Discovering Distributional Thesauri Semantic Relations

Velislava Stoykova

Institute for Bulgarian Language, Bulgarian Academy of Sciences,
52, Shipchensky proh. str., bl. 17, 1113 Sofia, Bulgaria
vstoykova@yahoo.com

Abstract. The paper[1] presents technique and analysis to discover distributional thesauri relations by using statistical similarity of different word's contexts. The application uses educational electronic text corpus and the Sketch Engine software statistical search to extract and compare word's collocations from the related text corpus. The semantic search used is based on the evaluation and comparison of common keyword's collocations by generation distributional thesauri word's semantic relations and words sketch differences. The results of the related search experiments for British Academic Spoken English corpus are evaluated and presented.

Keywords: Data Mining, Big Data, Hierarchical Categorization.

1 Introduction

The similarity search is widely known technique to extract semantically related words. It is used to evaluate not only synonyms but also to extract semantic relations between words in large electronic text corpora. Recent research in that area [1] extend the search techniques by combining semantic approaches and information retrieval approaches improving the search, so to deal with more complex semantic representations.

Thus, the technique is applied to evaluate semantic content of retrieved electronic textual documents by systematic analysis of structure of that documents. Additionally, the traditional approaches were improved with the technique of linking text-based content to image using joint information sources [2,3] for document classification.

Generally, the combined statistical similarity approaches were successfully applied for extracting and comparing words belonging to different thesauri by comparing their related contextual collocations. The existing applications improve the multilingual use and the universal scope of that approach [4].

Further, we are going to demonstrate the use of such technique by presenting and analyzing search results of generation and comparison of collocations

[1] The research described presents results obtained during COST-STSMIC1302-36988 "Natural Language Processing Keyword Search for Related Languages" of COST Action IC1302 "Semantic keyword-based search on structured data sources (KEYSTONE)".

of related words in British Academic Spoken English corpus using specialized software of Sketch Engine.

2 The Sketch Engine (SE)

The SE software [5] allows approaches to extract semantic properties of words and most of them are with multilingual application. Extracting keywords is widely used technique to extract terms of particular studied domain. Also, semantic relations can be extracted by generation of related word contexts through word concordances which define context in quantitative terms and a further work is needed to be done to extract semantic relations by searching for co-occurrences and collocations of related keyword.

Co-occurrences and collocations are words which are most probably to be found with a related keyword. They assign the semantic relations between the keyword and its particular collocated word which might be of similarity or of a distance. We use techniques of $T-score$, $MI-score$ and $MI^3-score$ for corpora processing and searching. For all, the following terms are used: N – corpus size, f_A – number of occurrences of keyword in the whole corpus (the size of concordance), f_B – number of occurrences of collocated keyword in the whole corpus, f_{AB} – number of occurrences of collocate in the concordance (number of co-occurrences). The related formulas for defining $T-score$, $MI-score$ and $MI^3-score$ are as follows:

$$\text{MI-Score} \quad \log_2 \frac{f_{AB}N}{f_A f_B}$$

$$\text{T-Score} \quad \frac{f_{AB} - \frac{f_A f_B}{N}}{\sqrt{f_{AB}}}$$

$$\text{MI}^3\text{-Score} \quad \log_2 \frac{f_{AB}^3 N}{f_A f_B}$$

The $T-score$, $MI-score$ and $MI^3-score$ are applicable for processing multilingual parallel corpora as well. Collocations have been regarded as statistically similar words [6] which can be extracted by using techniques for estimation the strength of association between co-occurring words.

The SE also offers further refinement of extracted semantic relations by evaluation of word's common collocations or evaluation of distributional thesauri semantic relations [7].

Further, we are going to present and analyze search results of extraction and evaluation of common collocations from British Academic Spoken English corpus using the SE software.

3 The British Academic Spoken English (BASE) corpus (BASE)

The British Academic Spoken English (BASE) corpus is a collection of transcripts of lectures and seminars recorded at University of Warwick and University of Reading in the UK during the period 1998-2005. It was created to analyze English for Academic Purposes [8] which allows also extraction of specific semantic relations between terms and definitions among subjects studied.

The texts included consist of 1 186 290 words and are distributed across four broad domain areas: (i) Arts and Humanities, (ii) Life and Medical Sciences, (iii) Physical Sciences and (iv) Social Studies and Sciences. The corpus is annotated according to Text Encoding Initiative Guidelines and recently was uploaded into SE allowing the use of its incorporated options for storing, sampling, searching and filtering texts according to different criteria.

4 Common Collocations Search Results

We are going to use the SE statistical options to extract word's semantic relations. The methodology includes generation of word's collocations [9] and their further comparison. For that, we are going to present search experiments and related results for the word *politics*.

Collocation candidates

Page 1 Go Next >

	Cooccurrence count	Candidate count	T-score	MI	MI3	log likelihood
P | N **electoral**	3	8	1.732	11.946	15.116	45.060
P | N **international**	17	192	4.119	9.864	18.039	202.593
P | N **politics**	10	116	3.159	9.825	16.469	117.978
P | N **race**	3	54	1.729	9.191	12.361	32.481
P | N **gender**	3	71	1.728	8.796	11.966	30.801
P | N **aspect**	3	80	1.728	8.624	11.794	30.072
P | N **influence**	5	138	2.230	8.575	13.219	49.856
P | N **Thompson**	3	99	1.727	8.317	11.487	28.775
P | N **analysis**	5	311	2.223	7.402	12.046	41.659
P | N **involved**	3	232	1.719	7.088	10.258	23.637
P | N **key**	3	260	1.718	6.924	10.094	22.954
P | N **interested**	3	281	1.717	6.812	9.982	22.489
P | N **power**	4	379	1.982	6.795	10.795	29.928
P | N **class**	3	295	1.716	6.742	9.912	22.199
P | N **book**	3	360	1.712	6.454	9.624	21.011
P | N **society**	3	365	1.712	6.434	9.604	20.928
P | N **local**	3	378	1.711	6.384	9.554	20.720
P | N **ideas**	3	412	1.709	6.260	9.430	20.207
P | N **European**	3	418	1.709	6.239	9.409	20.122
P | N **history**	5	794	2.202	6.050	10.694	32.325

Fig. 1. The collocation candidates of word *politics* from BASE corpus.

For our research, we use $MI - score$ and apply methodology already used to extract specialized collocations in mathematical domain [10]. Fig. 1 shows generated collocation candidates for the word *politics*. The received results present most frequent words which are most probably to be found with the word *politics*. They are: *electoral*, *international*, *gender*, etc.

The results include specialized terms that can be part of thesauri like *electoral politics*, *international politics* but also attributive collocations like *confrontational politics* which are based on the meaningful combination between word and its collocations [11].

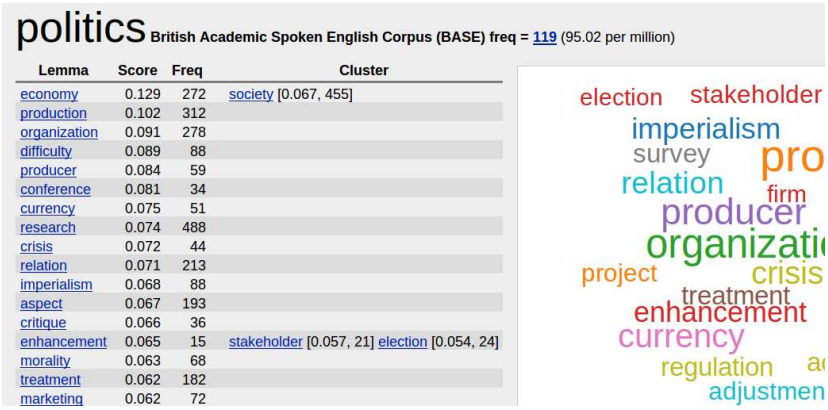

Fig. 2. The generated distributional thesauri and related semantic relations of word *politics* from BASE corpus.

The SE word search options can extract not only statistically similar words for building thesauri but also can compare words' collocations which belong to more than one thesauri – the so-called *common collocations*. Distributional thesauri search evaluates word's common collocations which share common semantic relations.

Generally, if two words have much collocations in common, they share semantic relations of *distributional thesauri* and will appear in each one's thesauri. The SE function to generate words' distributional thesauri can compare pairs of words and show how they collocate.

Fig. 2 shows generated distributional thesauri for the word *politics* and its semantically related words like: *economy*, *society*, *production*, *organization*, etc.

However, the results underlay two clusters of semantically similar words. The words from the first cluster present semantic relations of word *politics* and relate it with words *economy* and *society* by having common collocations which are not presented directly. The results, also, include words belonging to one and the same part-of-speech category which relate semantically.

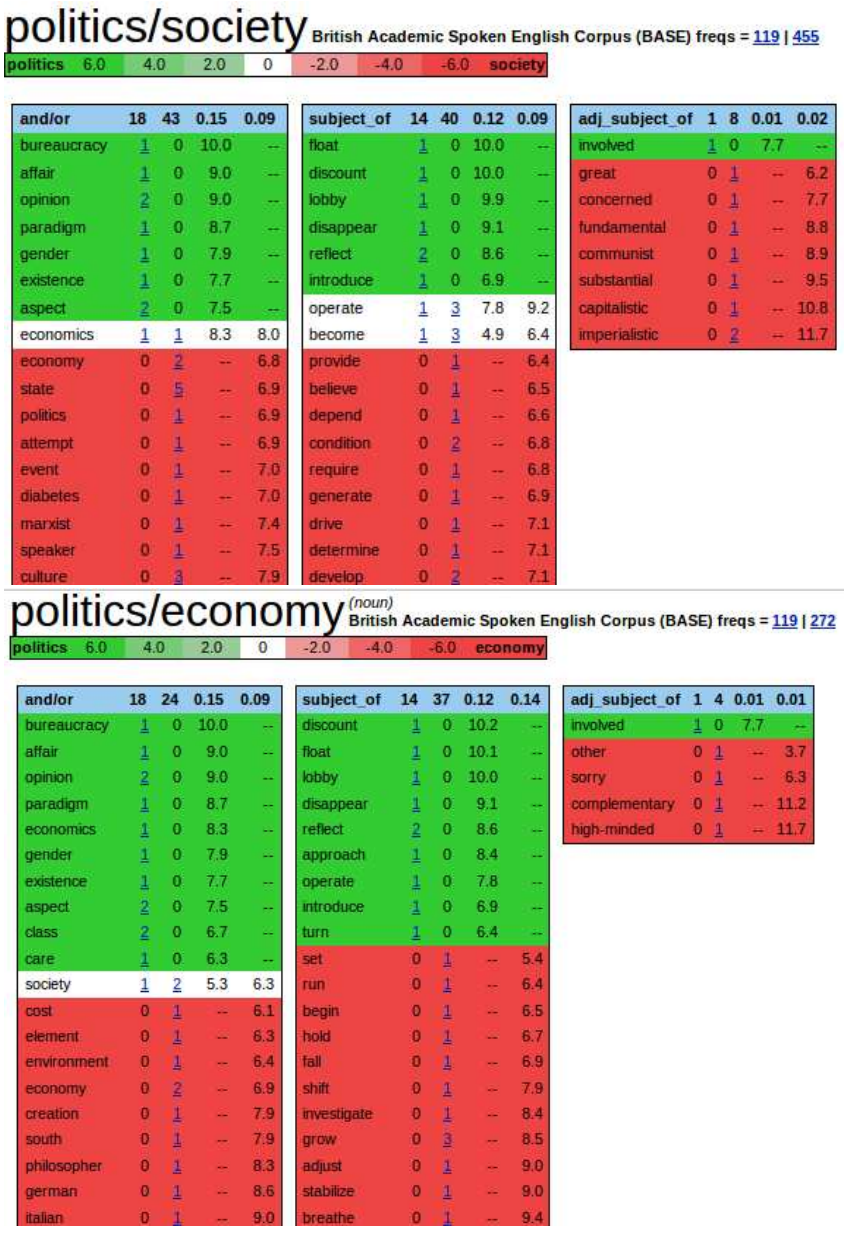

Fig. 3. The words sketch differences of pair words *politics*/*society* and *politics*/*economy* from BASE corpus.

The hidden semantic relations between words *politics* and *society* can be evaluated by generating their common words *sketch differences*, and using them to compare and contrast the two words by analyzing their collocations and by displaying their collocates. The results are presented at Fig. 3 and include relational semantic properties of words *politics* and *society* divided into categories based on grammatical relations.

Generally, the received results contain relations like and/or, $subject-of$, etc. Within them, the displayed words share common collocations with related keywords. Only the relation and/or gives as a result the same part-of-speech words ranked according to their statistical weight. It lists words like *bureaucracy*, $affair$, *opinion*, etc. which also semantically relate to word *politics*. The same relation connect words *state*, *speaker*, etc. which semantically relate to word *society*.

At the same time, the word *economics* and the word *economy* have similar weight (appear in the results for both *politics* and *society*), and are regarded as common collocations of both keywords which share with them hidden semantic relations (and/or).

The relation $subject-of$ lists words $float$, *discount*, *lobby*, etc. semantically related to word *politics*, and words *provide*, *believe*, *depend*, $require$, etc. semantically related to word *society*. However, the words *operate* and *become* have similar weight and are common collocations to both *politics* and *society* expressing their hidden semantic relations ($subject-of$).

Additionally, the sketch differences generated for pair words *politics*/*economy* (Fig. 3) include words which have the same grammatical relations (and/or, $subject-of$, etc.) with related keywords. The relation and/or lists more results for the word *politics* compared to that listed for pair words *politics*/*society* like $class$ and *care*. For the word *economy* that relation lists words *cost*, *element*, *environment*, *creation*, etc. However, the word *society* appear in the list of results for both *politics* and *economy* under the same relation and can be regarded as a hidden semantically related connection between them.

Consequently, the words *politics* and *society* relate to word *economy*, and all they form a cluster which share semantic relation of similarity and that words can be regarded as synonyms. The semantic relation was evaluated on the base of common collocation search by generation of words sketch differences.

The words from the second cluster present semantic relation between keyword *politics* and words *enhancement*, $stakeholders$ and *election* also by having common collocations which are not presented directly. The generated word sketch differences results for pair words *politics*/*enhancement* are presented at Fig. 4. They include relations and/or, $subject-of$, $modifier$, $object-of$, etc.

The results for relation and/or list more words compared to results for pair words *politics*/*society* and *politics*/*economy* among which are *policy*, *power*, *society*, *history*, *people*, etc. The resulted words also relate semantically to the keyword *politics*. That relation does not present common collocations between displayed words and pair words.

The results for relation $subject-of$ include more words compared to results for the same relation of pair words *politics*/*society* and *politics*/*economy* among which are words *became*, *call*, *look* which relate semantically to keyword *politics*.

Fig. 4. The words sketch differences of pair words *politics/enhancement* and *politics/stakeholders* from BASE corpus.

Under the same relation, the words *learn* and *explain* are connected semantically to word *enhancement*. However, the word *get* is a common collocation to both *politics* and *enhancement* connecting them by hidden semantic relation $(subject - of)$.

The results for relation $modifier$ display rich list of words which relate semantically to keyword *politics* among which are $international$, $confrontational$, *electoral*, etc. Generally, that relation express a connection between a term (*politics*) and its hyponyms (usually multi-word terms – *international politics*, *electoral politics*, etc.). The results displayed for the keyword *enhancement* give the word *stimulus* presenting a combination *stimulus enhancement*. The word *local* appear as a semantically related word in the list of results for both *politics* and *enhancement* and is considered as a common collocation which connects semantically both words. Thus, the resulting multi-word terms can be *local politics* and *local stimulus enhancement*. The other relations for the pair words *politics*/*enhancement* do not contain common collocations.

Fig. 5. The words sketch differences of pair words *politics*/*election* from BASE corpus.

The generated words sketch differences for pair words *politics/stakeholders* are presented also at Fig. 4. and include the same relations and/or, $subject-of$, $modifier$, $object-of$, etc.

The results for relations and/or and $subject-of$ do not contain any common collocations relating to pair words. However, the results for the relation $modifier$ display exactly the same results as those generated for the pair words *politics/enhancement*. Thus, the words which semantically relate to keyword *politics* give exactly the same combinations forming multi-word terms. The result for the keyword *stakeholders* include only the word *local*. The same word also semantically relates to word *politics* and is a common collocation for both *politics* and *stakeholders*. The related multi-word terms are *local politics* and *local stakeholders*.

Thus, the word *local* connects not only the pair words *politics/stakeholders* but also the pair words *politics/enhancement* forming a triple of semantically related words (*politics/enhancement/stakeholders*). The results of the other generated relations do not contain any common collocations.

Part of word sketch differences results for pair words *politics/election* are presented at Fig. 5. and include the same relations and/or, $subject-of$, $modifier$, $object-of$, etc. The results for relations and/or, $subject-of$ and $modifier$ do not include common collocations. The relation $object-of$ contains almost the same words as for pair words *politics/enhancement* and *politics/stakeholders* which relate semantically to keyword *politics*. However, only the word *organize* is listed among semantically related words to both *politics* and *election* and is considered as a hidden connection. The other generated relations of pair words *politics/election* do not include any common collocations.

Consequently, common collocations search reveals specific types of hidden semantic relations and evaluates distributional thesauri by statistical ranking and comparing common collocations of two semantically related words. The sketch differences generation enlarges the number of extracted semantically related words revealing their complex semantic structure and specific hierarchical relations.

5 Conclusion

The presented semantic search technique includes the use of SE approaches for common collocations word search evaluating distributional thesauri in electronic text corpus. It is extended also with generation of words sketch differences which use grammatical relations to sort and filter the results. The received results show that using above type of extended search, it is possible to reveal specific word's semantic and grammar features (part-of-speech) by evaluating underlaying relations between different words contexts.

The technique can be used for multilingual application [12] since it uses statistical search and standard grammatical relations. It is applicable also for terminology extraction and can be used for compilation of electronic or printed dictionaries as well.

References

1. Azzopardi, J. et al., *Back to the Sketch-Board: Integrating Keyword Search, Semantics, and Information Retrieval*, In: Cali A., Gorgan D., Ugarte M. (eds) Semantic Keyword-Based Search on Structured Data Sources, KEYSTONE 2016, LNCS, vol. 10151, 2017, 49–61, Springer.
2. Cristani, M., Tomazzoli, C., *A multimodal approach to exploit similarity in documents*, In: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), vol. 8481, 2014, Springer, 490–499.
3. Cristani, M., Tomazzoli, C., *A multimodal approach to relevance and pertinence of documents*, In: Lecture Notes in Computer Science (Subseries of Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9799, 2016, Springer, 157–168.
4. Stoykova, V., *Using Statistical Search to Discover Semantic Relations of Political Lexica – Evidences from Bulgarian–Slovak EUROPARL 7 Corpus*, In: I. Kotsireas, S. Rump and Ch. Yap (eds.), Mathematical Aspects of Computer and Information Sciences, Lecture Notes in Computer Sciences, vol. 9582, 2016, Springer, 335–339.
5. Killgarriff, A. et al., *The Sketch Engine: Ten Years On*, In: Lexicography, 2014, 1, 17–36.
6. Sinclair, J., *Corpus, Concordance, Collocations*, 1991, Oxford, OUP.
7. Killgarriff, A., Markowitz, F., Smith, S., Thomas, J., *Corpora and Language Learning with the Sketch Engine and SkELL*, In: Revue francaise de linguistique, 2015, 1, vol. XX, 61–80.
8. Thompson, P., *Changing the Bases for Academic Word Lists*, In: P. Thompson and G. Diani (eds.) English for Academic Purposes: Approaches and Implications, 2015, Newcastle-upon-Tyne, Cambridge Scholars, 317–342.
9. Gledhill, Ch., *Collocations in Science Writing*, 2000, Tuebingen.
10. Stoykova, V., Mitkova, M., *Conceptual Semantic Relationships for Terms of Precalculus Study*, WSEAS Transaction on Advances in Engineering Education, 2011, issue 1, vol. 8, 13–22.
11. Stoykova, V., *Extracting Academic Subjects Semantic Relations Using Collocations*, EAI Endorsed Transactions on Energy Web and Information Technologies, vol. 4, 17(14), 2017, http://dx.doi.org/10.4108/eai.4-10-2017.153161
12. Stankovic, R. et al., *Keyword-Based Search on Bilingual Digital Libraries*, In: Cali A., Gorgan D., Ugarte M. (eds) Semantic Keyword-Based Search on Structured Data Sources, KEYSTONE 2016, LNCS, vol. 10151, 2017, 112–123, Springer.