

Przetwarzanie języka naturalnego

Jan Daciuk

Katedra Inteligentnych Systemów Interaktywnych
Wydział ETI, Politechnika Gdańska

21 lutego 2015

Warunki zaliczenia

Jest **jedna** ocena za cały przedmiot: wykład i ćwiczenia laboratoryjne.
Obie części po 50% punktów. Min. 50% z każdej części.

Wykład: **egzamin** testowy, każdy otrzymuje indywidualny, losowo wybrany zestaw, 4 możliwe odpowiedzi, jedna prawidłowa, brak punktów ujemnych, można korzystać z dodatkowych materiałów, nie można korzystać z pomocy innych osób. Punkty przeskalowuje się do 45%. 5% punktów otrzymuje się za obecność.

Laboratorium: suma punktów za poszczególne ćwiczenia, przeskalowana na 50 punktów. Za ćwiczenie dostaje się zwykłą, tradycyjną ocenę. Zaliczenie następuje na zajęciach. W przypadku opóźnienia odejmowane jest pół stopnia za każdy rozpoczęty tydzień, w którym odbywają się zajęcia.

% pkt	oc.
96 – 100	5,5
90 – 95	5
80 – 89	4,5
70 – 79	4
60 – 69	3,5
50 – 59	3
0 – 49	2

- 1 Daniel Jurafsky, James Martin, *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Second Edition, Prentice Hall, 2008.
- 2 Christopher D. Manning, Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 2000.
- 3 Emmanuel Roche, Yves Schabes, *Finite-State Language Processing*, MIT Press, 1997.
- 4 Kwartalnik *Computational Linguistics* i materiały konferencji organizowanych przez *ACL* (Association for Computational Linguistics). Dostępne przez <http://acl.ldc.upenn.edu/> – *ACL Anthology*.

Literatura uzupełniająca dotycząca języka polskiego

- 1 Alicja Nagórko, *Zarys gramatyki polskiej*, Wydawnictwo Naukowe PWN, Warszawa, 1996.
- 2 Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński, Danuta Skowrońska, *Słownik gramatyczny języka polskiego*, Wydanie II, Warszawa 2012.
- 3 *Gramatyka współczesnego języka polskiego. Morfologia* pod redakcją Renaty Grzegorzczkovej, Romana Laskowskiego i Henryka Wróbla, tom 1 i 2, Wydawnictwo Naukowe PWN, Warszawa, 1998.
- 4 Mirosław Bańko, *Wykłady z polskiej fleksji*, Wydawnictwo Naukowe PWN, Warszawa, 2002.
- 5 Zygmunt Saloni, *Czasownik polski. Odmiana. Słownik*, Wiedza Powszechna, Warszawa, 2001.
- 6 Stanisław Mędak, *Słownik form koniugacyjnych czasowników polskich*, Universitas, Kraków, 2004.
- 7 Stanisław Mędak, *Słownik odmiany rzeczowników polskich*, Universitas, Kraków, 2003.

Język naturalny to język powstały na drodze rozwoju historycznego, zróżnicowany geograficznie i społecznie, przeciwstawiający się z jednej strony językom sztucznym (jak np. esperanto), z drugiej zaś językom formalnym i językom programowania. Od języków sztucznych różni się polisemią swoich wyrażań oraz tym, że podlega ciągłym zmianom.

Encyklopedia językoznawstwa ogólnego (ze skrótami), Ossolineum 1993

Język naturalny to np. polski, angielski, turecki, arabski, chiński, ale nie C++ czy język predykatów pierwszego rzędu.

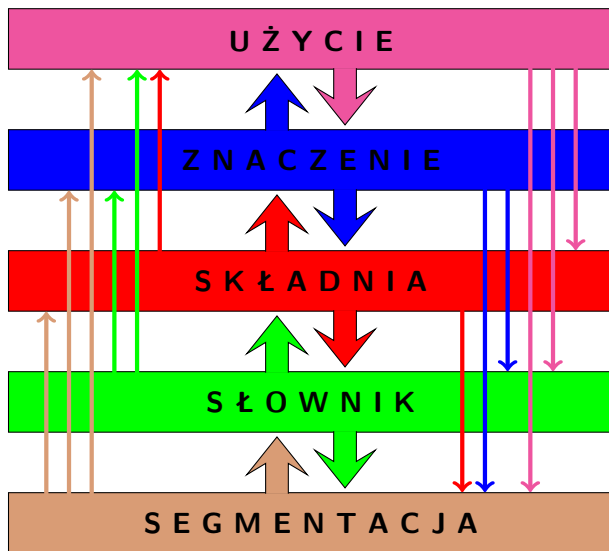
Przetwarzanie języka naturalnego to takie przetwarzanie tekstów (pisanych lub mówionych), które wykorzystuje specyficzne właściwości języka naturalnego.

Liczenie znaków w tekście nie jest przetwarzaniem języka naturalnego.
Liczenie zdań – tak.

Korekta pisowni
Tłumaczenie maszynowe
Wyszukiwanie dokumentów
Odpowiadanie na pytania
Obsługa programu/systemu
Rozstrzyganie autorstwa
Streszczanie
Klasyfikacja tekstów

Język naturalny jest naturalnym formatem przechowywania informacji i komunikowania się ludzi.

Poziomy przetwarzania języka naturalnego



Zbiór tekstów (ang. *corpus*, l.mn. *corpora*) może być **oznaczony** (ang. *tagged*) lub **nieoznaczony** (ang. *untagged*). **Rodzaj oznaczeń** (ang. *markup*) bywa różny, chociaż ostatnio w modzie jest użycie oznaczeń opartych na **XML**. Zbiory tekstów odgrywają kluczową rolę w nowoczesnych systemach przetwarzania języka naturalnego. Pozwalają zbierać różne statystyki, umożliwiają też stosowanie algorytmów uczenia maszynowego. Zbiory oznaczone są dużo bardziej przydatne niż nieoznaczone.

Dla języka angielskiego najbardziej znane zbiory tekstów to zbiór artykułów z czasopisma *Wall Street Journal* (WSJ) i *British National Corpus*. Dla języka polskiego wzorcowym zbiorem tekstów jest Korpus IPI PAN dostępny pod adresem <http://korpus.pl/index.php?page=download>

Segmentacja testu (1/3)

- 1 Co to jest słowo – ciąg liter? Popatrzmy:
 - *cóżeś mi uczynił*
 - *żebyś zdechł*
 - *obym dożył tej chwili*
- 2 Apostrofy:
 - ang. *it's a 'dog', dog's bone, dog's crazy, dogs' house*
 - fr. *qu'est-ce que c'est, aujourd'hui, l'amour, je l'aime*
- 3 Czy słowa połączone myślnikiem tworzą nowe słowo?
 - *W 1900 r. trafił do Niemieckiej Południowo-Zachodniej Afryki.*
 - *Zakład Przemysłowo-Drzewny „Henryków”*
 - *Żydowskie Stowarzyszenie Kulturalno-Oświatowe Tarbut*
 - *SS-man Fuss aresztował Jankiela za sabotaż*
 - *Kazimierz Opel ukrył 6-osobową rodzinę Górskich*
 - *musieli oni nie tylko wykazać się znajomością programu 2-letniej państwowej szkoły elementarnej...*
 - *Dochodząc w opowieści o PRL-u do takiego punktu,...*
- 4 Czy po polsku to jedno, czy dwa słowa?

Segmentacja testu (1/3)

- 1 Co to jest słowo – ciąg liter? Popatrzmy:
 - *cóżeś mi uczynił*
 - *żebyś zdechtł*
 - *obym dożył tej chwili*
- 2 Apostrofy:
 - ang. *it's a 'dog', dog's bone, dog's crazy, dogs' house*
 - fr. *qu'est-ce que c'est, aujourd'hui, l'amour, je l'aime*
- 3 Czy słowa połączone myślnikiem tworzą nowe słowo?
 - *W 1900 r. trafił do Niemieckiej Południowo-Zachodniej Afryki.*
 - *Zakład Przemysłowo-Drzewny „Henryków”*
 - *Żydowskie Stowarzyszenie Kulturalno-Oświatowe Tarbut*
 - *SS-man Fuss aresztował Jankiela za sabotaż*
 - *Kazimierz Opel ukrył 6-osobową rodzinę Górskich*
 - *musieli oni nie tylko wykazać się znajomością programu 2-letniej państwowej szkoły elementarnej...*
 - *Dochodząc w opowieści o PRL-u do takiego punktu,...*
- 4 Czy po polsku to jedno, czy dwa słowa?

1 Gdzie kończy się zdanie? Na kropce?

- *...nie ma prawdy innej, jak cała prawda; to też wszelkie zatajanie jest popełnianiem kłamstwa.*
- *Czy to nasza wina, że mamy takich władców? Myśmy ich sobie nie wybierali! W tysiącletniej afgańskiej historii żaden z władców nie został wyniesiony na tron z woli poddanych.*

2 Na kropce, średniku, wykrzykniku lub na znaku zapytania?

- *W 1885 r. znalazł się Stanach Zjednoczonych, następnie w Wielkiej Brytanii; w 1900 r. w Johannesburgu i Kapsztadzie. W 1900 r. trafił do Niemieckiej Południowo-Zachodniej Afryki. Zmarł prawdopodobnie w Brukseli w 1912 r.*

3 Czy kropka wyłącznie kończy zdanie?

- Skróty (także inicjały), liczby porządkowe (zapisane cyframi)?
- Czy kropka należy do skrótu, czy stanowi odrębny znak?
- Co ze skrótami na końcu zdania?

Kiedy kropka kończy zdanie?

- 1 Pierwsze przybliżenie: kiedy następne słowo zaczyna się wielką literą.
- 2 Ale: po kropce nie może następować znak przestankowy, kropka nie może kończyć skrótu, o którym wiadomo, że nie kończy zdania (wymaga następnego słowa, na ogół nazwy własnej).

Prawidłowe rozpoznawanie końca zdania wymaga rozpoznawania skrótów i nazw własnych oraz oznaczania części mowy, które z kolei wymagają dobrej segmentacji...W językach takich jak japoński lub chiński wyrazy zapisuje się bez odstępów. Jeżeli segmentacji dokonujemy w stosunku do mowy, to na wejściu mamy ciąg głosek...

Dobre wyniki daje podejście skupione na dokumencie. Badamy, w jakim charakterze użyte zostały słowa zakończone kropką w całym dokumencie.

- 1 Gregory Grefenstette, Pasi Tapanainen, *What Is a Word, What Is a Sentence? Problems of Tokenization*, w proceedings of the Third Conference on Computational Lexicography and Text Research COMPLEX'94, Budapest, 1994. Dostępne pod:
http://iling.torreingenieria.unam.mx/curso2002_2/lecturas/mltt-004.pdf.
- 2 Andrei Mikheev, *Periods, Capitalized Words, etc.*, Computational Linguistics Volume 28, Number 3, str. 289-318, September 2002. Dostępne pod:
<http://acl.ldc.upenn.edu/J/J02/J02-3002.pdf>.
- 3 David D. Palmer, Marti A Hearst, *Adaptive Multilingual Sentence Boundary Disambiguation*, Computational Linguistics, Volume 23, Number 2, str. 241-269, June 1994. Dostępne pod:
<http://acl.ldc.upenn.edu/J/J97/J97-2002.pdf>