Semantic clustering of Russian web search results: possibilities and problems

Andrey Kutuzov

Mail.ru Group, National Research University Higher School of Economics

Abstract. The present paper deals with word sense induction from lexical co-occurrence graphs. We construct such graphs on large Russian corpora and then apply the data to cluster the results of Mail.ru search according to meanings in the query. We compare different methods of performing such clustering and different source corpora. Models of applying distributional semantics to big linguistic data are described.

1 Introduction

The presented paper deals with the problem of semantic clustering of search engine results page (SERP). The problem arises from the obvious fact that many user queries are ambiguous in some way. Thus, search engines strive to diversify their results and to present such results that are related to as many query interpretations as possible. For example, Google search for the Russian word *'максим'* returns:

1. five results related to a popular singer,
2. two results for a magazine,
3. one result for `http://lib.ru`, Maxim Moshkow's electronic library,
4. one result for a proper name.

However these results are not sorted by their meaning and are returned simply according to their relevance ranking, which for many of them seems to be almost equal. The obvious way to cluster the results is by the words their snippets share. Unfortunately, often snippets for results belonging to one query sense do not have a single content word in common (except for the query itself, which is useless). Cf. two snippets for the first query meaning from the example above:

1. *'МакSим начинает самостоятельно заниматься своей карьерой, пишет новые песни. В этот период певица выступает как малобюджетный проект, ...'*
2. *'МакSим презентовала видеоклип «Я буду жить», получивший широкую огласку еще до момента появления видео в сети.'*

They do not have a single common word, but still belong to one meaning (popular singer).

Moreover, snippets for different query senses can share some words. Cf. two snippets from the same search engine results page. They share the word *'автор'* ('author'), however the first snippet relates to the first meaning, while the second snippet shows the third one:

1. *'МакSим (Марина Абросимова) – одна из самых популярных и коммерчески успешных певиц в России, являющаяся* ***автором*** *и исполнителем...'*
2. *'Работает с 1994 года. Книги и тексты, разбитые по жанрам и* ***авторам****.'*

That means that there is a need for more sophisticated way to cluster search results. We should somehow learn which senses the query has and with which words these meanings are (probabilistically) associated. One of the possible ways to solve this problem is by extracting co-occurrence statistics from large corpora. The idea behind this is that word meaning is in fact the sum (or the average) of its uses. So, meaning is a function of distribution (cf. [1]). Thus, if we know with which words the query typically co-occurs and how these neighbors are related to each other, then we know the 'sense set' of the query. After that we can somehow measure semantic similarity of each search snippet on the SERP with each of the senses and map them to each other. This information can then be used to either rank the results, or mark them with appropriate labels.

The structure of the paper is as follows. In Section 2 we briefly overview work previously done on the subject. Section 3 describes the process of building co-occurrence graphs from large Russian corpora. In Sections 4 and 5 we conduct an experiment on clustering SERPs with ambiguous queries from Mail.ru search engine with the help of the methods described before. The results are evaluated in Section 6. Section 7 draws conclusions concludes and provides suggestions for further research.

2 Related Work

As stated in the previous Section, we are inspired by a fundamental hypothesis than meaning depends on the distribution [1] and that frequency of linguistic phenomena (in our case, word co-occurrence) is important for determining these phenomena's place in the system of language [2]. Our work is also based on the idea that the senses of ambiguous lexical units should be induced from the data itself, not from a dictionary. No dictionary is perfect or comprehensive, because *'senses as identified in the dictionary identify points on a continuum of possibilities for how the word is used'* [5]. The only robust source of words' meanings in the text is the text itself. That's why we shift our focus away from selecting the most suitable senses from a pre-defined inventory towards discovering senses automatically from the raw data, which is natural text.

One of the first notes on practical application of this idea to word sense disambiguation and word sense induction is found in [3], where vector representations of word similarity derived from co-occurrence data are used. Broad review of contemporary (by 2012) state of the field is provided in [4].

The main source of methods for our present research is [6], which describes workflow for clustering web search results using graph analysis over co-occurrence networks. Specifically, we use the notion of query graph, consisting of query terms and words from search engine results page augmented with nearest neighbors and relations from a reference corpus. For partitioning query graph and clustering

query senses we employed *Curvature* algorithm [6] and *Hyperlex* algorithm proposed in [7].

3 Building Co-Occurrence Graph

The first thing we had to do was to select a text corpus to build the graph upon. It is well known that the larger the corpus is the more co-occurrence information it contains. However, increasing corpus size also leads to exponentially growing computation time. Thus, for the sake of time and because of the preliminary nature of our research, we restricted ourselves to three Russian corpora of smaller but still decent size:

1. Open Corpora[1] (1 million tokens), further *'OC'*;
2. Disambiguated fragment of Russian National Corpus[2] (1 million tokens), further *'RNC'*;
3. Corpus of random search queries from Mail.ru search engine[3] (2 million tokens), further *'QC'*.

The first two items are academic corpora of Russian texts, supposedly representing (written) language in general. They differ in that the first one consists of full texts published under various free and open licenses, while the second one is a random sample of sentences from the larger Russian National Corpus. Both of them come with morphological annotation.

The third corpus was taken for comparison. It is important in view of the aim of our research (to test semantic SERP clustering). Our intuition was that perhaps query corpus provides more 'real-life' sense inventory. It is two times as big as its counterparts, because 'connectivity' between its members is lower (see Table 2) and we had to compensate for this.

At the same time, it turned out that the first two corpora mixed into one give better results, thus below we will often refer to such 'meta-corpus' as *'Mix corpus'*.

Before constructing the graph itself, we preprocessed the corpora, namely:

1. Removed from *QC* all queries which did not contain Cyrillic characters (as apparently they are not Russian),
2. Processed *QC* with Freeling analyzer [8] to extract lemmas and morphological information for all tokens,
3. Removed stop words,
4. Removed all tokens except nouns, as we restrict ourselves to inducing only nominal senses (the same strategy was applied in [6]).

Sizes of preprocessed corpora are given in Table 1. Average query length in QC is 2.47 noun tokens per query.

[1] `http://opencorpora.org`
[2] `http://ruscorpora.ru`
[3] `http://go.mail.ru`

Table 1. Sizes of corpora participating in the experiment

Corpus	Size (tokens)
OC	490671
RNC	294849
Mix	785520
QC	1035483

After the corpus has been built, the process of constructing co-occurrence graph is rather straightforward: we create an empty graph and then populate it with vertexes denoting word types in the text (lemmas). After that for each lemma we find all its immediate neighbors in the corpus, that is, words to the left and to the right (sentence boundaries not crossed, queries considered to be 'sentences' as well). If two lemmas were neighbors at least one time, we draw an edge between corresponding neighbors.

Finally, we have an undirected graph in which noun lemmas are vertexes and co-occurrence relations are edges. For each edge we also calculate Dice coefficient [9]. It measures the 'strength' of the collocation, based on absolute frequency (c) of both words (w and w') and collocation (w,w'):

$$Dice(w,w') = \frac{2c(w,w')}{c(w) + c(w')} \tag{1}$$

One can also think about the graph as a matrix of Dice coefficient values for all possible pairs of lemmas in the corpus.

Table 2 gives an overview of the basic features of the graphs.

Table 2. Parameters of the graphs

Corpus	Vertexes	Edges	Average degree	Average path length	Clustering coefficient
OC	21881	257846	23.57	3.26	0.166
RNC	22467	163914	14.6	3.53	0.136
Mix	31984	395225	24.7	3.29	0.186
QC	85548	291033	6.8	4.07	0.16

One can see that the average degree of *QC* is lower in comparison with the other corpora (because queries are typically shorter that sentences in natural texts). That is one of the reasons for our decision to use a larger query corpus.

It should also be noted that all corpora comply to 'small world' definition [10], because their average path length is approximately the same as in a random graph with the same number of vertexes (N_V) and average degree (A_D), while clustering coefficient is significantly higher than it should be in the random graph.

For example if *Mix* corpus were a random one, its average path length would be equal to 3.24 ($= \frac{log(N_V)}{log(A_D)}$), very close to the actual value. However, in this case, its clustering coefficient should be 0.0015 ($= \frac{2 \times A_D}{N_V}$), which is significantly lower than the actual value. The same is true for all other corpora.

'Small world' nature of our graphs means that vertexes in them tend to bundle into clusters, which is typical of many real-world networks. This finding supports the idea of extracting senses from such clusters. It also additionally proves the applicability of graph sense induction methods to our corpora, as English-language graphs in the related publications also showed such properties.

4 Building Query Graph

We experimented with clustering search engine results page on a set of sixty ambiguous one-word Russian queries, taken from *Analyzethis* homonymous queries analyzer[4]. *Analyzethis* is a search engines evaluation initiative, offering various search performance analyzers, including one for ambiguous or homonymous queries. We crawled Mail.ru search for these queries, getting titles and snippets (10 for each result).

The procedure of semantic clustering starts with building the so called query graph. Here we closely follow [6].

First, we lemmatize all snippets and titles and remove stop words and the query word itself. Then we construct a graph G_q with all nouns from snippets and titles as vertexes. Then we use one of the large corpora graphs (those that we built in Section 3) to find words strongly connected to the query word and add these words to the query graph. We consider a connection 'strong' if it falls under the following constraints:

$$\begin{cases} \frac{c(q,w)}{c(q)} \geq 0.01 \\ Dice(q,w) \geq 0.005 \end{cases} \tag{2}$$

where c is absolute frequency in the corpus, q is the query and w is the word under analysis. Thresholds 0.01 and 0.005 were determined empirically while experimenting on the above mentioned ambiguous queries set. These thresholds produced most convincing sense clustering. However, the issue of choosing the thresholds is a subject for thorough evaluation in future.

Thus, we now have G_q with no edges and vertex set consisting of words from the search result and strong neighbors of the query word. After that, for each pair of words (w,w') in G_q we check if they co-occur in the large corpus. If they do and $Dice(w,w') \geq 0.005$, we connect these words in G_q with an edge with weight = *Dice(w,w')*. Finally, we delete disconnected vertexes (those with the degree equal to 0).

[4] `http://analyzethis.ru/?analyzer=homonymous`

5 Processing Query Senses and Results

With query graph at hand, we are ready to find which senses the query has. What we need is an optimal partition of the query graph, in which words related to different senses are in different parts of the graph. We apply two techniques for that, namely, *Curvature* from [6] and *Hyperlex* from [7].

5.1 Curvature

Curvature algorithm aims at finding vertexes from G_q with low local clustering coefficient. Our hypothesis is that these are words which serve as 'links' between different senses or 'uses' of the query. Then we remove vertexes with clustering coefficient below a certain threshold. It leads to the graph disjointing into several components related to different senses. Vertexes in these components represent lexical inventory of each sense. Disconnected vertexes are removed from the final graph.

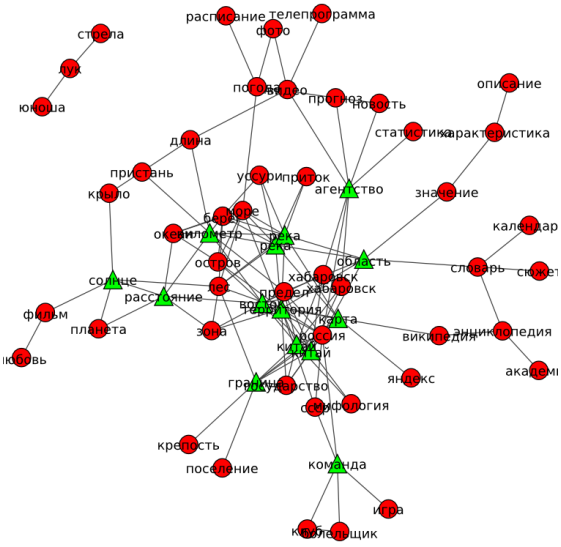

Fig. 1. Query graph for *'амур'* (*Curvature*)

Let us illustrate the process with the example of *'амур'* ('Amur') query. Figure 1 shows its query graph. It is already disconnected into two components and the meaning of love god (associated with words *'лук'*, *'стрела'* and *'юноша'*) is separated. However, other 'senses' of the query remain hidden in the giant component. Vertexes shown as triangles have low clustering coefficient and are thus marked for deletion.

So, we delete 'triangular' vertexes. Note that we chose threshold 0.3 – all vertexes with clustering coefficient below this are removed. It is also important that we do not delete vertexes with clustering coefficient = 0. This is because neighbors of such vertexes are not connected to anything except this vertex. If we remove it, a lot of disconnected vertexes will appear. Such clusters (consisting of only one word) do not make much sense. For example, the word *'лук'* on Figure 1 is characterized by clustering coefficient = 0. If we remove it, then the whole component representing 'love god' meaning disappears.

Figure 2 shows the query graph after removing vertexes with low clustering coefficient. We now have 6 components (note that the labels for these clusters are introduced by us, not by the algorithm):

1. River (all vertexes except enumerated below)
2. Love god (*'юноша, лук, стрела'*)
3. Hockey club (*'клуб, болельщик'*)
4. Movie (*'любовь, фильм'*)
5. Dictionary-1 (*'календарь, словарь, википедия, энциклопедия, академик, сюжет'*)
6. Dictionary-2 (*'значение, описание, характеристика'*)

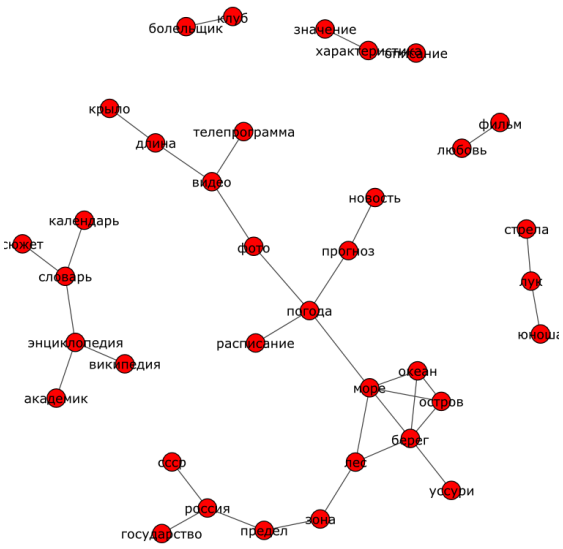

Fig. 2. Disjointed query graph (*Curvature*)

First 4 components clearly represent different meanings of the word *'амур'*. The last two are rather 'uses', typical contexts. However they can still be useful in clustering as they allow to keep encyclopedic results together.

5.2 Hyperlex

Hyperlex algorithm described in [7] introduces the notion of 'hubs' within the graph, meaning most inter-connected vertexes and employs the graph's maximum spanning tree. Just like the previous algorithm, it takes as an input the query graph G_q we prepared in Section 4 and the query itself.

First we create a list **L** with all vertexes from G_q sorted in decreasing order by their absolute frequency in the large corpus. Then for each item of this list we check if the corresponding vertex complies to the following constraints:

1. Vertex normalized degree is greater than or equal to 0.05,
2. Average Dice coefficient of vertex edges is greater than or equal to 0.007.

If the constraints are met, we add this word to the hub list, considering it to be a kind of a connector. Simultaneously, we remove this vertex and its neighbors from the list **L** and continue iterating. In case we meet a word which does not

satisfy the requirements above, we check whether the list of hubs has at least two elements. If it does, we stop iterating, if not, we continue to the next item. Note that it differs from the original *Hyperlex* algorithm, where one should stop no matter how long the hub list is. In our Russian material it sometimes caused the hub list to remain empty or contain only one item, which is useless.

After we have the list of hubs, we augment G_q with query vertex and connect this vertex to all hubs putting infinite (or very high) Dice coefficient on the corresponding edges. Then, we produce a maximum spanning tree from this graph. Maximum spanning tree is an attempt to keep all the vertexes connected while eliminating cycles and using as few edges as possible with as high weights (in our case it is Dice coefficient) on them as possible. In the spanning tree, there is only one path between any two vertexes and this path lies through edges with maximum Dice. Because the query vertex and the hubs are connected by edges with infinite Dice, they are sure to be the center of the spanning tree and directly linked.

At last we remove the query vertex from the spanning tree, producing disjointed subtrees with hubs as roots. These subtrees represent query meanings. Note that we also delete all disconnected vertexes (those with degree = 0).

Let us present an example of Hyperlex at work with the same query *'амур'*. Our corpus is *Mix*. Initial state of the query graph G_q is the same as in Figure 1.

We add the word *'амур'* to the graph G_q and connect it to vertexes selected as hubs: *'область, фильм, команда'*. The result is presented on Figure 3 with query vertex drawn as a diamond. For reference, vertexes which were introduced from the corpus and not from search results (*'сюжет'*, *'океан'*, etc) are drawn as triangles.

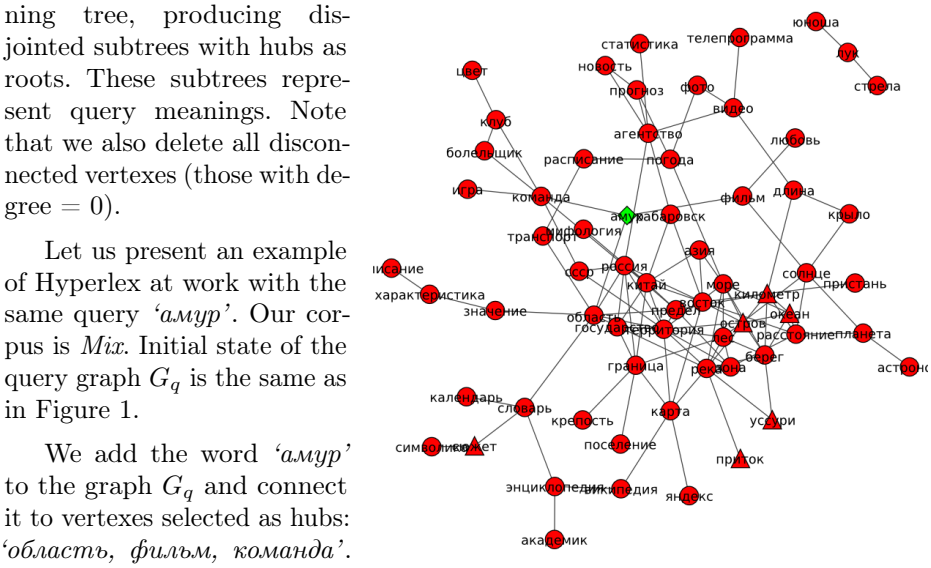

Fig. 3. Query graph for *'амур'* after augmenting it with the query vertex (*Hyperlex*)

Now we produce maximum spanning tree from G_q with Dice coefficient as weight measure. The tree is visualized on Figure 4. Note that it has much fewer edges than the initial G_q.

Finally, we remove the query vertex and all vertexes that become disconnected after this removal. As a result, we have a disjointed graph shown in Figure 5. The number of the components has grown from 2 to 4 (once again, labels are assigned by us):

1. Love god (*'юноша, лук, стрела'*)
2. Movie (*'любовь, фильм'*)
3. Hockey club (*'клуб, игра, болельщик, команда, цвет'*)
4. River (all the remaining vertexes)

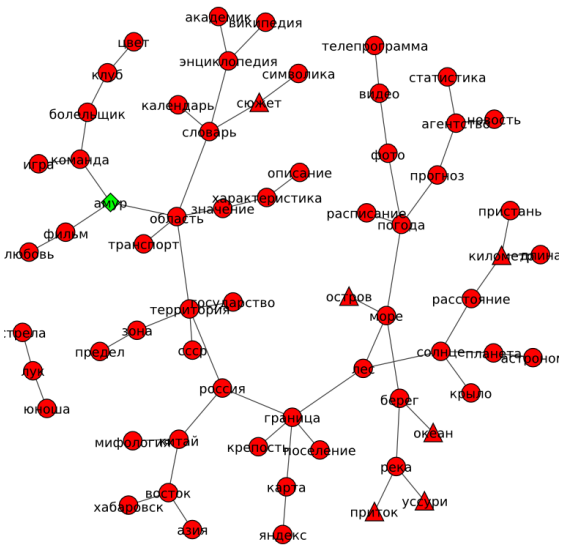

Fig. 4. Maximum spanning tree for *'амур'* query graph (*Hyperlex*)

One can see that *Hyperlex* successfully extracted the same four important meanings as the previous algorithm. At the same time, unlike *Curvature*, it managed to avoid two 'encyclopedic' clusters (obviously in common for too many queries) and leave their vertexes in the 'river' cluster. Also, *Hyperlex* is better because it describes 'hockey club' cluster in a richer way, using 5 relevant words instead of 2.

One can again note that in fact what we call 'senses' are not senses like meanings in the dictionaries. We agree with Jean Veronis who argues that co-occurrence networks reflect 'uses' rather than senses. So, what we have are typical environments where the word is used, and these environments are only loosely connected to what a lexicographer would call 'senses' or 'meanings'. However, we are fine with that, as we assume that clustering SERP according to 'typical uses' is at least equally important as clustering according to 'proper senses'. Perhaps, these senses are in fact less related to real-life, as even linguists sometimes have trouble matching the 'senses' found in a dictionary and the occurrences found in a corpus [7]. Additionally, as has already been stated, dictionary senses are always limited and by design cannot cover new semantic trends and subtle meanings quickly appearing and disappearing in the modern world. Thus, theoretically typical uses are more relevant for clustering than academic dictionary senses. To strictly prove it for the Russian material, one needs manually clustered data set (see Section 6), and we leave it for further research.

5.3 Mapping Results to Senses

Once we possess the sense inventory for the query, we can combine it with bags-of-words for each search result to finally perform SERP clustering. We do that in a rather straightforward way.

Given a set of senses represented by a lemma set each and a set of results (snippet and title) represented by lemma sets as well, for each pair of result (r) and sense (s) we calculate similarity measure sim. It is a simple number of lemmas in common for both sets divided by the number of lemmas in the result:

$$sim(r,s) = \frac{r \cap s}{length(r)} \tag{3}$$

Then we choose the sense with maximum similarity and link this sense to the result. Thus, each result receives some sense, and is 'understood'.

In the future we plan to explore other means of calculating similarity measure as well, for example, counting tokens not types or considering weights on edges in the intersection.

6 Evaluation of SERP clustering

Generally, evaluation of clustering is a rather harsh task. Perhaps, the best way to do this is to employ human assessment, but for the time being we limited ourselves to simple evaluation of the correctness of cluster number (that is, number of meanings).

Analyzethis service provides data about how many senses of an ambiguous query are there in the SERP. Thus we consider it to be an expert opinion and check how strong is our deviation from this 'gold standard'. For example, if *Analyzethis* believes that there are three senses present on the SERP, and our clustering algorithm puts all the results into one cluster, this signals that the algorithm is not optimal. The same is true if the number of clusters is, for example,

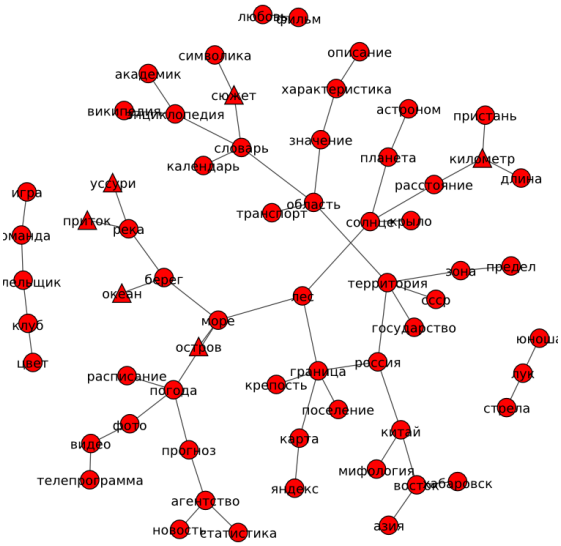

Fig. 5. Maximum spanning tree after removing query vertex (*Hyperlex*)

eight. The less our deviation from *Analyzethis* assessment is, the better. So, in fact we check that the employed algorithms do not produce senseless results (too many or too few meanings). We once again note that in order to evaluate the contents of the clusters themselves, one needs manually clustered SERPs for ambiguous queries. To our knowledge, there is no such a data set for Russian. We are working on creating it.

For the time being, we compared the number of clusters for each of ambiguous queries in four different settings (two corpora and two word sense induction methods). Then we calculated average deviation of our clustering number from that of *Analyzethis*. Table 3 provides the results of this comparison. Note that the average number of senses per query in *Analyzethis* data set was 2.65.

Table 3. Evaluation of SERP clustering (average deviation from *Analyzethis* assessment in number of senses and in percent from the average number of senses in the set)

Corpus	**Curvature**	**Hyperlex**
Mix	1.636 (61%)	**1.288** (49%)
Query	1.742 (66%)	1.379 (52%)

It is clear that *Hyperlex* consistently outperforms *Curvature*, and that *Mix* corpus does the same with the query corpus. *Hyperlex* victory comes as no surprise, as it uses maximum spanning tree notion, which seems to allow deeper grasping of graph structure. The victory of *Mix* corpus (which is smaller than the query corpus) is much less expected. We believe that there are two reasons for this:

1. As we have already mentioned, the query corpus is less 'dense' because of low length of queries. Thus, there are fewer edges and less data for algorithms.
2. Query corpus was lemmatized with Freeling while *Mix* corpus consists of manually annotated corpora. Glitches and outright errors of Freeling could impact graph quality. This can be fixed in the future either by improving Freeling or by using another lemmatizer.

Thus, at the moment, using *Mix* corpus and *Hyperlex* algorithm of word sense induction seems to be the best option. However, things surely can be different if we employ larger corpora (which we plan to do in the future).

7 Conclusion and future work

We showed that state-of-the-art methods of word sense induction and search results clustering based on semantic graphs do work for Russian data.

Application of such methods can lead to search engine results presentation getting closer to actual semantics of the results, not simply term frequency

ranking. For a user, it would mean the possibility to immediately grasp which results in the SERP are actually related to the query sense, and which other senses exist. The power of this approach can be increased by wider employment of Semantic Web paradigm: semantically marked up web pages are represented by generally better and clearer snippets. Such snippets, in turn, should provide better data for graph-based word sense induction algorithms.

We plan to experiment on more types of query graph processing and launch a full-scale human evaluation of results. Also, it seems profitable to use not only separate words, but also compound phrases, as well as to construct graphs with not only immediate neighbors, but also with second-order co-occurrences (neighbors of neighbors). Additionally, experiments with larger query corpora may lead to new and inspiring insights in this field.

References

1. Harris, Z.S.: Distributional structure. Springer (1970)
2. Bybee, J.: Frequency of use and the organization of language. Oxford University Press, USA (2006)
3. Schütze, H., Pedersen, J.O.: Information retrieval based on word senses. In: Proceedings 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR 1995). (1995) 161–175
4. Navigli, R.: A quick tour of word sense disambiguation, induction and related approaches. In: SOFSEM 2012: Theory and practice of computer science. Springer (2012) 115–129
5. Kilgarriff, A.: Dictionary word sense distinctions: An enquiry into their nature. Computers and the Humanities **26**(5-6) (1992) 365–387
6. Marco, A.D., Navigli, R.: Clustering and diversifying web search results with graph-based word sense induction. Computational Linguistics **39**(3) (2013) 709–754
7. Véronis, J.: Hyperlex: lexical cartography for information retrieval. Computer Speech & Language **18**(3) (2004) 223–252
8. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, European Language Resources Association (ELRA) (may 2012)
9. Smadja, F., McKeown, K.R., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics **22**(1) (1996) 1–38
10. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. Nature **393**(6684) (1998) 440–442