EAI Endorsed Transactions Preprint Research Article/Editorial

Extracting Academic Subjects Semantic Relations Using Collocations★

Velislava Stoykova*

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

Abstract

The paper presents approach to analyze semantic content of academic subjects and its internal relations using statistically-based techniques for collocation extraction from large electronic educational text corpus. It offers a survey and analysis of some related corpus-based approaches to extract conceptual relations used for educational purpose and presents a technique for semantic search of collocations. The results of extended keyword search from British Academic Spoken English corpus using Sketch Engine searching software are presented. They are analysed with respect to types of generated keyword's collocations and semantic relations which they assign.

Received on 24 October 2016; accepted on 10 September 2017; published on XXXX

Keywords: Data mining, Big data, Knowledge discovery

doi:10.4108/eai.12-9-2017.153069

1. Introduction

Recent developments in the use of Internet technologies enlarge the scope of approaches to adopt various types of search, so to improve the effectivness and correctly related to the query seach results. The Internet users are tend to search using various types of keywords, so to get as much as closer to searched keywords outcome.

Hovever, for the effective search results a complex techniques are developed combining information retrieval and semantic approaches [1]. The techniques use related data structure which are capable to deal with complex semantic representations, so to filter and extract the required knowledge. Moreover, that technologies are with multilingual application, since they use various statistical approaches and knowledge extraction.

Further, we are going to present results of research aimed at analyzing semantic content of academic subjects based on the use of statistically-based keyword search techniques in academic educational electronic text corpora.

2. Related Approaches

Keywords generation is widely known technique to define semantic textual relations. It, also, is used to extract semantic relations between words, text coherence, etc. The tradition of application of electronic text corpora for education uses different academic texts [2] (including from Wikipedia [3]) and approaches extended and improved with statistically-based techniques to extract sophisticated semantic relations.

Various techniques were successfully applied for wide range of languages including Chinese [4, 5]. The results of existing applications significantly improved their universality with respect to domain application and multilingual scope [6]. The related techniques used are based on adoption of various metrics for extraction different types of word semantic relations by applications using estimation of word similarity measure.

Further, we are going to present and analyse applications of such techniques by comparing and discussing several results of extended keyword search with collocations in educational electronic text corpus

★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)"
*Email: vstoykova@yahoo.com

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

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

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

Figure 1. The formulas of Sketch Engine's statistical scoring.

using statistical functions of Sketch Engine specialized software.

3. The Sketch Engine (SE)

The SE software [7] 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.

The statistical approaches used by SE to search for co-occurrence and collocated words are based on defining probability of their co-occurrences and collocations. 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 presented at Fig. 1. 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 [8] which can be extracted by using techniques for estimation the strength of association between co-occurring words. Recent developments improved that techniques with respect to application areas including language learning [9].

Further, we shall present and analyse results for extracting collocations using SE software and compare related results with respect to semantic types of received collocations and related texts sources.

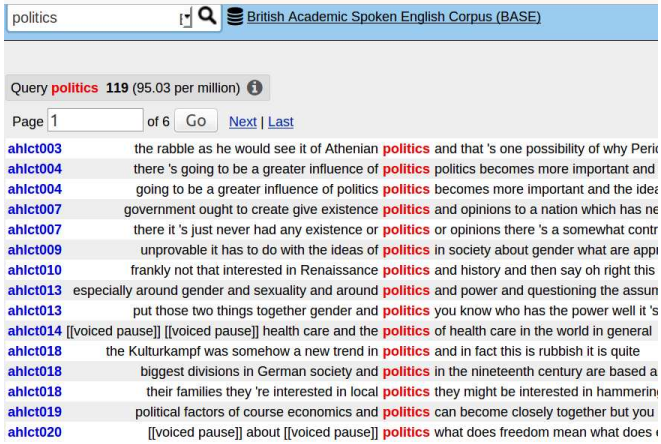


Figure 2. The concordance of keyword *politics* from BASE corpus.

4. The British Academic Spoken English (BASE) Corpus

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 analyse English for Academic Purposes [10].

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 the SE allowing the use of its incorporated options for storing, sampling, searching and filtering texts according to different criteria.

5. Keyword Search Results

For our research, we shall use SE standard options to generate keyword's concordances, distribution, collocations, grammatical and semantic relations. We shall present general methodology by demonstrating related results for keyword *politics*.

Concordances present all occurrences of given keyword with its related quantitative contexts. Fig. 2 presents all occurrences of keyword *politics* within BASE corpus with its related contexts. The generated results show that keyword *politics* has 119 occurrences but do not give information about its frequency distribution which is also important structural criterion.

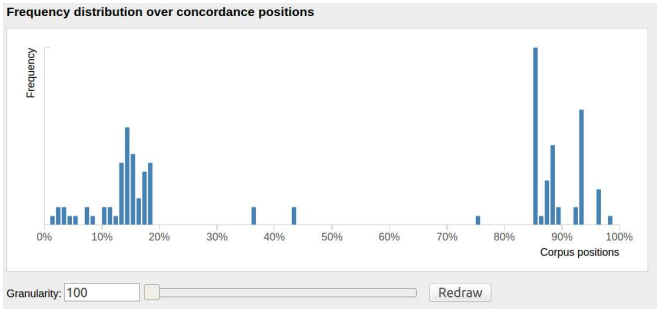


Figure 3. The frequency distribution of keyword *politics* over BASE corpus.

	doc.id	Frequency
P | N	sslct018	19
P | N	sslct031	14
P | N	ahlct023	11
P | N	sslct024	9
P | N	ahlct029	7
P | N	ahlct028	6
P | N	ahlct025	6
P | N	sslct022	5
P | N	sslct034	4
P | N	ahlct024	3
P | N	ahlct022	3

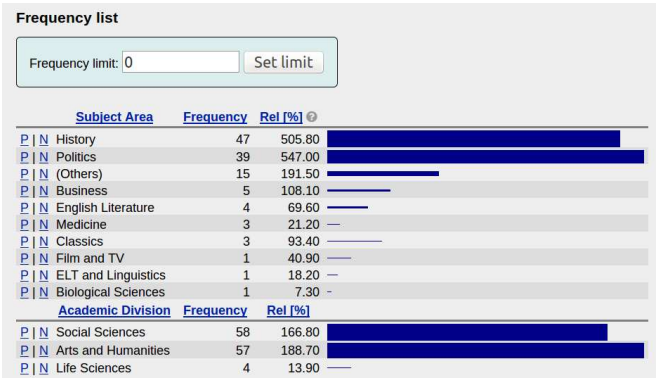
Figure 4. The frequency distribution of keyword *politics* over concordance position.

The SE has options to evaluate different types of keyword distribution. Thus, Fig. 3 shows frequency distribution of keyword *politics* over whole BASE corpus. The received results lead to conclusion that keyword is not coherent within whole corpus and is frequent only in certain texts. More detailed information about frequency distribution of keyword *politics* is obtained by generation of keyword distribution over its concordances.

The related results are presented at Fig. 4 and show that distribution of keyword is more coherent over certain concordance position and can be detected with respect to different thematic part of BASE corpus. Thus, the keyword occurs mostly in texts from domain of Social Studies and Sciences and Arts and Humanities texts of whole corpus.

Another SE option for detecting keyword frequency distribution is the generation of keyword's distribution over subject areas. Fig. 5 shows the distribution of keyword *politics* over subject area.

The generated results show that keyword *politics* is occurred in texts from History, Politics, Business, English Literature, etc. subject areas. The results presented at Fig. 3, Fig. 4 and Fig. 5 show that keyword *politics* can be occurred within structured texts of related specific domains, areas or subjects.



	Subject Area	Frequency	Rel [%]
P | N	History	47	505.80
P | N	Politics	39	547.00
P | N	(Others)	15	191.50
P | N	Business	5	108.10
P | N	English Literature	4	69.60
P | N	Medicine	3	21.20
P | N	Classics	3	93.40
P | N	Film and TV	1	40.90
P | N	ELT and Linguistics	1	18.20
P | N	Biological Sciences	1	7.30
	Academic Division	**Frequency**	**Rel [%]**
P | N	Social Sciences	58	166.80
P | N	Arts and Humanities	57	188.70
P | N	Life Sciences	4	13.90

Figure 5. The frequency distribution of keyword *politics* over subject area.

However, concordance and frequency distribution do not give semantic information about keyword's meaningful combinations. For that, a semantic filtering is needed by extending keyword search with collocations.

The SE offers several statistical approaches to generate collocations of a related keyword. However, for our analysis we shall use only that presented in Section 3. Thus, we apply $MI - score$ which was already used in [11] for parallel bilingual collocations generation.

Fig. 6 shows generated collocation candidates for keyword *politics* from BASE corpus. The results show that most frequent words which are most probably to be occurred together with keyword *politics* are: *electoral*, *international*, *gender*, etc.

The SE allows more elaborated keyword search over structured data to extract both grammatical and semantic relations. The related techniques are based on the idea that word association measures extract not only collocations but also other types of associations between a lexical unit and a grammatical word or between two semantically related words (hypo- or hyperonyms). Thus, we use SE's word sketch option to generate most frequent grammatical relations of keyword *politics*.

Fig. 7 shows generated results which include following keyword's relations together with their frequent collocations: as $modifier$, as $pp - obj - of$, as and/or, as $obj - of$, etc. Thus, the SE keyword search can extract not only statistically similar words for building thesauri but also can define their semantic relations.

Generally, the search is performed over structured data and gives results with respect to related structures. For example, concordance search gives all occurrences of keyword with related contexts within the whole corpus. Distributional frequency search gives distribution of keyword within the whole corpus, within the domain sub-corpora or within the subject areas. The collocation candidates search gives as a result list of words which

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

Figure 6. The collocation candidates of keyword *politics* from BASE corpus.

politics British Academic Spoken English Corpus (BASE) freq = 119 (95.02 per million)

unary rels			modifier			pp_obj_of			and/or			object_of		
	56	0.47		35	0.29		18	0.15		17	0.14
Sfin	18	0.15	international	18	10.94	touchstone	1	9.61	bureaucracy	1	10.00	decode	1	10.68
Swh	8	0.07	*of international politics*			domination	1	8.87	affair	1	9.00	erase	1	10.35
poss	8	0.07	electoral	3	10.52	edition	1	8.87	opinion	2	8.96	borrow	1	8.81
VPing	3	0.03	confrontational	1	9.12	agenda	2	8.44	paradigm	1	8.65	join	1	8.11
Swhether	1	0.01	reproductive	1	9.12	conduct	1	8.08	economics	1	8.31	organize	1	7.63
VPto	1	0.01	enlightened	1	9.12	aspect	3	8.05	gender	1	7.91	reject	1	7.62
			domestic	2	9.10	currency	1	8.04	existence	1	7.67	enter	1	7.04
			politicocentric	1	9.09	analysis	5	8.00	aspect	2	7.53	believe	1	6.63
			orthodox	1	8.82	reality	1	7.70	class	2	6.68	define	1	6.48
			elite	2	8.81	institution	1	7.45	care	1	6.10	influence	1	6.42
			athenian	1	8.69	importance	1	7.29	policy	1	5.62	understand	1	5.70
			organized	1	8.64	study	2	6.43	power	1	5.37	say	2	4.00
			renaissance	1	8.62	centre	1	6.41	society	1	5.32	see	1	3.41
			wing	1	8.55	influence	1	6.35	history	1	4.52	give	1	3.36
			broader	1	8.54	element	1	6.27	people	1	2.82	know	1	3.10
			comparative	1	8.47	historian	1	6.05				get	1	2.10

Figure 7. The grammatical relations of keyword *politics* from BASE corpus.

are most probably to be found with a related keyword. The results include both attributive collocations like *electoral politics* and specialized collocations like *international politics*.

The semantic relations search takes into account keyword's grammar features and gives as a result all possible semantic relations of keyword and its related collocations. Consequently, different types of collocations search generates keyword's semantic profiles which describes both semantic and grammar features.

6. Conclusion

The approach presents search and retrieval over educational electronic text corpus, which use SE statistically-based techniques for extending the keyword search to evaluate frequency distribution, collocation candidates, grammatical and semantic relations.

The analyzed keyword search results show that using different types of extended search, it is possible to capture keyword's constraints (lexical, grammatical, syntactic or semantic) which govern word combinations selection in related semantic context. In that way, it is possible to extract general keyword's semantic relations. The precision of statistical filtering is used to extract, range and isolate semantically-related keyword combinations and to receive more semantically relevant to keyword results.

References

[1] J. Azzopardi 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] Gledhill, Ch., *Collocations in Science Writing*, Tuebingen (2000).

[3] Hunter, B., Austin, R., *Wiki'd Transformations: Technology Supporting Collaborative Learning*, Higher Education in Transformation Conference, pp. 522–534, Dublin, Ireland (2015)

[4] Li, F., *Construction on English Translation Corpus System for Chinese Cultural Classics*, International Conference on Education, Sports, Arts and Management Engineering (ICESAME 2016), pp. 1155–1160, Atlantis Press (2016)

[5] He, Z., *Design on Corpus System for Classical Chinese Teaching in Middle School*, 5th International Conference on Social Science, Education and Humanities Research (SSEHR 2016), pp. 692–697 Atlantis Press (2016)

[6] Lossio-Ventura, J. A., Jonquet, C., Roche, M., Teisseire, M., *Biomedical Term Extraction: Overview and a New Methodology*, Information Retrieval Journal, vol. 19, issue 1, 59–99, Springer (2016)

[7] Killgarriff, A. et al., *The Sketch Engine: Ten Years On*, Lexicography, 1, 17–36 (2014)

[8] Sinclair, J., *Corpus, Concordance, Collocations*, OUP, Oxford (1991)

[9] Baisa, V., Suchomel, V., *SkELL: Web Interface for English Language Learning.*, Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, pp. 63–70, Tribun EU, Brno (2014)

[10] Thompson, P., *Changing the Bases for Academic Word Lists*. In: P. Thompson and G. Diani (eds.) English for Academic Purposes: Approaches and Implications, pp. 317–342, Cambridge Scholars, Newcastle-upon-Tyne (2015)

[11] V. Stoykova, *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, 335–339, Springer.