The Method of Synonyms Extraction from Unannotated Corpus

Alexander Alexandrovich Pak, Sergazy Sakenovich Narynov, Arman Serikuly Zharmagambetov
LLC AlemResearch
Almaty, Kazakhstan
aa.pak83@gmail.com

Sholpan Nazarovna Sagyndykova, Zhanat Elubaevna Kenzhebayeva, Irbulat Turemuratovich Utepbergenov
Institute of ICT
Almaty, Kazakhstan
i.utepbergenov@gmail.com

Abstract — The structuring of large volumes of e-documents assumes the organization of text on several levels, namely paragraphs, sentences, phrases, words. Methods of lexical paradigms extraction using statistical analysis were developed long ago. In this paper we attempt to move from lexical correlatives to the list of synonyms on various levels of generalization on the basis of local and global contexts' statistics.

Keywords — Extracting synonym algorithm, construction of a semantic map concepts, categorize the topics of texts, e-documents, Data Mining

I. INTRODUCTION

The constructing from the collections of the documents of hierarchical classes are used in information retrieval systems for more demonstrable view of texts. Some tasks of information retrieval such as topic identification and message sequencing (i.e. the messages related to the particular topic) require intelligent approaches for the processing of textual data. It is obvious that similar messages can have various thesauruses, styles, and the structure of narration. Thus, points corresponding to the similar messages can be located far away from each other. In the case of topic's development, every new related message can have its own thesaurus thereby increasing the distance between topic's related messages. The obvious solution in this case is to use synonyms dictionary, i.e. to transform the words' space into the meanings' space thereby decreasing the negative effect of points' divergence.

The possibility of structuring with coarse division was shown in the article WEBSOM [1]. The presented result is the clustering of scientific documents by categories (fuzzy logics, time series and others). The approach is based on the clustering of the averaged vector of local context, i.e. left and right words next to a headword. Fig. 1 shows semantic map obtained from processing with the help of WEBSOM algorithm. The authors were not able to produce a clean diagram of classes. In addition, the method of word encoding assumes the 90-dimensional random vector, which, in our opinion, caused mistakes in classes and applied extra restrictions in case of the big data scale. Also it is worth noting that another connectionist approach is based on deep learning [2]. The main idea is to use the algorithm of recursive self-organizing map (RSOM) with alternating compaction and mapping of neural network response to the stream of input symbols. Fig. 2 shows the resulting semantic map.

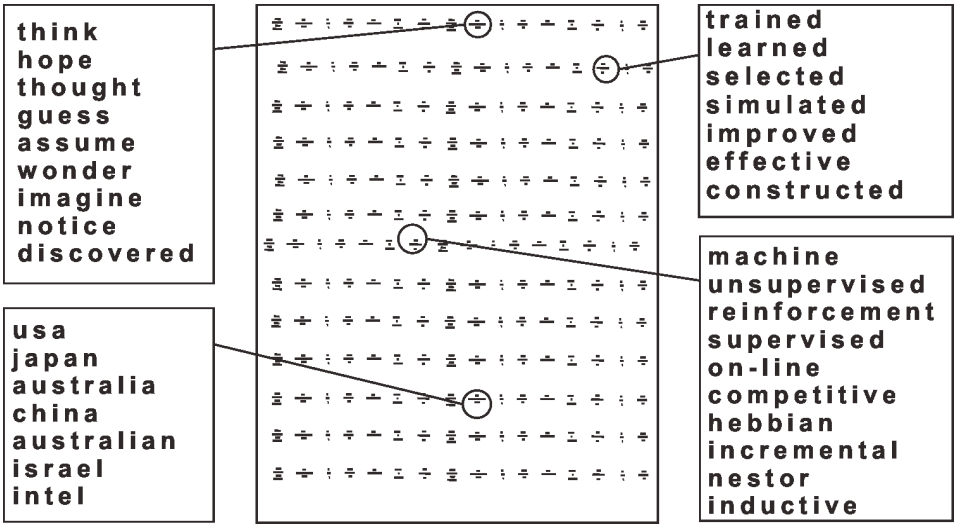

Fig. 1. Semantic map of words on the basis of WEBSOM algorithm. On the figure there are the cleanest classes. The figure is taken [1].

ISBN: 978-1-4799-6376-8/15/$31.00 ©2015 IEEE

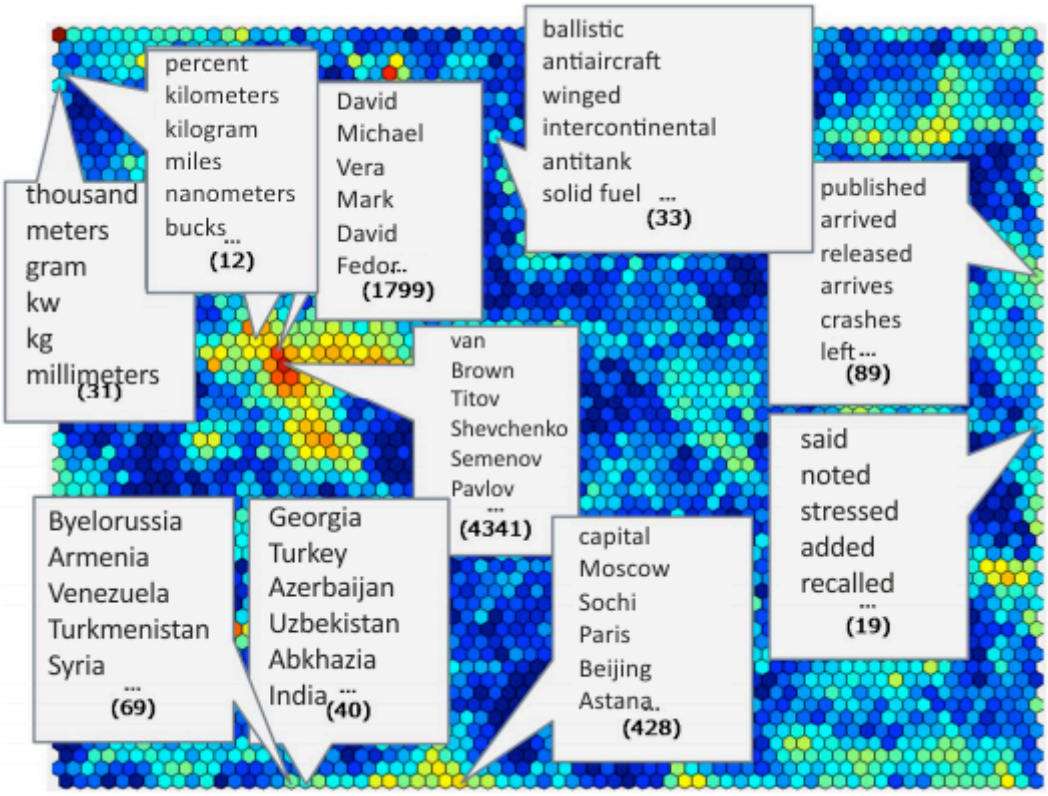

Fig. 2. Semantic map of Golem. Semantically close concepts are clustered together on the map. The figure is taken from [2].

The common feature of semantic maps mentioned above is the absence of level of parameter generalization. Therefore, the lexical correlatives can be erroneously grouped by local context ito the classes of words, such as:

1. Hyponym – A word of more specific meaning than a general or superordinate term applicable to it.

2. Hypernym – A word with a broad meaning that more specific words fall under; a superordinate.

3. Homonym – Each of two or more words having the same spelling but different meanings and origins.

4. Antonym – A word opposite in meaning to another.

In psycho-linguistic research [3][4] it was shown that the global context can be useful for resolving polysemy and homonymy.

II. THE RANGE STATISTICS OF LOCAL AND GLOBAL CONTEXTS

We propose to further develop the approach of local links (Model of local links, MLL) mentioned in [1][2], namely to enhance the model with the help of global context in the manner of [5]. Assume W be the sequence of words in some sentence $W \quad \{w_1, w_2, \ldots, w_k\}$, let's also denote the headword through w, and words from the left side w, right side w.
I is the sequence of the sentences W, $I \quad \{W_1, W_2, \ldots, W_k\}$
We denote G as the global context, i.e. it's the set of words $G \quad \{w_1, w_2, \ldots, w_k\}$ extracted from I, namely left-sided $W_{1-m}, W_{2-m+1}, \ldots, W_{1-1}$ and right-sided $W_{1+1}, W_{1+2}, \ldots, W_{1+m}$ sentences.

We denote B as the base string of words with m-length and ordered by alphabetic.

If we denote C as a local context that maps the set of left-sided words $\{w\}$ and right-sided words $\{w\}$ onto the base string B, then the proximity of two headwords w_1, w_2 in the space of local context is expressed by

$$cor(w_1, w_2) \quad 1 \quad \sum_{i=1}^{n} \frac{(a_i \quad b_i) < \Delta_{max} ? (a_i \quad b_i) : \Delta_{max}}{n\,\Delta_{max}} \quad (1)$$

Further, based on (1) the matrix of similarity of local and global averaged contexts can be built, so that the matrix for Hierarchical Agglomerative Clustering (HAC) [6] can be used on the next step. The advantage of HAC is the availability of an extra parameter of generalization's level. The result can be presented in the form of a dendrogram, as shown below.

III. NUMERICAL EXPERIMENTS

We agree with [5] that the presence of the global context of headword allows improving the synonyms' extraction. Histograms of the ranges' differences of two pairs of words, namely semantically analogous "school and academy" and dissimilar "school-baron" are shown on Fig. 3.
Fig. 3 shows the histograms of global contexts of similar and dissimilar words (1st line) that have a difference in kind, i.e. the left histogram is different from right, the relations of left and right bins at the histograms are different. In particular, for "school-academy" the height of left bin is higher than right bin, and for "school-baron" the case is vice-versa. The left bin is the amount of similar context words and the right bin is an amount of words that didn't match. The histograms of local context (2nd and 3rd lines) differ quantitatively, i.e. in the case of "school-baron" the height of left bin is smaller by an order of magnitude than the height of the right bin while for "school-academy" the height of the left bin is only about twice smaller than the right one. The mentioned features can be useful for thin and qualitive clusterization.

The results of algorithm testing presented on Figure 4 as the dendrogram of classes that contains the qualitative lists of synonyms: «булка, булочка, бутерброд, пирог» - "white bread, bun, burger, cake"; «бутылка, бокал, баночка, бутылочка, ампула» - "bottle, glass jar, little bottle, vial". However, some classes were formed in a wrong way: «бульон, ароматный, виноградный, апельсиновый» - "broth, fragrant, grape, orange" – different part of speech; «багажник, бумажник, аптечка» - "boot, wallet, first aid kit" – i.e. from semantically unrelated words.

IV. CONCLUSION

Thus, based on our results we conclude that:

1) Methods of clustering using an algorithm WEBSOM shows certain shortcomings;

2) Clustering method based on semantic maps also shows shows certain limitations;

3) Extraction algorithm is proposed to improve the methods considered synonymous with the help of the global context;

4) A numerical experiment to test the proposed method, which showed improved results semantic clustering due to the fact that takes into account the global context of the word and has a parameter generalization;

5) The proposed extraction algorithm synonym can be used to classify the topics of texts, as well as the construction of semantic concepts map.

The following problems must be solved in subsequent work to address the task of topic identification: 1) the identification of homonyms 2) processing of idiomatic expressions 3) improving the quality of word clustering.

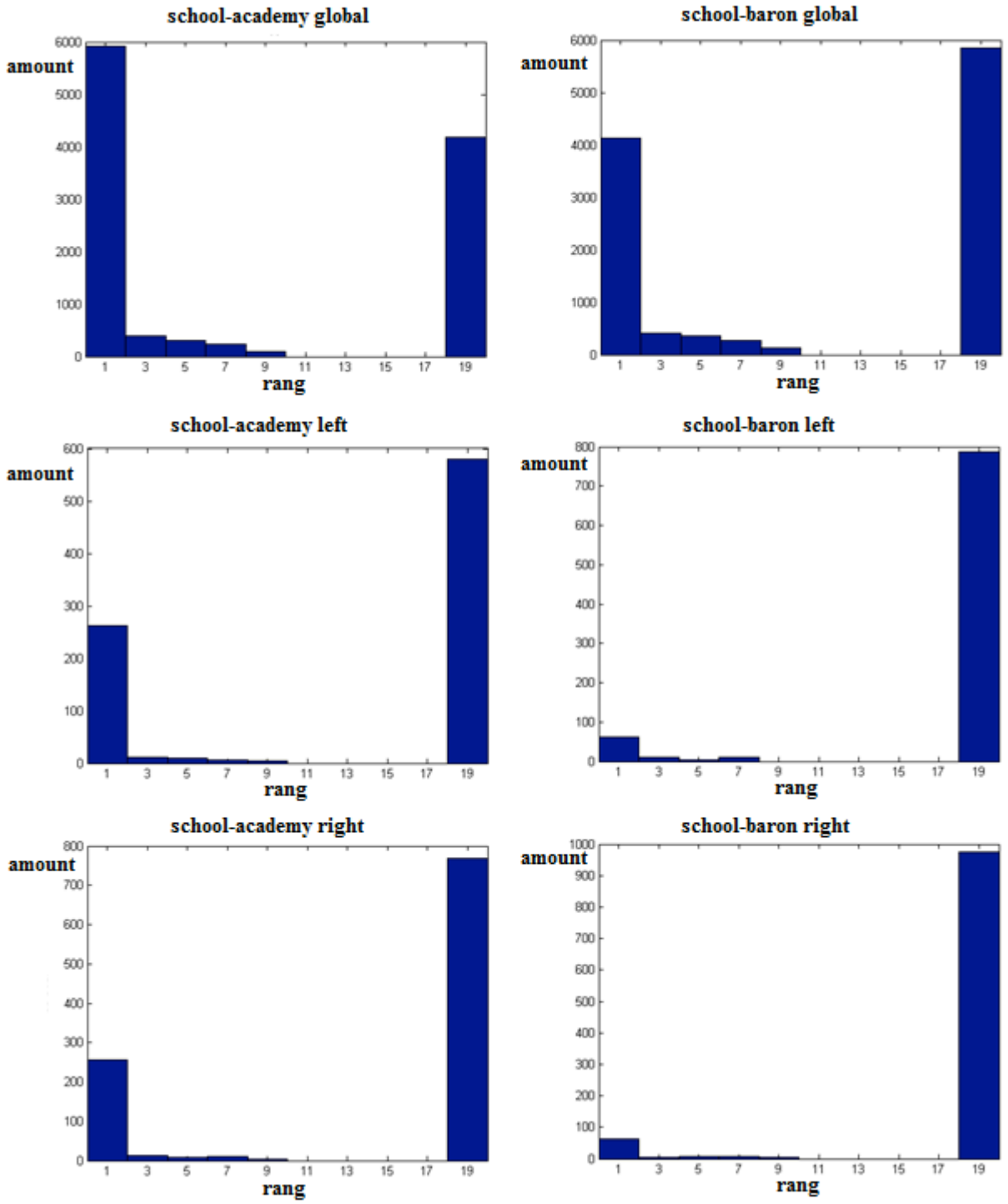

Fig. 3. The histogram of the distribution of ranges' differences of two pairs of words. The left-sided column of plots is for "school-academy". The right-sided column of plots is for "school-baron". The first pair of plots is global context, the second one is left-sided local context, the third one is right-sided local context.

ISBN: 978-1-4799-6376-8/15/$31.00 ©2015 IEEE

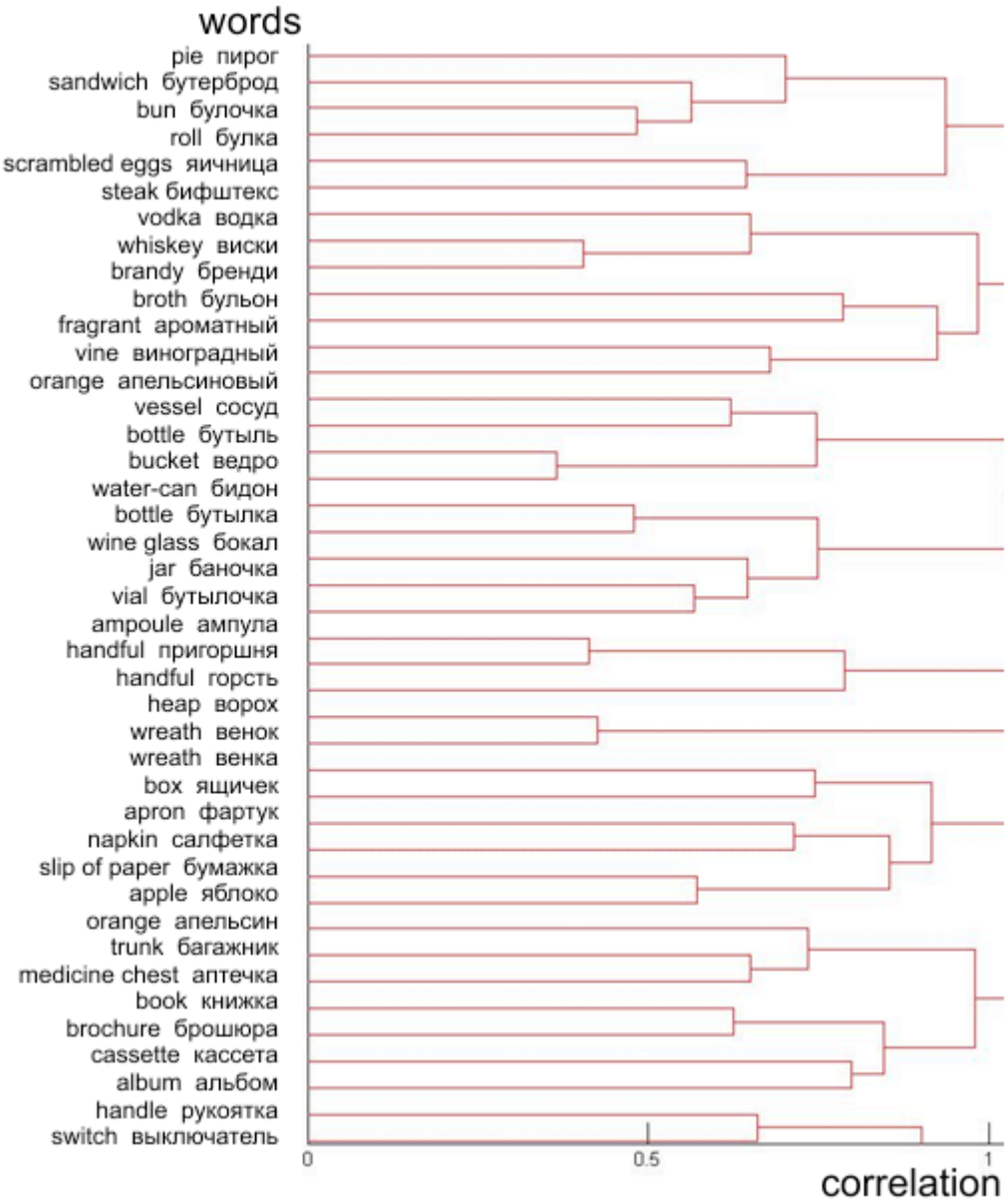

Fig. 4. The results of the extraction of lexical correlatives of Russian language.

REFERENCES

[1] T. Honkela, S. Kaski, K. Lagus, T. Kohonen, **"**Newsgroup Exploration with WEBSOM Method and Browsing Interface", *Social Media Group.* http://smg.media.mit.edu/classes/SocialVis03/honkela96tr.pdf

[2] С.А. Шумский, Глава 3. Язык и мозг: Как человек понимает речь. [авт. книги] Владимир Георгьевич Редько. Подходы к МОДЕЛИРОВАНИЮ МЫШЛЕНИЯ. Москва : УРСС, 2014

[3] P. Li, C. Burgess, K. Lund, "The acquisition of word meaning through global lexical co-occurrences", http://blclab.org/wp-content/uploads/2013/02/clrf00.pdf

[4] D.J. Hess, D. J. Foss, P. Carroll, "Effects of global and local context on lexical processing during language comprehension", *Journal of Experimental Psychology,* 124(1) , pp. 62-82. 1995

[5] E.H. Huang, R. Socher, C.D. Manning, A.Y. Ng, "Improving Word Representations via Global Context", http://nlp.stanford.edu/pubs/HuangACL12.pdf

[6] H. Zhao, Z.J. Qi, **"**Hierarchical Agglomerative Clustering with Ordering Constraints", http://dsl.cs.ucdavis.edu/~zhf/homepage/PID1045339.pdf