Zadanie NLP 3: Atrybuty

1 Opis

Zadanie wyznaczenia atrybutów polega na przypisaniu do każdej wzmianki (anotacji) wartości dwóch atrybutów, określających jej związek z pacjentem. Przykładowo w zdaniu ‘Wywiad rodzinny: ojciec – nadciśnienie’ wzmianka o nadciśnieniu musi nosić odpowiednie oznaczenie dla uwzględnienia informacji, że nadciśnienie dotyczy członka rodziny, a nie samego pacjenta. Możliwe jest to z wykorzystaniem cech kontekstu wzmianki – w tym wypadku słów ‘rodzinny’ i ‘ojciec’.

W ramach projektu uwzględniono następujące atrybuty:

- `Status`, określający stan danego zdarzenia (np. zabiegu) lub dolegliwości (np. choroby, objawu) względem pacjenta, z wartościami:
  - `Historical insignificant` – zdarzenie nastąpiło na tyle dawno, że ma małe znaczenie dla aktualnego stanu pacjenta,
  - `Family` – zdarzenie dotyczy innej osoby niż pacjent,
  - `Current` (domyślna) – w pozostałych przypadkach.
- `Source`, określający źródło informacji o danym zdarzeniu, z wartościami:
  - `Declared` – informacja pochodzi z deklaracji pacjenta i jest niepewna,
  - `Confirmed` (domyślna) – w pozostałych przypadkach.

Uwaga 1: Wartości `Current` i `Confirmed` pasują do zdecydowanej większości przypadków i są domyślne, tj. brak określenia atrybutu dla jakiejś wzmianki oznacza te właśnie wartości.

Uwaga 2: Atrybuty `Status` i `Source` są niezależne i oba mają zastosowanie dla każdej wzmianki.

2 Specyfikacja

Specyfikacja zadania zdefiniowana jest jako przekształcanie plików w formacie programu BRAT (https://brat.nlplab.org/standoff.html), w których przechowywane są anotacje w projekcie.

Wejście

- Plik TXT zawierający czysty tekst,
- Plik ANN zawierający znakowanie wzmianek (znaczniki ‘T’) pochodzące od anotatora lub algorytmu NER, np.

  ```
  T110 Treatment 4781 4821 operacji usunięcia tłuszczaków z żołądka
  ```

Wyjście

- Plik ANN otrzymany na wejściu, ale z dodanymi atrybutami (znaczniki ‘A’) przypisanymi do odpowiednich wzmianek, np.:

  ```
  A10  Status T110 Historical_Insignificant
  ```

Przykładowe pliki wejściowe i wyjściowe załączono do niniejszego opisu.

3 Ewaluacja

Ewaluacja algorytmu wykrywania relacji będzie polegała na usunięciu znaczników ‘A’ z ręcznie oznakowanych plików, przekazaniu ich do automatycznego wyznaczenia atrybutów i porównaniu wyników z tym, co wprowadzili anotatorzy. Dla każdego z atrybutów obliczona zostanie miara **dokładności** określająca, w jak wielu wzmiankach wartość tego atrybutu pochodząca z algorytmu zgadza się z tą wyznaczoną przez anotatorów, przy czym pomija się wzmianki, gdzie oba źródła używają wartości domyślnej.

4 Możliwe rozwiązania

Rozwiązanie tego problemu musi polegać na analizie kontekstu, w którym pojawiają się słowa takie, jak ‘wywiad rodzinny’, ‘ojciec’, ‘deklaruje’ ‘10 lat temu’, ‘stwierdzono’. Część kontekstu może zostać uwzględniona z pomocą anotacji, np. oznakowanie daty powiązanej z chorobą może rozstrzygnąć jej historyczność.

Podobnie jak w pozostałych zadaniach, możliwe są tu rozwiązania bazujące na ręcznym przygotowaniu reguł lub zastosowaniu uczenia maszynowego. Pierwszym krokiem może być sporządzenie listy słów najczęściej poprzedzających wzmianki o nie-domyślnej wartości atrybutu. Dodatkową trudność stanowi subiektywność wyboru właściwej wartości np. przy ocenie historyczności, która znajduje odwzorowanie w niskiej zgodności między znakującymi.