Intelligent Information Systems
ISBN 978-83-7051-580-5, pages 91–98

Extracting Semantic Knowledge from Wikipedia

Jan Wicijowski and Bartosz Ziółko

Department of Electronics
AGH University of Science and Technology
Kraków, Poland
{wici,bziolko}@agh.edu.pl
www.dsp.agh.edu.pl

Abstract

A semantic model of Polish used as a last level validation of hypotheses in an automatic speech recognition system is described. Semantic modelling is a key technique to improve automatic speech recognition of Polish which is a non-positional language. However, a vast amount of text and efficient methods are necessary to train a useful model. Wikipedia archives were used to develop and test a semantic analysis method. Wikipedia archives are stored in a simple XML with special mediawiki markup describing structure and formatting of articles. This markup allows us to set the boundaries of topics precisely and on a few levels: article, section, paragraph and sentence.

Keywords: Polish, corpora collecting, semantic analysis, natural language processing, speech recognition

1 Introduction

The automatic speech recognition (ASR) system as a whole consists of three separated levels: an acoustic recogniser, grammatical validator and semantic validator. The input to the final, semantic part of the system, which is described in this paper, are hypotheses of word sequences of arbitrary length, typically sentences. The output of the semantic model validation is a real number, which measures a likelihood that a given sequence of words could appear in the language.

The model of the language we have chosen is based on vector space analysis as shown by Salton *et al.* (1975). The grammatical structure of the language is approximated in an underlying layer by n-gram model—most notably, the order of words is extensively covered. The semantic model ignores the word order completely and operates on chunks of written texts, so it fits into the category of bag-of-words models.

The human perception system is based on catching context, structure and understanding combined with recognition procedure. One of the most popular methods of applying it in automatic speech processing is latent semantic analysis (LSA) (Bellegarda, 1997). Its philosophy is that the meaning of a small part of text, like a paragraph or a sentence, can be expressed by the sum of the meanings of its words. LSA uses a word-paragraph matrix which describes the occurrences

of words in topics. It is a sparse matrix with rows that correspond to topics and columns that correspond to words appearing in the topics. The elements of the matrix are proportional to the number of times the words appear in each document, where rare words are upweighted to reflect their relative importance.

LSA has found already a few applications. One of them is automatic essay and answers grading (Kakkonen *et al.*, 2006; Kanejiya *et al.*, 2003). LSA can be also used in modelling global word relationships for junk e-mail filtering or pronunciation modelling (Bellegarda, 70–80). Another possible application is for word completion (Miller and Wolf, 2006). LSA can be combined with the n-gram model (Coccaro and Jurafsky, 1998) or maximum entropy model (Deng and Khudanpur, 2003). There are other methods of analysing semantic information, like topic signatures (Agirre *et al.*, 2004) and maximum entropy language models (Khudanpur and Wu, 1999). The idea of topic signatures is to store concepts in context vectors. There are simple methods to automatically acquire for any concept hierarchy i.e. they were used to approximate link distances in WordNet.

2 Wikipedia structure and its processing

Polish Wikipedia (Wikipedia, 2001–) was chosen as the source of broad-range language, due to its rich semantic content, structured, hierarchical design, and relatively high quality of language (Ziółko *et al.*, 2010).

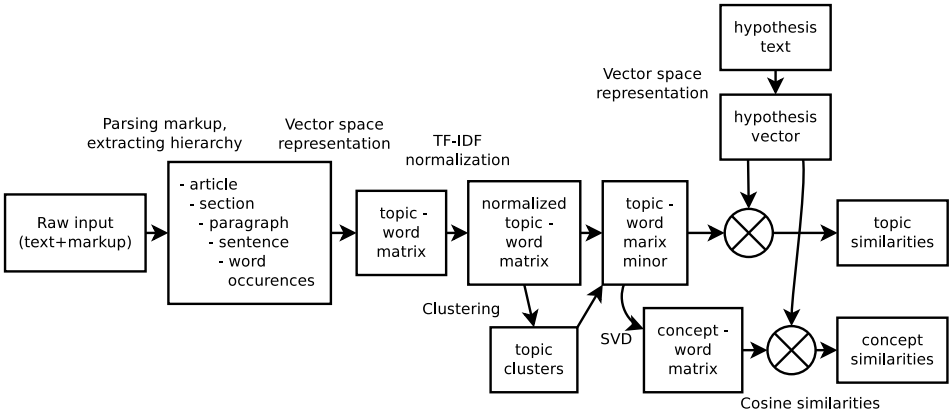

FIGURE 1: Text corpus processing pipeline.

Figure 1 presents the processing pipeline, which transforms both the corpus text and the text hypotheses into vector space model. First, the data is acquired from a database dump; it is available for download and is the method of dissemination of whole corpus suggested by its maintainers (Wikipedia, 2010b). The dump consists of relatively flat XML structure with separate articles formatted with mediawiki markup. The markup is very rich, and it's grammar complexity exceeds virtually all programming language grammars. Is spite of numerous efforts of formalization, the markup is only defined by the official engine implementation (Wikipedia, 2010a). Due to the inherent complexity of the formatting, which is

intermingled with the language data, there was the need to resort to third-party parser (PediaPress, 2007) to filter out the formatting.

The raw text is then parsed into hierarchical structure: articles, sections, paragraph and sentences. A topic-word matrix is constructed, on the base on a selected level of decomposition. The topic, as used afterwards in the text, is one of the hierarchical entities (consequently: sentence, paragraph, section and article), and the words are separate, lemmatised lexical entities. The cells of the constructed matrix carry the number of appearances of a word in the topic.

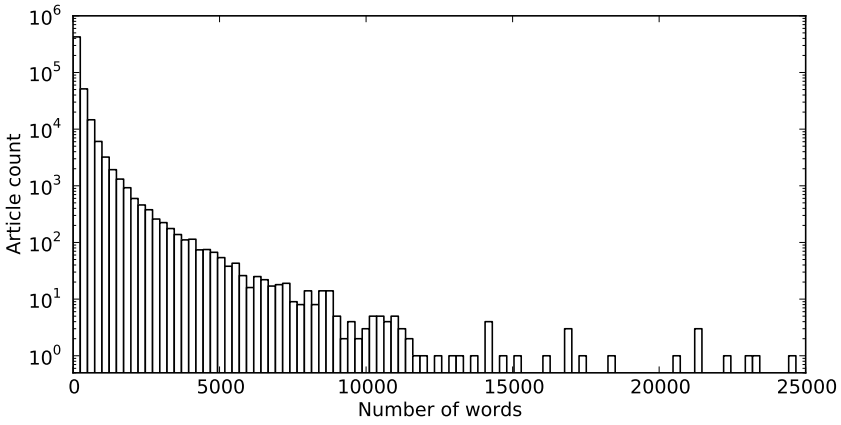

FIGURE 2: Histogram of Wikipedia articles sizes measured in word count.

3 Topic-word matrix denoising

A key issue to tackle, was to remove obscure words and topics, which we considered insignificant even for a broad topic range speech recognition. Wikipedia texts contain a lot of rare words, most commonly proper names, foreign words, and errors. About 54% of words occurred in single article, which follows the empirical Zipf's law (Kanter and Kessler, 1995). Among them we can find such strings as "printfie", "fimbriata", "shchedrin", "eksperyzy", "pggg", "euclidiinae", "benzodiazepinowego", "yzf". They are either not Polish words or very rare or obscure, and are not included in further calculations, as their appearance is so infrequent, that they cannot commit to the statistical model of a language—they are treated as noise.

The denoising process was conducted orthogonally on the articles. We have filtered out the articles with high probability of containing large amount of unstructured text, like fact tables, list, numerical data, etc. In the parsing stage vast majority of such articles were recognised by the conventional Wikipedia prefix, like "Kategoria:" or "Wikipedia:". Nevertheless, a significant proportion of similar articles were untouched—they were recognised later by the assumption, that table markup usually contains raw data. This proved successful, as 5.5% prevailing

articles were filtered, with a threshold on the non-tabular text set as minimum 20 words.

4 Normalisation

The matrix is then normalised by means of term frequency and inverse document frequency (TF-IDF) weighing (Salton and Buckley, 1988), which takes into account the topic length, as well as the popularity of a word in the corpus. The values of the cells in the transformed matrix correspond to word frequencies. The normalization scheme we have picked is *tfidf* - ltc, which was proven to yield good performance by many researchers, as described by Liu *et al.* (2009).

5 Topics clustering

In order to reduce the search space to interesting words, a clustering of topics is performed. It is conveyed using k-means vector quantisation (VQ) method. The parameters of the clustering process can vary, ranging from very imprecise, erroneous clustering, to a precise one. K-means time complexity is $O(n)$, where n is the number of samples to cluster (Mercer, 2003), so it fits the domain of corpus processing. It has also proven to yield high-quality results, comparable to SVD one (Dhillon and Modha, 2001). One may wish the clustering groups to be as specific as possible. Nevertheless, in the project it was discovered, that rough and fast clustering, which converges fast to suboptimal solution is sufficient to separate the topics with distinct meaning, as measured by human acceptance. The clustering method uses comprise of classical k-means, k-means with fractorization (Cutting *et al.*, 1992) and Learning Vector Quantization (Kohonen, 1998).

6 Concept space transformation

The system employs the transformation of topic-word vector space into concept-word space, which is achieved by singular value decomposition (SVD). The full SVD of the topic-word matrix is impossible to calculate using the available resources, as its algorithmic complexity is $O(mn^2)$ where m and n are the dimensions of the matrix. We employ truncated SVD calculations, in which only a specified number of singular vectors are kept. The SVD process is roughly described as to simultaneously group the topics of similar meaning and create a mapping between words and the topic groups—the concepts. It is further explained in numerous articles (Berry *et al.*, 1995; Husbands *et al.*, 2001).

7 Speech hypotheses reordering

The hypotheses of speech come in a form of a text. They are processed in analogous way to the corpus texts. The word indices and IDF weights are taken from the corpus. The achieved vector can be then compared with the topic-word matrix, or with the concept-word matrix. The comparison is achieved by calculating of cosine distance between vector hypotheses and concept/topic vectors which form

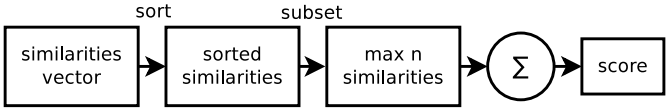

FIGURE 3: Text hypothesis validation.

the matrix. As a result, we obtain a vector of similarities of the hypothesis to the semantic entities.

A method of hypothesis scoring is to take n highest elements from the similarities vector and sum them. The resultant score can be then compared to the scores of other hypotheses. Good results are obtained with n equal 1 or 3 with both topic-word matrix and concept-word matrix. We devised a few hypotheses of sentences, that were not included in training data. The consecutive hypotheses, their mistaken forms and scores are presented in Table 1. The scores are measured as the sum of 3 largest elements of topic-hypothesis similarities vector, with the matrices constructed from 4% sample of Wikipedia articles and the topic on the sentence level. The concept space validation using truncated SVD was limited to a few thousands most frequent words and does not yield valuable examples.

8 Results

The correct sentences are typically semantically coherent while wrong recognitions are often not. This knowledge can improve speech recognition if a good semantic model is applied on a list of hypotheses of a particular sentence. Table 1 shows some examples of how semantically reasonable sentences score higher then wrong hypotheses. The ratio between semantic score between a correct hypotheses and a wrong one oscillates between 3.33 and 1.04. Sometimes the score values are similar but the correct sentence is scored higher in all examples we have produced. Our experiment supports an opinion that semantic analysis can improve speech recognition in many cases.

9 Software used

The database back-end used to store the matrices is the public-domain SQLite3, due to its simplicity and speed. The computational part of the system is build upon NumPy (Jones *et al.*, 2001–) project with Fortran LAPACK library (Anderson *et al.*, 1999) and SVDLIBC (Rohde, 2004–05). A lemmatizer was build upon the Morfeusz morphological analyser library (Woliński, 2006). Mediawiki markup is parsed by mwlib (PediaPress, 2007) library. Python scripts are used as a glue between these low level libraries, serving as a rapid application development environment.

TABLE 1: Scores of exemplary hypotheses scores measured as the sum of 3 largest elements of topic-hypothesis similarities. Correct sentences go first, followed by wrong sentences in which one word was replaced with an acoustically similar one. *Italics* are used to stress the wrong words. Original sentences are in Polish, with English translations in brackets.

Hypothesis	Score
zrobiłem (I made) pranie (washing) brudnych (dirty) fartuchów (aprons)	**1.20**
zrobiłem *branie (taking)* brudnych fartuchów	0.36
na (on) terenie (the area) tym (this) dotychczas (so far) zaobserwowano (observed) występowanie (existence) 39 gatunków (species) ssaków (mammals) w tym (including) aż (as many as) 11 drapieżnych (predators)	**2.69**
na terenie tym dotychczas zaobserwowano występowanie 39 gatunków *saków (sacks, old fashioned)* w tym aż 11 drapieżnych	2.58
na terenie tym dotychczas *zaabsorbowano (absorbed)* występowanie 39 gatunków ssaków w tym aż 11 drapieżnych	2.56
wczoraj (yesterday) poszłam (I went) do (to) koleżanki (a friend)	**0.78**
wczoraj *po szlam (for ooze)* do koleżanki	0.75
kocham (I love) cię (you) kochanie (a love) moje (of mine) to (this) rozstania (breaks) i (and) powroty (returns)	**1.80**
kocham cię kochanie moje *gronostaja (ermine)* i powroty	1.70
zygmunt III waza (proper name of a king, also a vase) pierwszy (first) władca (ruler) wywodzący się (coming) ze (from) szwedzkiej (swedish) dynastii (dynasty) wazów (family name)	**1.65**
zygmunt III waza pierwszy *radca (counsel)* wywodzący się ze szwedzkiej dynastii wazów	1.60
zygmunt III waza *piersi (breast)* władca wywodzący się ze szwedzkiej dynastii wazów	1.56
zygrfyd (proper name) III waza pierwszy władca wywodzący się ze szwedzkiej dynastii wazów	1.56

10 Conclusions

Wikipedia is a very comprehensive and valuable source of freely available semantic knowledge. It is provided in a format which clearly distinguishes between consecutive articles, sections and paragraphs. It allows to choose a scope of text data analysis precisely and treat all three types of subdivisions as topic representation. The fourth option is to use ends of a sentence as the change of a topic, which is easily implementable. Applying semantic analysis on Wikipedia data gives a model which reorders speech recognition hypotheses in a way that correct hypotheses move upwards.

11 Acknowledgements

This work was supported by MNISW grant OR00001905.

References

E. Agirre, E. Alfonseca, and O. Lopez de Lacalle (2004), Approximating hierarchy-based similarity for WordNet nominal synsets using Topic Signatures, *Proceedings of the 2nd Global WordNet Conference. Brno, Czech Republic.*

E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen (1999), *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, ISBN 0-89871-447-8 (paperback).

J. R. Bellegarda (1997), A latent semantic analysis framework for large-span language modeling, *Proceedings of Eurospeech*, 3:1451–1454.

J. R. Bellegarda (70–80), Latent semantic mapping, *IEEE Signal Processing Magazine*, September:70–80.

Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien (1995), Using linear algebra for intelligent information retrieval, *SIAM Rev.*, 37(4):573–595, ISSN 0036-1445, doi:http://dx.doi.org/10.1137/1037127.

N. Coccaro and D. Jurafsky (1998), Towards better integration of semantic predictors in statistical language modeling, *Proceedings of ICSLP-98, Sydney.*

Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey (1992), Scatter/Gather: a cluster-based approach to browsing large document collections, in *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 318–329, ACM, New York, NY, USA, ISBN 0-89791-523-2, doi:http://doi.acm.org/10.1145/133160.133214.

Y. Deng and S. Khudanpur (2003), Latent semantic information in maximum entropy language models for conversational speech recognition, *Proceedings of the HLT-NAACL 03*, pp. 56–63.

Inderjit S. Dhillon and Dharmendra S. Modha (2001), Concept decompositions for large sparse text data using clustering, *Mach. Learn.*, 42(1/2):143–175, ISSN 0885-6125, doi:http://dx.doi.org/10.1023/A:1007612920971.

P. Husbands, H. Simon, and C. H. Q. Ding (2001), On the use of the singular value decomposition for text retrieval, pp. 145–156.

Eric Jones, Travis Oliphant, Pearu Peterson, *et al.* (2001–), SciPy: Open source scientific tools for Python, URL `http://www.scipy.org/`.

T. Kakkonen, N. Myller, and E. Sutinen (2006), Applying part-of-speech enhanced LSA to automatic essay grading, *Proceedings of the 4th IEEE International Conference on Information Technology:Research and Education (ITRE 2006). Tel Aviv, Israel*, pp. 500–504.

D. Kanejiya, A. Kumar, and S. Prasad (2003), Automatic evaluation of students' answers using syntactically enchanced LSA, *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, 2:53–60.

I. Kanter and D. A. Kessler (1995), Markov Processes: Linguistics and Zipf's Law, *Phys. Rev. Lett.*, 74(22):4559–4562, doi:10.1103/PhysRevLett.74.4559.

S. Khudanpur and J. Wu (1999), A maximum entropy language model integrating n-grams and topic dependencies for conversational speech recognition, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Phoenix, AZ.*

Teuvo Kohonen (1998), Learning vector quantization, pp. 537–540.

Ying Liu, Han Tong Loh, and Aixin Sun (2009), Imbalanced text classification: A term weighting approach, *Expert Syst. Appl.*, 36(1):690–701, ISSN 0957-4174, doi: http://dx.doi.org/10.1016/j.eswa.2007.10.042.

Daniel Mercer (2003), Clustering large datasets, Technical report, Linacre College.

T. Miller and E. Wolf (2006), Word completion with latent semantic analysis, *18th International Conference on Pattern Recognition, ICPR, Hong Kong*, 1:1252–1255.

PediaPress (2007), mwlib MediaWiki parsing library, URL `http://code.pediapress.com`.

Doug Rohde (2004–05), SVDLIBC, URL `http://tedlab.mit.edu/~dr/SVDLIBC`.

G. Salton and C. Buckley (1988), Term-weighting approaches in automatic text retrieval, in *Information Processing and Management*, pp. 513–523.

G. Salton, A. Wong, and C. S. Yang (1975), A vector space model for automatic indexing, *Commun. ACM*, 18(11):613–620, ISSN 0001-0782, doi:http://doi.acm.org/10.1145/361219.361220.

Wikipedia (2001–), Wikipedia, wolna encyklopeida, URL `http://pl.wikipedia.org`.

Wikipedia (2010a), MediaWiki — Wikipedia, The Free Encyclopedia, URL `http://en.wikipedia.org/w/index.php?title=MediaWiki&oldid=357436575`, [Online; accessed 22-April-2010].

Wikipedia (2010b), Wikipedia:Database download — Wikipedia, The Free Encyclopedia, URL `http://en.wikipedia.org/w/index.php?title=Wikipedia:Database_download&oldid=357397182`, [Online; accessed 22-April-2010].

M. Woliński (2006), Morfeusz – a practical tool for the morphological analysis of Polish, in *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*, pp. 503–512, Springer.

B. Ziółko, D. Skurzok, and M. Ziółko (2010), Word N-Grams for Polish, *The Tenth IASTED International Conference on Artificial Intelligence and Applications, AIA 2010.*