Analiza języka naturalnego
Sprawozdanie 1

Weronika Pawlak
Agata Skibińska

2 Listopad 2020

1 Segmentator

1.1 Zasada działania segmentatora

Segmentator działa według kilku zasad:

1. Segmentator rozdziela tekst według znaków białych: tab, spacja, enter
2. Segmentator wydziela jako osobne segmenty znaki interpunkcyjne ze zbioru:
   (a) . (kropka)
   (b) , (przecinek)
   (c) ! (wykrzyknik)
   (d) ? (znak zapytania)
   (e) : (dwukropek)
   (f) ; (średnik)
   (g) ( ) (nawiasy prawe i lewe)
   (h) [] (nawiasy kwadratowe)
   (i) "" (cudzysłów)
3. Segementator nie rodziela skrótowców jednoliterowych typu w. lub o.
4. Segmentator nie rozdziala skrótowców wieloliterowych z podanej listy: np., itd., tzn., itp., m.in., ect., tys., lek., lic., inż., ul., al., def., dot., śp., tzw., tys.

1.2 Wnioski

Zadanie napisania segmentatora jest samo w sobie dość skomplikowane biorąc pod uwagę złożoność języka polskiego. Powstały w ramach tego ćwiczenia segmentator radzi sobie tylko w najprostszych przypadkach. Ze względu na dużo więcej różnych sposobów zapisywania wyrazów, aby używać go do innych zadań, np. tagowania, należałoby dodać do niego dużo więcej reguł, użyć do niego bazy danych skrótowców j. polskiego.

2 Porównanie działania tagerów dla j.polskiego

2.1 Metodologia badań

W zadaniu zbadano działanie tagerów:

1. WCRFT2 (https://ws.clarin-pl.eu - dostępny pod API Clarin-pl)
2. MorphoDiTa (https://ws.clarin-pl.eu - dostępny pod API Clarin-pl)

	Wcrtf2	**MorphoDita**	**KRNNT**
POS accuracy (Subtask A score)	64.5684%	89.7979%	**92.6242%**
POS accuracy (known words)	65.7766%	91.2451%	**93.4597%**
POS accuracy (unknown words)	6.4171%	20.1426%	**52.4064%**
Lemmatization accuracy (Subtask B score)	90.7122%	96.4917%	**97.0359%**
Lemmatization accuracy (known words)	91.3488%	97.2521%	**97.7039%**
Lemmatization accuracy (unknown words)	60.0713%	59.8930%	**64.8841%**
Overall accuracy (Subtask C score)	77.6403%	93.1448%	**94.8300%**

Tablica 1: Wyniki działania tagerów w porównianiu z gold standard konkursu PolEval 2017

3. KRNNT (https://github.com/kwrobel-nlp/krnnt)

Działanie tagerów porównano na tekście z konkursu PolEval 2017. Sprawdzano skuteczność tagerów poprzez zadanie wydzielenia części mowy (ang. POS tagging) oraz lematyzacji (ang. lemmatization). Skuteczność tagowania oceniono za pomocą podanym przez organizatorów skryptu (http://2017.poleval.pl/task1/tagger-eval.py).

2.2 Wyniki

Porównując działanie trzech tagerów, zdecydowanie najlepiej radził sobie tager KRNNT. Jeśli chodzi o zadanie wydzielenia części mowy najbardziej odbiegał od reszty na niekorzyść tager wcrtf2. MorphoDita pozwoliła osiągnąć porównywalne do KRNNT wyniki.

3 Naiwny klasyfikator Bayesa

Do wykonania zadania został wykorzystany MultinomialNB z biblioteki sklearn.

3.1 Preprocessing

- Z przeparsowanych plików tekstowych wyciągane są słowa wraz z tagami
- Tagi zostały podzielone na poszczególne części mowy, zgodnie z następującą refułą:
  - verb = ['fin', 'bedzie', 'aglt', 'praet', 'impt', 'imps', 'inf', 'pcon', 'pant', 'ger', 'pact', 'ppas', 'winien']
  - noun = ['subst', 'depr']
  - adj = ['adj', 'adja', 'adjp', 'adjc']
- Dane są filtrowane do jednej części mowy

`https://www.sketchengine.eu/polish-nkjp-part-of-speech-tagset/`
`http://nkjp.pl/poliqarp/help/ense2.html`

3.2 Zasada działania

- Ze słów występujących w tekstach budowany jest słownik (określonej długości)
- Dla każdego z tekstów wyznaczany jest wektor bag of words
- Lista wektorów bag of words -> X
- Lista tytułów przeanalizowanych tekstów -> y

3.3 Wyniki

3.3.1 Słownik długości 1000

	wcrft2		MorphoDiTa		KRNNT	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**
Czasowniki	0.546	0.523	0.568	0.544	0.566	0.542
Rzeczowniki	0.809	0.802	0.815	0.809	0.814	0.807
Przymiotniki	0.716	0.704	0.703	0.692	0.705	0.694

Tablica 2: Metryki klasyfikatora w zależności od tagera i części mowy

3.3.2 Słownik długości 5000

	wcrft2		MorphoDiTa		KRNNT	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**
Czasowniki	0.576	0.554	0.597	0.573	0.597	0.574
Rzeczowniki	0.856	0.852	0.859	0.855	0.859	0.855
Przymiotniki	0.752	0.741	0.751	0.739	0.75	0.739

Tablica 3: Metryki klasyfikatora w zależności od tagera i części mowy

3.4 Wnioski

Niezależnie od tagera, jakość klasyfikacji jest najlepsza, gdy klasyfikator bazuje na rzeczownikach. Jest to efekt, którego można było się spodziewać, rzeczowniki najlepiej oddają tematykę tekstu. Przy wykorzystaniu czasowników do budowy słownika, jakość klasyfikacji znacznie spada. Czasowniki nie są tak bardzo zależne od tematu. Można wyróżnić zbiór tych bardziej specyficznych np. ['kondensować', 'przezbroić', 'naelektryzować', 'zaryglować', 'perforować', 'zislamizować'] jednak w porównaniu z neutralnymi czasownikami takimi jak np. ['stanowić', 'być', 'spotkać', 'wybierać', 'oceniać', 'podejść'] jest ich stosunkowo niewiele. Rozkład czasowników może sugerować piewien zakres tematów, jednak nie jest to tak jednoznaczne jak w przypadku rzeczowników. Wykorzystanie przymiotników sprawdza się nieznacznie gorzej niż wykorzystanie rzeczowników. Przymiotnik określa rzeczownik, zatem rozkłady przymiotników i rzeczowników są w pewien sposób od siebie zależne. Przy słowniku ograniczonym do 5000 najczęściej występujących słów, wyniki są nieco lepsze niż przy zakresie 1000 słów. Nie musi to jednak oznaczać, że wydłużanie słownika powoduje poprawę jakości klasyfikacji. Zbyt długi słownik mógłby zawierać dużo przypadkowych słów, nie oddających w żadnym stopniu charakteru tekstu. Wyniki dla wszystkich trzech tagerów nie są bardzo rozbieżne, co oznacza, że różnice w tagowaniu części mowy są niewielkie.