Studia i Materiały Informatyki Stosowanej, Tom 4, Nr 6, 2012
str. 43-46

ANALIZA ROZMIESZENIA WYRAZÓW W ZDANIACH W CELU DETEKCJI CZASOWNIKÓW

Artur Niewiarowski

Środowiskowe Studium Doktoranckie IPPT PAN - UJ
ul. Reymonta 4, 30-059 Kraków
e-mail: aniewiarowski@pk.edu.pl

Streszczenie: *Artykuł przedstawia analizę wyników próby detekcji czasowników wyprowadzonych przez mechanizm typu text mining, oparty o model cech wyrazów w zdaniach, bazujący na strukturze relacyjnej bazy danych. Podjęta została próba stworzenia mechanizmu wykrywającego czasowniki w oparciu o ich rozmieszczenie statystyczne w zdaniach. W artykule przeanalizowane zostały dokumenty tekstowe polskie i niemieckie, będące felietonami o tematyce z różnych dziedzin życia, w reprezentatywnej liczbie 50 artykułów polskich i 50 artykułów niemieckich.*

Data mining, database mining, text mining, Structured Query Language, natural language.

Analysis of the distribution of word sentences for the purpose of detecting verbs

Abstarct: *This paper presents analysis of the results of detection of verbs deduced by a text-mining mechanism based on the model of the characteristics of words in sentences, based on a relational database structure. Attempt is made to build a mechanism to detect the words based on their statistical distribution in sentences. In article where analyzed Polish and German feuilletons of the various fields of life, in a representative number of 50 Polish articles and 50 German articles.*

Keywords: *Data mining, database mining, text mining, Structured Query Language, natural language.*

3. WSTĘP

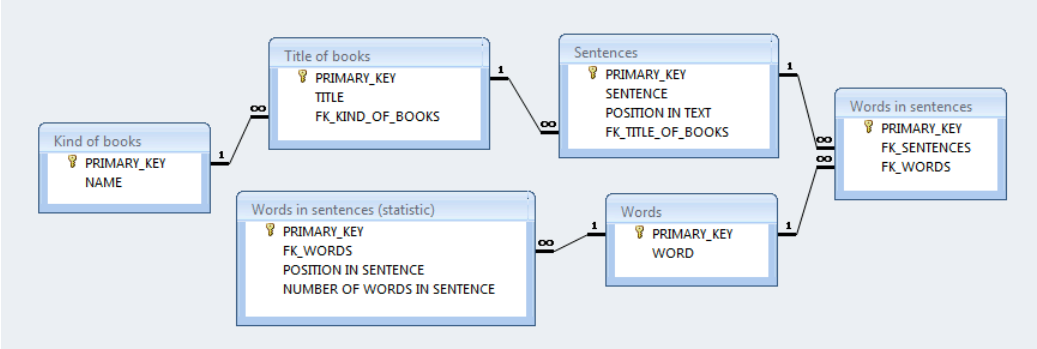
Jednymi z podstawowych projektów realizowanych na całym świecie w pracowniach lingwistyki komputerowej jest stworzenie słowników *wordnet*, w których zawarte są m.in. definicje wyrazów[1]. Ze względu na złożoność problematyki budowy takiego słownika, istnieje wiele metod rozpoznawania wyrazów, które są nieustannie udoskonalane [1][2].

Zadaniem postawionym w ramach powyższych badań jest analiza możliwości stworzenia mechanizmu (metody) syntezy zdań, opartego na badaniu wyłącznie rozmieszczenia wyrazów w zdaniach (tworzących korpus), który umożliwiałby oszacowanie przynależności poszczególnych wyrazów do części mowy z naciskiem na czasowniki. Pozytywny wynik eksperymentu posłużyłby do zbudowania mechanizmu automatycznej (tj. bez znacznej ingerencji człowieka) detekcji części mowy w celu budowy bazy danych słów ze wskazaniem na rodzaj części mowy. Poszczególnym wyrazom przyporządkowane zostały podstawowe właściwości je charakteryzujące, tj. rozmieszczenie w zdaniach. Model zaimplementowany został w relacyjnej bazie danych. Dzięki temu obliczenia

[1] Nazwa jednego z projektów realizowanych w Polsce: *Półautomatyczna konstrukcja zasobów leksykalnych przez rozpoznawanie relacji semantycznych na podstawie danych morfo-syntaktycznych i semantycznych w korpusach tekstu*. Grupa Technologii Językowej Politechniki Wrocławskiej.

statystyczne są łatwiejsze w realizacji, gdyż wykorzystywany jest tutaj potencjał systemu bazodanowego (czytelne rozmieszczenie danych i wbudowane funkcje matematyczne agregujące). W celu analizy skuteczności algorytmu, badania zostały wykonane na bazie dokumentów polskich i niemieckich, ze względu na odmienne zasady budowy zdań w obu językach. Dokumentami poddanymi analizie są felietony reprezentujące różne dziedziny życia, tj. politykę, zdrowie, sport i odkrycia naukowe. Dla języka polskiego przeanalizowanych zostało 11905 unikalnych wyrazów, w 5962 zdaniach, będących częścią 50 felietonów (w sumie 120879 znaków). Dla języka niemieckiego zbadanych zostało 14370 unikalnych wyrazów, w 7380 zdaniach, będących częścią 50 felietonów (w sumie 150020 znaków).

3. MODEL ANALIZY DANYCH



Rysunek 5. Ilustracja modelu rozmieszczenia danych w tabelach z uwzględnieniem relacji**.**

Rysunek 5 przedstawia strukturę tabel odpowiednio przechowujących fragmenty danych tekstowych przeanalizowanych zdań. Tabela *Kind of books* jest słownikiem zawierającym typy (rodzaje) możliwych dokumentów. W celach badań jest to wyłącznie jeden typ – felieton. Tabela ta jest w relacji jeden do wielu z tabelą *Title of books*, która przechowuje wyłącznie tytuły dokumentów. Klucz główny tabeli jest w relacji jeden do wielu z odpowiednim kluczem obcym w tabeli *Sentences*. Tabela *Sentences* przechowuje pełne zdania oraz informację o pozycji zdania względem całego dokumentu. Istotną tabelą w całej strukturze jest również tabela *Words* przechowująca unikalne wyrazy, które odpowiednio są przyporządkowywane poprzez tabelę pośredniczącą *Words in sentences* zdaniom w tabeli *Sentences*. Tabela *Words in sentences (statistic)* przechowuje informacje o każdym wyrazie oraz jego pozycji w zdaniu, jak również liczbie wszystkich wyrazów w zdaniach. Reasumując cechami danego wyrazu są: jego pozycja w zdaniu, liczba wyrazów w zdaniu, w którym się znajduje, pozycja zdania w dokumencie.

3. IMPLEMENTACJA MECHANIZMU[2]

Umieszczenie danych w tabelach polegało na wyodrębnieniu zdań, na bazie znaków podziału (tj. kropek, pytajników, wykrzykników i średników) z uwzględnieniem najważniejszych skrótów zawierających kropki. Fragmenty tekstu ujętego w nawiasach były ignorowane w celu uniknięcia ewentualnych przekłamań budowy zdań.

Na podstawie danych zawartych w tabelach, wygenerowana została tabela wynikowa: *tmp_stat*, przechowująca statystyki rozmieszczenia wyrazów. Kod SQL, na bazie którego wygenerowane zostały wyniki znajduje się poniżej.

```
CREATE TABLE tmp_stat
SELECT
word, --słowo
```

[2] W celu implementacji mechanizmu zastosowany został system bazodanowy MySQL. Tabele działają w oparciu o silnik MyISAM posiadający bardzo szybkie algorytmy przeszukiwania danych.

```
avg_pos_of_words, --poz. uśr.
round(avg_pos_of_words * 100,1), --poz. uśr. w %
pos_1, --min. poz.
pos_1_proc, --min. poz. w proc.
pos_2, --maks. poz.
pos_2_proc, --maks. poz. w proc.
num_of_pos, --liczba próbek
(pos_2 - pos_1) AS diff_maxmin, --różnica maks-min
(pos_2_proc - pos_1_proc) AS diff_maxmin_proc --
różnica maks. - min. w %

FROM
(SELECT *,
(SELECT AVG(pos_in_sent / words_in_sent) FROM
words_in_sentences_statistic WHERE
words_in_sentences_statistic.ID_word =
words.ID_word) avg_pos_of_words,

(SELECT MIN(pos_in_sent) FROM
words_in_sentences_statistic WHERE
words_in_sentences_statistic.ID_word =
words.ID_word) pos_1,

(SELECT MIN(pos_in_sent/words_in_sent)
FROM words_in_sentences_statistic WHERE
words_in_sentences_statistic.ID_word =
words.ID_word) pos_1_proc,

(SELECT MAX(pos_in_sent/words_in_sent) FROM
words_in_sentences_statistic WHERE
words_in_sentences_statistic.ID_word =
words.ID_word) pos_2_proc,

(SELECT MAX(pos_in_sent) FROM
words_in_sentences_statistic WHERE
words_in_sentences_statistic.ID_word =
words.ID_word) pos_2,

(SELECT COUNT(pos_in_sent) FROM
words_in_sentences_statistic WHERE
words_in_sentences_statistic.ID_word =
words.ID_word) num_of_pos

FROM words ORDER BY avg_pos_of_words) C;
```

Kod SQL jest złożony i zawiera wiele podzapytań i funkcji agregujących. Wykonanie kodu na bardzo dużej liczbie danych pomiarowych jest czasochłonne[3].

[3] Wygenerowanie wyników w oparciu o bazę ok. 6000 zdań, na maszynie o procesorze 8 – rdzeniowym i 24 GB pamięci RAM trwało ok. 30 min.

3. WNIOSKI

Fragment tabeli *tmp_stat* dla języka polskiego znajduje się poniżej.

słowo:	poz. uśr:	poz. uśr. w %	min. poz.:	min. poz. w proc.:	maks. poz.:	maks. poz. w proc.:	liczba próbek:	różnica max-min:	różnica max-min w %:
.	1.00000000	100%	0	0%	31	100%	2284	31	0%
w	0.44932974	44.9%	1	3.7%	28	93.3%	1091	27	89.6%
i	0.48006514	48%	1	5%	26	91.7%	846	25	86.7%
nie	0.42971790	43%	1	3.3%	27	91.7%	754	26	88.3%
się	0.52816117	52.8%	2	8.7%	29	96.7%	674	27	88%
na	0.47564423	47.6%	1	5.9%	22	90.9%	670	21	85%
z	0.49931408	49.9%	1	4.8%	22	91.3%	595	21	86.5%
że	0.15654999	15.7%	1	3.2%	14	77.8%	548	13	74.6%
to	0.40002906	40%	1	4.8%	21	92.9%	542	20	88.1%
do	0.52707101	52.7%	1	4.4%	22	89.5%	346	21	85.1%
jest	0.47119687	47.1%	1	9.1%	17	88.9%	300	16	79.8%

Fragment tabeli *tmp_stat* dla języka niemieckiego znajduje sie poniżej.

słowo:	poz. uśr:	poz. uśr. w %	min. poz.:	min. poz. w proc.:	maks. poz.:	maks. poz. w proc.:	liczba próbek:	różnica max-min:	różnica max-min w %:
.	1.00000000	100%	0	100%	42	100%	2951	42	0%
die	0.34546935	34.5%	1	2.9%	33	92.6%	2141	32	89.6%
der	0.44066629	44.1%	1	2.4%	38	94.9%	1893	37	92.5%
und	0.46709360	46.7%	1	2.4%	35	93.9%	1330	34	91.6%
in	0.42802904	42.8%	1	3.7%	30	90.3%	923	29	86.6%
das	0.35232031	35.2%	1	4.8%	35	92.3%	760	34	87.6%
den	0.43931603	43.9%	1	4.2%	31	93.9%	605	30	89.8%
von	0.49117399	49.1%	1	3.9%	34	91.9%	590	33	88%

Tabele zawierają następujące kolumny: badane słowo, uśredniona pozycja wyrazu po analizie wszystkich zdań podana w postaci liczby oraz w postaci procentowej, minimalna pozycja w zdaniu wyrazu wyrażona w postaci liczby oraz procentowej względem danego zdania, maksymalna pozycja, jaką osiągnął wyraz w postaci liczby i w postaci procentowej względem danego zdania, w którym wystąpił, liczba próbek wyrazów zebranych ze wszystkich zdań (im większa, tym dokładniejsze są wyniki) oraz różnice pomiędzy minimalną wartością pozycji wystąpienia a wartością maksymalną.

Dodatkowo dla celów analizy, w słowo wliczone zostały również znaki interpunkcyjne. Jak wynika z obliczeń, kropka zajmowała zawsze ostatnią pozycję w zdaniu. Poniższe tabele przedstawiają dane posortowane od wyrazu najczęstszego do najrzadszego na podstawie liczby próbek.

Wyniki badań wskazują, że oszacowanie części mowy w języku polskim (w tym z naciskiem na czasowniki) nie jest możliwe na bazie wyłącznie tego typu modelu i zwykłej statystyki wystąpień wyrazów. Pojawia się zbyt duża rozbieżność pozycji danych wyrazów w zdaniach. Podobna sytuacja tyczy się języka niemieckiego. Pomimo znanych reguł budowy zdań w języku niemieckim, mówiących o tym, że orzeczenie jako część zdania jest na drugim miejscu w zdaniu, to m.in. zjawisko rekcji[4] czasownika, uniemożliwia bez dodatkowego rozbioru logicznego zdań, oszacowania na bazie modelu cech wyrazów.

3. PODSUMOWANIE

Wyniki badań pokazują, że nie jest możliwe utworzenie bazy danych słów z uwzględnieniem części mowy, na bazie wyłącznie rozmieszczenia wyrazów w korpusie zdań. Natomiast przedstawiony model może znaleźć zastosowanie w analizie interakcji wyrazów w sąsiednich zdaniach, jak również w celu obliczania i zapamiętywania wspólnych wystąpień wyrazów w zdaniach, ponieważ model zawiera informacje o rozmieszczeniu zdań względem siebie (a w nich poszczególnych wyrazów). W połączeniu z popularnymi metodami grupowania danych tekstowych (opartych o odległości pomiędzy wektorami zważonych terminów), w których nie jest brana pod uwagę kolejność występowania terminów (czyli wyrazów i grup wyrazowych) [3], można zbudować mechanizm podpowiedzi wprowadzanych tekstów, na wzór tych stosowanych w popularnych wyszukiwarkach internetowych (np. Google, Yahoo); jak również zastosować w algorytmach analizujących podobieństwo dokumentów tekstowych lub tłumaczących teksty na języki obce [4][5].
Istnieje również możliwość implementacji modelu w celu dedukcji części zdań przy użyciu rachunku prawdopodobieństwa (modele probabilistyczne w NLP). Techniki związane z rachunkiem prawdopodobieństwa (m.in. prawdopodobieństwa warunkowego, sieci Bayesa) są często używane do przetwarzania języka naturalnego (w tym do badania konstrukcji zdań i przewidywania wystąpień wyrazów). Zastosowanie tego typu metod i modeli, to m.in. rekonstrukcja niekompletnych (uszkodzonych) tekstów [6][7].

[4] Zjawisko dla czasowników wymagających dopełnienia w odpowiednim przypadku lub połączenia z dopełnieniem za pomocą przyimka.

Literatura

1. Broda B., Derwojedowa M., Piasecki M., Szpakowicz S. Corpus-based Semantic Relatedness for the Construction of Polish WordNet. Proceedings of the 6th Language Resources and Evaluation Conference (LREC'08), 2008.
2. Maziarz M., Piasecki, M., Szpakowicz S. 2012. "Approaching plWordNet 2.0". Proceedings of the 6th Global Wordnet Conference. Matsue, Japan, January 2012.
3. Piasecki M., Broda B. "Semantic similarity measure of Polish nouns based on linguistic features". Business Information Systems 10th International Conference, Poznań, Lecture Notes in Computer Science, vol. 4439, Springer, 2007.
4. Kao Anne, Poteet, Steve R. (Editors). "Natural Language Processing and Text Mining". Springer, ISBN 184628175X.
5. Doddington, G. "Automatic evaluation of machine translation quality using n-gram co occurrence statistics". Proceedings of the Human Language Technology Conference (HLT), San Diego, 2002, CA pp. 128—132.
6. George A. Miller, Elizabeth A. Friedman. The reconstruction of mutilated english texts. Harvard University, Cambridge, Massachusetts, USA, 1957. Elsevier.
7. Manning C., Schutze H. (1999). "Foundations of Statistical Natural Language Processing". Cambridge, MA: MIT Press. ISBN 978-026213360