SPRAWOZDANIE
Analiza języka naturalnego, Ćwiczenie 1

Marcel Cielinski, Kornel Romański

1. Segmentator

Program napisany w celu segmentacji tekstu działa zgodnie z poniższymi zasadami:

- zastępuje znaki nowej linii znakami spacji,
- każdy znak spacji jest miejscem rozdzielenia segmentów tekstu,
- jeżeli dwa następujące po sobie segmenty są zdefiniowane w zbiorze związków wyrazowych, to są łączone w jeden segment,
- znaki interpunkcyjne (zdefiniowane przez użytkownika w odpowiednim zbiorze) znajdujące się na początku lub na końcu wyrazu oddzielane są do pojedynczych segmentów,
- wyjątkiem jest znak kropki: jeżeli kropka znajduje się na końcu wyrazu, a wyraz ten nie znajduje się w zdefiniowanym przez użytkownika zbiorze skrótów, to kropka oddzielana jest do osobnego segmentu, a po niej dodawany jest znak ‘<eos>’ oznaczający koniec zdania, natomiast jeśli wyraz został zdefiniowany jako skrót, to nie jest rozdzielany,
- jeżeli znak interpunkcyjny znajduje się wewnątrz wyrazu, a wyraz nie został zdefiniowany w zbiorze związków wyrazowych, to jest on dzielony na segmenty.

Przykład działania:

- tekst - ‘Sklep sportowy otwarty w godz. 15-16 (codziennie). Do nabycia np. biało-czerwone stroje i czerwona kartka.’,
- zbiór skrótów - ('np.', 'godz.'),
- zbiór znaków interpunkcyjnych - ('(', ')', ‘-’),
- zbiór związków wyrazowych - ('czerwona kartka', 'biało-czerwone'),
- wynik segmentacji - ['Sklep', 'sportowy', 'otwarty', 'w', 'godz.', '(', '15', '-', '16', 'codziennie', ')', '.', '<eos>', 'Do', 'nabycia', 'np.', 'biało-czerwone', 'stroje', 'i', 'czerwona kartka', '.', '<eos>'].

Program dokonuje podstawowej segmentacji tekstu, jednak dla poprawności działania powinien zostać rozbudowany o obsługę takich przypadków jak: daty, przenoszenie wyrazu do nowej linii.

2. Wykorzystane tagery

W dalszej części zadania zapoznano się z trzema tagerami morfo-syntaktycznymi dla języka polskiego. Wybrane przez nas to:

- *morphoDita*
- *wcrft2*
- *krrnnt*

Do analizy tekstu za pomocą pierwszych dwóch narzędzi, wykorzystaliśmy usługę sieciową udostępnioną przez CLARIN-PL. Natomiast trzecie z nich zostało zainstalowane lokalnie.

3. Porównanie tagerów

Wybrane przez nas analizatory morfo-syntaktyczne zostały przetestowane na zbiorze testowym z konkursu PolEval.

Napisano własny skrypt do oceny stopnia zgodności segmentacji dla danych zwróconych przez tager i danych gold standard. Skrypt iteruje po wynikach dla danych gold standard i sprawdza czy na tej samej pozycji w danych wynikowych aktualnie rozpatrywanego tagera znajduje się taki sam wydzielony segment. Jeśli nie to szuka takiego segmentu w sąsiedztwie aktualnej pozycji (liczba przeszukiwanych sąsiednich pozycji jest równa różnicy w liczbie wydzielonych segmentów w wynikach gold oraz wynikach tagera). Jeżeli znalazł odpowiedni segment, to do odpowiednich list zapisywane są tag z wyników gold oraz tag przypisany przez rozpatrywany tager. Jeżeli nie został znaleziony odpowiedni segment, to jako wynik tagera zapisywano wartość '_', która jest interpretowana jako brak wyniku / błędna segmentacja. Tak zgromadzone tagi zostały porównane za pomocą miary dokładności klasyfikacji (accuracy), gdzie tagi z wyników gold były rozpatrywane jako wartości rzeczywiste, a wyniki tagera jako wartości przewidywane.

Wyniki prezentują się następująco:

TAGER:	morphoDita	wcrft2	krrnnt
DOKŁADNOŚĆ:	0.5814	0.4778	**0.5945**

4. Klasyfikacja tekstów z Wikipedii

Na potrzeby ostatniego podzadania, porównano wpływ działania poszczególnych tagerów jako narzędzi wstępnego przetwarzania na wyniki klasyfikacji tekstów (korpus Wikipedii z CLARIN-PL) za pomocą naiwnego klasyfikatora Bayesowskiego.

Do przeprowadzenia klasyfikacji konieczne było przypisane do segmentów odpowiednich części mowy na podstawie znalezionych tagów. Po przeanalizowaniu zwracanych tagów przyjęto następujące zasady:

- wyraz jest rzeczownikiem, jeżeli tag rozpoczyna się oznaczeniem 'subst',
- wyraz jest przymiotnikiem, jeżeli tag rozpoczyna się oznaczeniem 'adj',
- wyraz jest czasownikiem, jeżeli tag rozpoczyna się jednym z oznaczeń: ('fin', 'bedzie', 'aglt', 'praet', 'impt', 'imps', 'inf', 'pcon', 'pant', 'ger', 'pact', 'ppas').

Porównywano następujące podejścia:

- stosowany tager:
  - morphoDita

    - wcrft2
    - krrnnt
- wybrana część mowy:
    - rzeczownik (noun)
    - czasownik (verb)
    - przymiotnik (adjective)
- liczba najczęściej występujących słów w wektorze wejściowym do klasyfikacji (do wektorowej reprezentacji tekstu wykorzystano technikę *bag of words*):
    - 1000
    - 10000
    - 100000

Wyniki prezentują się następująco:

max_features	tagger	part_of_speech	accuracy
1000	morphoDita	noun	0.5577
1000	morphoDita	adjective	0.4538
1000	morphoDita	verb	0.2784
1000	wcrft2	noun	0.5445
1000	wcrft2	adjective	0.4545
1000	wcrft2	verb	0.2757
1000	krnnt	noun	0.5608
1000	krnnt	adjective	0.4446
1000	krnnt	verb	0.2743
10000	**morphoDita**	**noun**	**0.7592**
10000	morphoDita	adjective	0.6028
10000	morphoDita	verb	0.3681
10000	**wcrft2**	**noun**	**0.7416**
10000	wcrft2	adjective	0.5957
10000	wcrft2	verb	0.3549
10000	**krnnt**	**noun**	**0.7674**
10000	krnnt	adjective	0.5960
10000	krnnt	verb	0.3695
100000	**morphoDita**	**noun**	**0.7667**
100000	morphoDita	adjective	0.6051
100000	morphoDita	verb	0.3681
100000	**wcrft2**	**noun**	**0.7545**
100000	wcrft2	adjective	0.5957
100000	wcrft2	verb	0.3549
100000	**krnnt**	**noun**	**0.7999**
100000	krnnt	adjective	0.5967
100000	krnnt	verb	0.3695

Wnioski:

- segmentacja przy wykorzystaniu analizatora *krnnt* przełożyła się na najlepsze wyniki klasyfikacji,

- wykorzystanie rzeczownika przekłada się na znaczne lepsze rezultaty w stosunku do stosowania pozostałych badanych części mowy jako wejście klasyfikatora,
- zwiększanie wektora wejściowego pozytywnie wpływa na jakość klasyfikacji.