Wrocław, 24.06.2017
Bartosz Rybicki

Research Engineer: Zadanie rekrutacyjne - Raport

**Treść**: Opracuj prototyp podziału zdania złożonego na zdania pojedyncze dla tekstów potocznych dla języka polskiego.

## Moja propozycja rozwiazania - „algorytm”

„Zdania pojedyncze mają jedno orzeczenie. Zdania złożone powstają w skutek połączenia zdań pojedynczych, najczęściej spójnikiem” - taka podstawowa wiedza wystarcza żeby spróbować zmierzyć się z tym zadaniem. Poniżej przedstawiam sposób zastosowania tego podejścia wraz z drobnymi „poprawkami”, które pomagaja rozwiązać szczególne przypadki w których zdania proste nie są rozdzielone spójnikiem.

Plan (algorytm) na rozwiązanie zadania jest następujący:

- podzielić tekst na zdania a te na tokeny
- otagować tokeny (oraz otagować hashtagi jako rzeczowniki)
- rozbić każde zdanie złożone na zdania-kandydatów (dzielimy wg. spójników)
- każde zdanie-kandydat które zawiera choć jedno orzeczenie uznajemy za zdanie proste, natomiast zdanie-kandydata które nie zawiera orzeczenia łaczymy z sąsiednim zdaniem-kandydatem
- zdania proste, które zawierają k > 1 czasowników dzielimy na k zdań pojedynczych (dzielenie zaraz po wystąpieniu czasownika – to pewna heurystyka, bo nie mam konkretnego pomysłu jako to zrobić dobrze, pewnie trzeba by się bliżej przyjrzeć tokenom i w ten sposób zadecydować o miejscu podziału)

## Problemy jakie napotkałem oraz pomysły na ich rozwiązanie

Problem: zastąpienie w tekście wejściowym liter ze znakami diakrytycznymi odpowiednimi literami ‘zwykłymi’ - co istotnie rzutowało na proces tagowania, który jest w całym zadaniu kluczowy

Rozwiązanie: wytrenowanie tagera na mieszance zdań napisanych poprawnie oraz zdań w których (niektóre losowo wybranie) litery ze znakami diakrytycznymi zostały zastąpione odpowiednimi literami ‘zwykłymi’

Problem: niepoprawna pisownia słów tj. przekręcanie klejności liter lub ich pomijanie ew. zastępowanie niepoprawnymi literami – problem z tagowaniem

Rozwiązanie: zastąpienie słowa nieobecnego w słowniku przez możliwie podobne słowo ze słownika (poszukiwania odpowiednego słowa można ograniczyć do słów które są sensowne jeśli weźmiemy pod uwagę kontekst tj. poprzednie słowo / słowa z tego zdania – landuage modeling: n-gram language model lub neural language model). Jako miary podobieństwa można wykorzystać odległość Hamminga lub odległość edycyjną

Problem: nie-słowa – hashtagi itp. (można je odrazu otagować jako nie-czasowniki oraz nie-spójniki)

Rozwiązanie: stworzenie specjalnego taga np. „hashtag” i wykorzystanie przy tagowaniu. Oczywiście to rozwiazanie wymaga wytrenowania tagera na zbiorze poszerzonym o zdania w których występują nie-słowa.

Problem: named-entities – jeśli elementem encji nazwanej jest spójnik lub czasownik to nie powinien być on brany pod uwagę w procesie dzielenia zdania złożonego na zdania pojedyncze.

Rozwiązanie: rozpoznanie named-entities oraz odpowiednie ich otagowanie, co później pomoże ‘ignorować’ odpowiednie spójniki oraz czasowniki w procesie dzielenia zdania złożonego.

Podsumowując, główne problemy jakie napotkałem były związane z niepoprawnym tagowaniem tokenów co wynikało głównie z tego że tekst jest napisany językiem potocznym. Konsekwencje niepoprawnego otagowania są oczywiste – problemy z poprawnym wyborem miejsca podziału. Wydaje mi się, że podane wyżej sposoby rozwiązania tego problemu byłby skuteczne i istotnie wpłynełby na poprawienie jakości wyników zwracanych przez algorytm opisany w początkowym fragmencie tego dokumentu.