaniu jakichkolwiek metod automatycznej analizy i przetwarzania danych. Niejednoznaczności te muszą być rozwiązywane na każdym z poziomów z możliwie dużą dokładnością, ponieważ błędy z wcześniejszych etapów przetwarzania mają wpływ na dokładność tych etapów, które realizowane są w kolejnych krokach potoku przetwarzania.

Na koniec krótka uwaga techniczna. W przykładach poniżej używamy – podobnie jak autorzy książki – pakietu spaCy, który w łatwy sposób umożliwia realizowanie wielu zadań związanych z przetwarzaniem języka naturalnego. W chwili pisania tego tekstu istnieją co najmniej dwa modele języka polskiego do spaCy: model oficjalny, ogłoszony przez twórców spaCy latem 2020[1], oraz model przygotowany przez IPI PAN jesienią

[1] https://spacy.io/models/pl.