**EURASIAN JOURNAL OF MATHEMATICAL
AND COMPUTER APPLICATIONS**
ISSN 2306–6172
Volume 5, Issue 3 (2017) 73 – 86

A CONTRASTIVE APPROACH TO TERM EXTRACTION: CASE-STUDY FOR THE INFORMATION RETRIEVAL DOMAIN USING BAWE CORPUS AS AN ALTERNATIVE COLLECTION

Nugumanova A.B, Apayev K.S. , Baiburin Ye.M., Mansurova M.Ye.

Abstract This work describes a case-study on contrastive term extraction from a popular tutorial on information retrieval by C.D. Manning, P. Raghavan and H. Schutze. The results on extraction of terms are evaluated with the help of a reference index of terms compiled by the authors themselves. Execution of the case-study was motivated by 2 factors. Firstly, this work is part of a major project on automatic creation of an ontology in the field of information retrieval. Secondly, the authors also study applicability of the British Academic Written English (BAWE) as an alternative collection necessary for a contrastive extraction of terms. BAWE is a well-balanced and sufficiently representative corpus containing high-quality academic texts on 35 academic disciplines. The aim of this work is to show that the use of a contrastive approach with the help of such a balanced and representative alternative collection as BAWE allows to quite effectively solve the problem of automatic term recognition.

Key words: term extraction, contrastive approach, information retrieval, automatic term recognition, BAWE corpus.

AMS Mathematics Subject Classification: 68P20

1 Introduction

The British Academic Written English (BAWE) was created as a joint project of three British universities: University of Warwick, University of Reading and Oxford Brooks University. The aim of the project was to collect the best examples of written works of undergraduate and graduate students of the mentioned universities [1]. Thus, the corpus included about 3000 works on 35 academic disciplines in four fields of science: art and humanitarian sciences, life sciences, physical sciences and social sciences. At present, the corpus is available for downloading from the Oxford Text Archive as Resource number 2539 [2]. As this, it contains 2761 documents each of which is provided with a detailed abstract including data such as the code of work, its title, course, date of writing, genre of work, academic discipline, the estimate obtained, and the number of words. Also, it contains information on the author of each work, in particular, such abstract contains data on a student's sex, his/her date of birth, the first language, country where he/she is from, etc.

Initially, the corpus was created for studying language peculiarities inherent to the works of British higher educational institutions [3]. In particular, the style, vocabulary, genre diversity of academic written works, dependency of style and genre on the field of science and discipline were studied according to the examples collected in the corpus.

Later, the corpus has been widely used not only by linguists but by all those who are interested in studying the written English language.

In the field of natural language processing, the BAWE corpus has started to be used as a test collection practically from the moment of its publication in open access. So, the pilot version of the corpus consisting of only 500 documents was used in [4] for performing experiments on automatic identification of the authors' gender. According to the experimental results, the gender of 81% of authors was identified correctly. In [5], the corpus was used for carrying out experiments on topic modeling. To verify their method, the authors used the texts of the BAWE corpus referring to the field of Arts and Humanities. In [6], the authors used the texts of the corpus for automatic determination of theme in English sentences. The system Theme Analyzer worked out by these authors automatically determined not only the theme-rhematic structure of each sentence, but also the included syntactic nodes, thematic roles, etc.

One of interesting BAWE corpus application practices is its use as an alternative collection of documents necessary for comparison with another collection which is interesting for the researcher. In [7], the authors use BAWE together with a collection of texts containing descriptions of ritual actions to extract keywords associated with this domain. The authors use the well-proven contrastive approach identifying the keywords of the domain from the point of view of their different occurrence within the domain and beyond it. The words which are often used within the domain and extremely rarely beyond it are considered to be keywords. In this cited case, "within the domain" means in the texts describing rituals and "beyond it" means in the texts of alternative collection, i.e. in the texts of BAWE corpus. In [8], the authors use BAWE for comparison with another corpus, too, which is, according to their words, "a direct, practical and fascinating way of studying the characteristics of corpora and types of texts". The authors of this work analyze the top 100 keywords of each of the corpora under study and compare these lists with each other.

The aim of this work is to show that the use of a contrastive approach with the help of such a balanced and representative alternative collection, with us considering BAWE as such, allows to effectively solve the problem of automatic recognition of terms contained in the domain collection of texts. In this work, the textbook "Introduction to Information Retrieval" [9] is used as a domain collection of texts. The textbook is available in electronic form on the website of Stanford University [10] and is provided with an author's index of terms which is used in experiments as a gold standard for evaluating the precision and recall of extraction of terms.

According to the set up aim, the work is further presented as follows. Section 2 presents the meaningful analysis of the BAWE corpus, gives a brief description of each of the four sections of the corpus. Section 3 describes the essence of the contrastive approach to extraction of terms and a detailed process of preprocessing of texts necessary for using the considered approach. Section 4 describes the performed experiments and analyzes their results. Section 5 contains the conclusion and plan of further works.

Table 1: Distribution of BAWE corpus texts in the fields of science

No.	Field of science	Number of texts
1	Arts and Humanities (AH)	705
2	Life Sciences (LS)	683
3	Physical Sciences (PS)	596
4	Social Sciences (SS)	777
	Total	2761

2 The meaningful analysis of BAWE corpus

A contrastive approach is a common name of methods that identify terms based on the contrastive in their distribution within a domain (in the texts of a domain collection) and beyond it (in the texts of alternative collection, i.e. texts collected from different domains not related to the domain under consideration). The main criteria for the quality of an alternative collection are its representativeness and balance. Representativeness means that an alternative collection should cover as many texts as possible from as many domains as possible not related to the target domain. Balance means that different domains in the alternative collection should be presented in equal proportions.

From the point of view of the mentioned criteria, the corpus BAWE is quite representative (124516 words) and balanced (4 fields of science are presented by approximately equal number of texts). Table 1 shows distribution of BAWE texts in the fields of science and Figure 1 presents a diagram illustrating the balance of BAWE.

In [1], the authors describe the corpus composition in detail and give statistics on distribution of texts in all possible sections: according to subject disciplines, genres, years, courses, etc. Figure 2 presents histograms of distribution of corpus texts in each of the four fields of science with details on academic subjects. It is seen that the "Engineering" discipline accounts for the greatest number of texts in BAWE (238), then comes "Biology" (169) and in the third place is "Business" (146).

We analyzed distribution of words in the three most representative disciplines and built their clouds (see Figure 3-6). As the number of all words is very great for visualization, we used only words with the frequency of use not less than 70. Before building of clouds, the texts were subjected to preprocessing: at first tokenization was executed (breaking texts into words and other tokens), then lemmatization (reduction of words to normal forms), then we deleted number, punctuation marks and stop words. Table 2 presents pairwise intersections of top 100 key words for each of the three academic disciplines under consideration.

Simultaneously, among the extracted top 100 words, we underlined the top words common for all three disciplines (see Table 3). These are scientific terms such as "result", "system", "factor", "process", "table" and so on. If we extend this approach to all disciplines of BAWE corpus, we can speak about the prospects of automatic building of a dictionary of general scientific and inter-branch vocabulary based on BAWE corpus. We know such works on automatic or semiautomatic building of dictionaries of

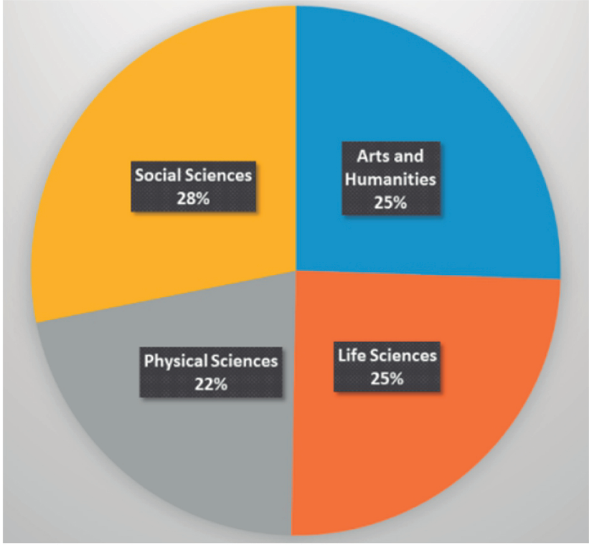

Figure 1: Diagram of distribution of BAWE corpus texts in the fields of science

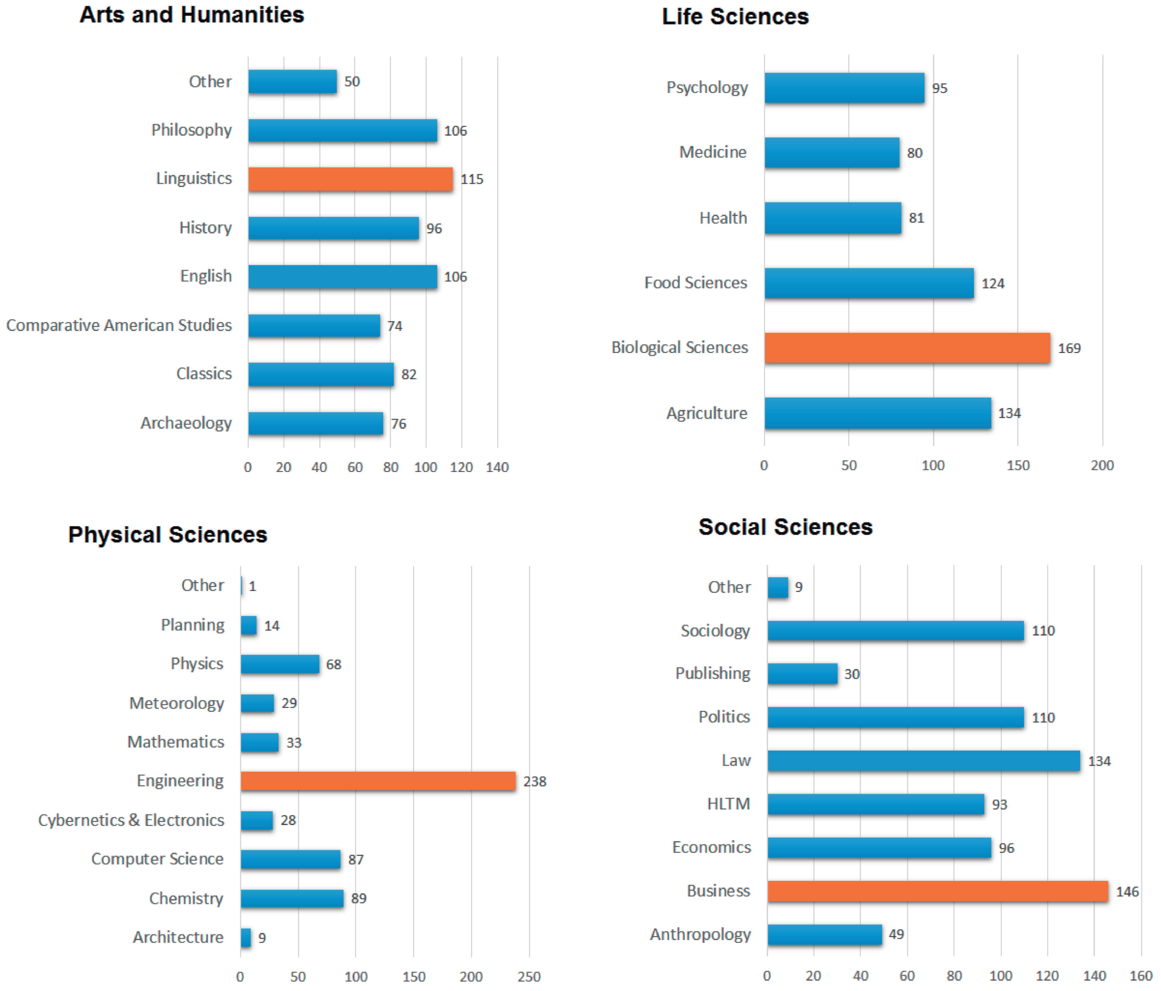

Figure 2: Distribution of BAWE corpus texts on academic subjects

Table 2: Intersections of top key words for "Engineering", "Biology" and "Business" academic disciplines

Engineering and Biology - 26 words	Engineering and Business - 47 words	Biology and Business - 35 words
activity, change, control, development, factor, figure, form, group, high, important, increase, level, need, number, order, process, product, production, quality, rate, result, role, study, system, table, time, year	analysis, based, business, case, change, company, control, cost, current, customer, development, factor, figure, financial, good, group, high, important, increase, information, level, management, market, model, need, number, order, performance, point, power, price, problem, process, product, profit, project, rate, result, service, strategy, system, table, team, term, time, work, year	area, change, control, data, development, effect, energy, experiment, factor, figure, formula, group, high, higher, important, increase, level, method, need, number, order, picture, process, product, rate, required, result, small, stage, system, table, temperature, time, type, year

general scientific vocabulary based on corpora of scientific texts, for example, we can refer to work [11] for the English language and [12] - for Russian.

3 A contrastive approach to extraction of terms

There are a great number of methods using a contrastive approach to extraction of terms. For example, a detailed review of similar methods is given in [12]. In this work, we use two methods of contrastive extraction of key words: 1) one of the simplest methods based on the use of Pearson criterion (chi-squared test) and 2) one of the most effective at present methods based on calculation of measure TF-DCF proposed

Table 3: Common academic words, extracted from intersections of top-words for "Engineering", "Biology" and "Business" academic disciplines

No.	Word	No.	Word	No.	Word
1	change	8	important	15	product
2	control	9	increase	6	rate
3	development	10	level	17	result
4	factor	11	need	18	system
5	figure	12	number	19	table
6	group	13	order	20	time
7	high	14	process	21	year

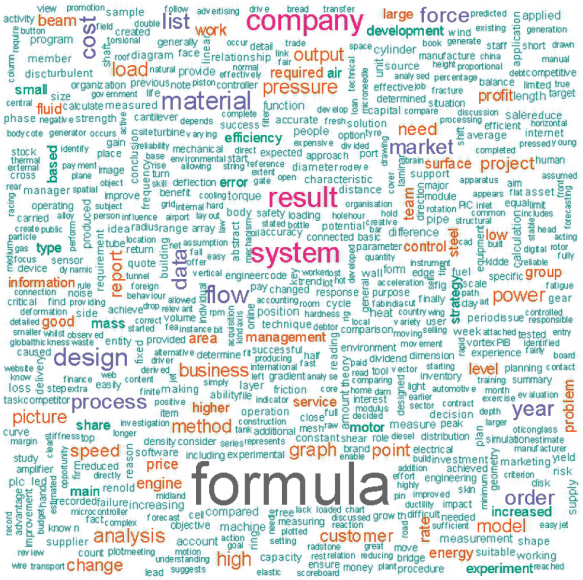

Figure 3: Cloud of words based on the texts of the "Engineering" discipline

in [13].

Chi-squared test refers to the class of statistical tests evaluating significance of discrepancy between the experimental data and theoretical model. If the criterion value is below the critical value, one accepts a hypothesis on agreement of the data with the model (Null hypothesis), if it is higher, the hypothesis is rejected. As applied to extraction of key words from the texts of a subject domain, chi-squared test determines the significance of discrepancy between distribution of words in the texts of a domain and the texts of an alternative collection.

Let the word t occurs in A texts, not occur in B texts in a subject collection, and in an alternative collection the same word occurs in C texts and does not occur in D texts. Then, the formula of chi-squared test is as follows (its detailed conclusion is given, for example, in [14]):

$$Chi^2(t) = \frac{(A+B+C+D)*(A*D-C*B)^2}{(A+C)*(B+D)*(A+B)*(C+D)} \qquad (1)$$

The critical value of the test is calculated on the basis of a special table (for example, at the error probability level in 1% it is equal to 6.6).

Thus, to define a list of key words and word combinations of a domain, first it is necessary to select all the words and most frequently used words (bigrams, trigrams and so on) from the texts of a domain. Then, for each selected word or word combination it is necessary to calculate the values of A, B, C and D and substitute in formula (1). Then, it is necessary to filter out all words for which the formula will return the criterion value less than the critical one. The criterion is symmetrical, i.e. it selects key words for both a domain and an alternative collection, therefore it is also necessary to discard all the words for which $A < B$. The obtained in this way list can be considered as one of the variants of the list of terms of a subject domain.

Measure TF-DCF develops the idea of fines and rewards in the basic construction

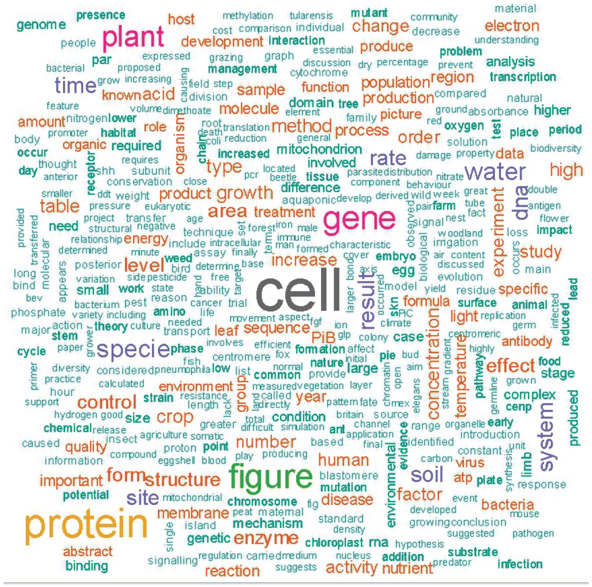

Figure 4: Cloud of words based on the texts of the "Biology" discipline

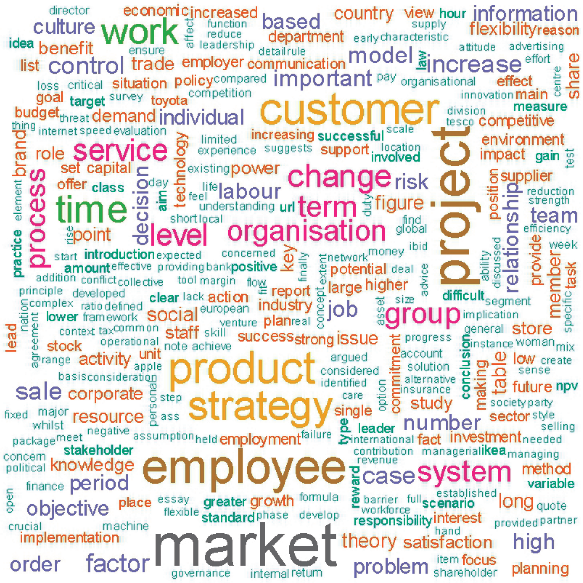

Figure 5: Cloud of words based on the texts of the "Business" discipline

Table 4: The examples of reference terms from the book "Introduction to Information Retrieval", contained in the author's index (taken in a random order)

No.	One-word terms	Two-word terms	Three-word terms
1	accumulator	authority score	ad hoc retrieval
2	break-even	auxiliary index	binary independence model
3	BSBI	average-link clustering	blind relevance feedback
4	lemmatization	Bayes risk	click through log analysis
5	likelihood	cumulative gain	maximum likelihood estimation
6	LSA	data-centric xml	multivariate Bernoulli model
7	NLP	support vector	natural language processing
8	regression	term frequency	principal left eigenvector
9	regularization	term-document matrix	unigram language model
10	Reuters-21578	word segmentation	vector space model

of TF-IDF formula and proposes a new variant of this formula called term frequency-disjoint corpora frequency. The absolute frequency of the word usage in a subject domain is used as a reward and the product of absolute frequencies of the word usage in a set of other subject domains is used as a fine:

$$TF - DCF(t) = \frac{f_t^D}{\prod_{g \in G} 1 + \log\left(1 + f_t^g\right)} \tag{2}$$

where f_t^D and f_t^g are frequencies of use of the word t in a subject and an alternative collections, respectively, G is asset of all such alternative collections. The authors prove experimentally that their method is the best one among a number of contrastive methods. They justify the use of the product in the denominator of formula by the fact that the fine must grow in a geometric progression for each use of the word in a new subject domain. Thus, this measure is well suited for the case when there are several alternative subject domains (as in our case when collection of BAWE is formed from the texts of 35 academic disciplines). For measure TF-DCF, a critical value is not given, it must be determined empirically. In other respects, the principle of selection of terms is the same as for chi-squared test.

In this work, we will extract one-, two- and three-word terms and then compare the list of extracted terms with the reference author's index. The reference author's index contains 603 terms including 174 one-word terms, 335 two-word terms, 78 three-word terms, 14 four-word terms and 1 six-word term. The examples are given in Table 4.

Due to the presence of the reference list we can evaluate the *Precision* and *Recall* of the considered contrastive methods for extraction of terms. For this, it is necessary to calculate the values as shown in Table 5.

The precision and recall can be evaluated according to the following formulae:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Table 5: Support values for calculating precision and recall

Notation	Title	How is it determined
TP	True Positive	Number of extracted terms that are included in the reference list
FP	False Positive	Number of extracted terms that are not included in the reference list
FN	False Negative	Number of terms of the reference list that are not included in the number of extracted terms

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

The obtained values of precision and recall of extraction of terms can be combined in a single index called F-measure with the help of average harmonic:

$$F1 = 2\frac{Precision * Recall}{Precision + Recall} \tag{5}$$

4 Experimental work

The experiments on extraction of terms were performed in R using libraries tm and quanteda [15, 16]. The both collections (the chapters of the book and texts of BAWE corpus) were loaded into R and then transformed for convenient processing (they were presented in the form of sparse documents-to-terms matrices).

The collection based on the chapters of the book "Introduction to Information Retrieval" was parsed from the website of Stanford University where it was in the public domain in the form of html-pages [10]. When parsing, the html-tags were deleted and the book content was exported to 245 text files according to the number of chapters and paragraphs of the book. The collection based on the texts of BAWE corpus was downloaded from the website of Oxford archive of texts where it was in open access, too, in the form of an archive of text files [2].

The text of each file was lemmatized with the help of Wordnet Lemmatizer which was part of the package NLTK - an open library of programs for symbolic and statistical processing of a natural language [17]. When extracting terms, post-tagging was not used, correspondingly, the search by lexical patterns which are characteristic of two- and three-word terms was not used. Undoubtedly, this considerably decreased the precision of extraction of terms as, for example, two-word terms are mainly characterized by lexical templates of the form A+N (adjective + noun), while we selected all two-word combinations (bigrams). The same concerns three-word combinations (trigrams).

Table 6 presents top-30 one- and two-word terms extracted with the help of measure TF-DCF, when using all four sections of BAWE as alternative collections. Not all of the extracted entities are terms according to the author's version. For example, in the

Table 6: Top-30 one- and two-word terms extracted with the help of TF-DCF, when using all 4 sections of BAWE as alternative collections

Rank	Term	Rank	Term	Rank	Term
1	postings list	11	relevant document	21	machine learning
2	query term	12	term document	22	term frequency
3	information retrieval	13	language model	23	IDF
4	text classification	14	crawler	24	number document
5	web search	15	nonrelevant	25	Rocchio
6	document collection	16	multinomial	26	document query
7	relevance feedback	17	single-link	27	complete-link
8	training set	18	Reuters-RCV1	28	set document
9	KNN	19	SVM	29	IR system
10	inverted index	20	linear classifier	30	centroid

Figure 6 – Cloud of concepts of the "Information retrieval" domain

author's reference list there are bigrams such as "machine learning" (though, according to our opinion, this is an explicit term) or "set document" (this is undoubtedly not a term). Evidently, these examples indicate high complexity of the problem of extraction of terms.

In Table 7, we present top-30 one- and two-word terms extracted with the help of chi-squared test. Though, at first sight, the two lists of terms in Tables 6 and 7 do not differ much, in fact, at close comparison, it is seen that chi-squared test was less effective. For example, the top-30 included such words as "algorithm", "compute", "vector", "Boolean" which are not terms referring to only the field of information retrieval, these terms are also widely represented in such fields as mathematics, computer sciences, engineering.

Thus, as expected, when using the method TF-DCF, the indices of precision, recall and F-measure proved to be higher than when using chi-squared test (see Tables 8-9). Therefore, in the course of formation of ontology in the field of information retrieval, the terms selected with the help of measure TF-DCF were used as basic constructions for concept building. Figure 6 shows a cloud of ontological concepts extracted according to measure TF-DCF. It should be noted that the description of how domain concepts were formed from the extracted terms is beyond the scope of this work. The result of comparing quality indicators of extraction of terms with the help of measure TF-DCF depending. On the number of alternative collections (Table 10) appeared to be noteworthy. Maximum quality indicators were obtained when using maximum number of alternative domains - 4. An important condition was not to let alternative domains cross the target domain. Thus, when part of texts referring to computer sciences (i.e. a domain related to the domain of information search) was excluded from the alternative collection, both precision and recall of extraction of terms increased.

Another noteworthy fact is that the recall of extraction of terms decreases with the increase in the number of alternative texts. For example, when using 3 collections,

Table 7: Top-30 one- and two-word terms extracted with the help of chi-squared criterion, when using all 4 sections of BAWE as alternative collections

Rank	Term	Rank	Term	Rank	Term
1	query	11	inverted index	21	IR system
2	retrieval	12	postings list	22	term occur
3	query term	13	vector	23	centroid
4	information retrieval	14	Boolean	24	naive
5	algorithm	15	document collection	25	IDF
6	posting	16	classifier	26	relevance feedback
7	compute	17	text classification	27	relevant document
8	vector space	18	retrieval system	28	naive Bayes
9	search engine	19	term frequency	29	machine learning
10	web search	20	IR	30	nonrelevant

Table 8: Maximum quality indices of term extraction using 4 alternative collections $(AH + LS + SS + PS)$ after deduction of discipline "Computer Science"

Index	One-word terms		Two-word terms		Three-word terms	
	Chi-squared	TF-DCF	Chi-squared	TF-DCF	Chi-squared	TF-DCF
Precision	0.1818	0.24	0.1654	0.2018	0.1443	0.1271
Recall	0.3086	0.24	0.1284	0.2627	0.1489	0.2447
F-measure	0.2288	0.24	0.1446	0.2283	0.1466	0.1673

Table 9: Quality indices of one-, two- and three-word term extraction using 4 alternative collections $(AH + LS + SS + PS)$ after deduction of discipline "Computer Science"

Index	Chi-squared (with threshold 24)	TF-DCF (with threshold 5.5)
Precision	0.1045	0.2196
Recall	0.2913	0.2470
F-measure	0.1539	0.2325

Table 10: Dependence of precision, recall and F-measure on the number of alternative collections when term extracting with the help of measure TF-DCF (threshold 5.5)

Index	2 alternative collections $(LS + SS)$	3 alternative collections $(AH + LS + SS)$	4 alternative collections $(AH + LS + SS + PS)$	4 alternative collections $(AH + LS + SS + PS)$ without Computer sciences
Precision	0.2110	0.2156	0.2218	0.2196
Recall	0.2675	0.2623	0.2419	0.2470
F-measure	0.2359	0.2367	0.2314	0.2325

Figure 7 – Dependence of F-measure on threshold value of measure TF-DCF

the number of terms which were absent in alternative collections included such words as "index", "stemming", "entropy" while, when using 4 collections, these terms were considered to be ordinary words equally spread in both target and alternative collections. Correspondingly, recall of coverage of terminology decreased due to exclusion of these words from the list of terms. In a whole, the quality of extraction of three- and two- word terms is lower than that of one-word terms because, as we mentioned above, lexical templates were not used.

Of great interest is the problem of choice of a threshold value for both chi-squared test and measure TF-DCF. Figure 7 shows how the values of F-measure change with the change in the threshold value of measure TF-DCF, when extracting one-, two- and three-word terms. According to these results, the optimum threshold value giving maximum of F-measure is in the range from 3 to 6 (4.4 - for one-word terms; 5.6 - for two-word terms; 3.2 - for three-word terms).

5 Conclusion

In this work, we used a contrastive approach to automatic recognition of one-, two- and three-word terms used in the book "Introduction to Information retrieval". The British corpus of academic written English was used as an alternative collection. The carried out experiments allowed to make the following 3 conclusions:

1. When increasing the number of alternative collections, the precision and recall of extraction of terms increase, and it is important criteria which take into account not a cumulative distribution of terms in alternative collections but their individual frequencies in each separate collection.

2. Alternative collections should not be related to the target collection under consideration.

3. The British corpus of academic written English completely satisfies the above conditions and, despite its comparatively small size, can successfully be used as

a set of alternative collections.

It should also be emphasized that, when carrying out the experiments, we did not use any lexical-syntactical templates but completely relied on statistical approaches, this explaining not very high indicators of precision and recall of extraction of terms. If the approaches under consideration are complemented with lexical-syntactical analysis, the quality of extraction of terms will significantly increase.

Acknowledgement

This work is partially supported by the grant of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan No. 0115PK00779 "Development of intellectual high-performance information-analytical search system of processing semi-structured data" (2015-2017).

References

[1] Heuboeck A., Holmes J., Nesi H., *The BAWE corpus manual.* Technical report, Universities of Warwick, Coventry and Reading, 2007.

[2] British Academic Written English Corpus. Retrieved from http://ota.ahds.ac.uk/headers/2539.xml

[3] Ebeling S.O., Heuboeck A., *Encoding document information in a corpus of student writing: the British Academic Written English corpus.* Corpora, Vol. 2 Issue 2 (2007). P. 241-256.

[4] Doyle J., Keselj V., *Automatic categorization of author gender via n-gram analysis.* The 6th Symposium on Natural Language Processing, SNLP, 2005. P. 1-5.

[5] Allahyari M., Kochut K., *Automatic topic labeling using ontology-based topic models.* Machine Learning and Applications (ICMLA), IEEE 14th International Conference. IEEE, 2015. P. 259-264.

[6] Park K., Lu X., *Automatic analysis of thematic structure in written English.* International Journal of Corpus Linguistics, Vol. 202 Issue 1 (2015). P. 81-101.

[7] Reiter N. et al., *Adapting standard NLP tools and resources to the processing of ritual descriptions.* ECAI, 2010. P. 39.

[8] Kilgarriff A., *Getting to know your corpus.* International Conference on Text, Speech and Dialogue. Springer Berlin Heidelberg, 2012. P. 3-15.

[9] Manning C.D. et al., *Introduction to information retrieval.* Cambridge: Cambridge university press, 2008. 496 pages.

[10] Introduction to Information Retrieval. Retrieved from https://nlp.stanford.edu/IR-book/

[11] Da Sylva L., *Corpus-based derivation of a "basic scientific vocabulary" for indexing purposes.* Journal of Linguistics. Vol. 45 Issue 1 (2009). P. 167-201.

[12] Nugumanova A. et al., *A New Operationalization of Contrastive Term Extraction Approach Based on Recognition of Both Representative and Specific Terms.* International Conference on Knowledge Engineering and the Semantic Web. Springer International Publishing, 2016. P. 103-118.

[13] Lopes L., Fernandes P., Vieira R., *Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf*. Knowledge-Based Systems. Vol. 97 (2016). P. 237-249.

[14] Nugumanova A. et al., *Automatic keywords extraction from the domain texts: Implementation of the algorithm based on the MapReduce model*. Current Trends in Information Technology (CTIT), 2013 International Conference on. IEEE, 2013. P. 186-189.

[15] Feinerer I., Introduction to the tm Package Text Mining in R. Retrieved from https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

[16] Benoit K., Nulty P., *quanteda: Quantitative Analysis of Textual Data*. R package version 0.9. 2016.

[17] Perkins J., *Python 3 Text Processing with NLTK 3 Cookbook*. Packt Publishing Ltd., 2014.

A.B. Nugumanova,
D. Serikbayev East Kazakhstan State Technical University,
Ust-Kamenogorsk, Kazakhstan,
Email: yalisha@yandex.kz,
K.S. Apayev,
D. Serikbayev East Kazakhstan State Technical University,
Ust-Kamenogorsk, Kazakhstan,
Email: kapaev@ektu.kz,
Ye.M. Baiburin,
D. Serikbayev East Kazakhstan State Technical University,
Ust-Kamenogorsk, Kazakhstan, Email: ebaiburin@gmail.com,
M.Ye. Mansurova,
al-Farabi Kazakh national University,
Almaty, Kazakhstan,
Email: mansurova.madina@gmail.com