Zadanie NLP 2: Relacje

1 Opis

Zadanie wyznaczania relacji polega na powiązaniu niektórych par anotacji (wzmianek) w tekście w sposób odpowiadający ich powiązaniu znaczeniowymu. Przykładowo w zdaniu ‘W dniu 10.12.2013 wykonano cewnikowanie serca.’, zawierającym wzmianki typu `Date` i `Investigation`, należy utworzyć powiązanie typu `Dat`, oznaczające, że wyróżniona data wskazuje czasy wykonania tego zabiegu. Zadanie to nie jest trywialne, gdyż np. w zdaniu omawiającym historię choroby może wystąpić wiele zjawisk:

- schematyczne wyliczenie naprzemiennie dat i przypisanych im schorzeń,
- podanie pojedynczej daty, a następnie listy schorzeń, których ona dotyczy,
- obecność schorzeń bez przypisanej daty,
- stosowanie różnych konwencji zapisu: data poprzedzająca lub następująca po powiązanym schorzeniu, używanie nawiasów, dwukropków, dywizów itp.

Dopiero rozwikłanie wszystkich tych problemów pozwoli na ustalenie chronologii choroby.

Analogiczne trudności dotyczą wszystkich typów relacji, które w ramach projektu są następujące:

- `Inv`, łącząca badanie (`Investigation`) z jego wynikiem (`Investigation result`),
- `Neg`, łącząca określenie zaprzeczenia (`Negation`) z zaprzeczanym fragmentem (`Symptom`, `Condition`, `Behaviour`, `Treatment`, `Investigation`, `Investigation result`, `Drug` lub `Drug dose`),
- `Drg`, łącząca lek (`Drug`) z dawkowaniem (`Drug dose`),
- `Dat`, łącząca określenie daty wystąpienia zdarzenia (`Date`) z określeniem zdarzenia (`Symptom, Condition, Behaviour, Treatment, Investigation` lub `Drug`),
- `Alg`, łącząca anotację uczulenia (typu `Condition`) z substancją uczulającą (`Drug`).

Uwaga: Każda z relacji jest skierowana, czyli rozróżnia element źródłowy od docelowego, jednak w tym modelu anotacji kierunek jest jednoznacznie wyznaczany przez typy wzmianek, np. zawsze od `Drug` do `Drug dose`, a nie odwrotnie.

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.

  ```
  T41  Negation 2715 2721  Neguje
  T42  Symptom 2722 2734   zasłabnięcia
  ```

Wyjście

- Plik ANN otrzymany na wejściu, ale z dodanymi relacjami (znaczniki ‘R’) pomiędzy odpowiednimi wzmiankami, np.:

  ```
  R13  Neg Arg1:T41 Arg2:T42
  ```

Przykładowe pliki wejściowe i wyjściowe załączono do niniejszego opisu.

3 Ewaluacja

Ewaluacja algorytmu wykrywania relacji będzie polegała na usunięciu znaczników ‘R’ z ręcznie oznakowanych plików, przekazaniu ich do automatycznego wyznaczenia relacji i porównaniu wyników z tym, co wprowadzili anotatorzy. Obliczone zostaną miary:

- **Pokrycie** określa, jak wiele z relacji oznaczonych przez anotatorów zostało także wykrytych przez automatyczny algorytm.
- **Precyzja** określa, jak wiele z relacji zwróconych przez algorytm ma swój odpowiednik ręcznych anotacjach.

4 Możliwe rozwiązania

Architekturę oczekiwanego rozwiązania można podzielić na kilka etapów:

1. Podział tekstu na fragmenty, np. zdania.

Pierwszy krok ma na celu podział całości na fragmenty tak, aby pomiędzy wzmiankami pochodzącymi z różnych fragmentów wykluczone były relacje. Najbardziej oczywistą jednostką podziału wydaje się zdanie, gdyż np. negacja w jednym zdaniu (np. przez słowo ‘nie’) nie ma zastosowania do wzmianek w dalszych zdaniach. Należy jednak tę hipotezę zweryfikować na praktycznych przykładach, szczególnie uwzględniając treści o różnego rodzaju formatowaniu, podziały na linie, wyliczenia itp.

2. Wyodrębnienie kandydatów na relacje.

Drugi krok jest bardzo prosty, gdyż ma na celu wybranie wszystkich par wzmianek w danym fragmencie, pomiędzy którymi mogłaby zachodzić relacja. Pary te są wyznaczane jednoznacznie przez typy wzmianek, np. jeśli we fragmencie występuje wzmianka typu `Negation` i `Symptom`, to należy utworzyć kandydata do relacji `Neg`.

3. Wybór relacji do realizacji spośród kandydatów

Ostatni krok jest kluczowy, gdyż tylko część z kandydatów rzeczywiście będzie zrealizowana. Przykładowo, jeśli w zdaniu występuje 5 dat i 6 chorób, oznacza to 30 kandydatów na relacje, ale w praktyce należy oczekiwać realizacji kilku z nich. Rozstrzygnięcie, czy dany kandydat łączący dwie wzmianki zostanie zrealizowany może uwzględniać przykładowo następujące czynniki:

- treść wzmianki źródłowej, np. ‘Neguje:’ będzie miało większy zasięg niż ‘bez’,
- wystąpienie wzmianki źródłowej przed lub po wzmiance docelowej,
- odległość (tj. liczba słów lub znaków) między wzmiankami,
- obecność innych wzmianek w tekście pomiędzy wzmianką źródłową a docelową,
- typ relacji,
- interpunkcyjne znaczniki zasięgów słów, np. dywizy, nawiasy, myślniki, przecinki itp.

Reguły tego typu będą charakterystyczne dla tekstów z danego źródła (np. szpitala czy systemu sprawozdawczego) i mogą zostać przygotowane ręcznie na podstawie obserwacji przykładów lub poprzez uczenie maszynowe na zbiorze oznakowanych dokumentów.