Tomsk State University of Control Systems and Radioelectronics

ТУСУР TUSUR UNIVERSITY

RITSUMEIKAN

IICST 2013

INNOVATIONS IN INFORMATION AND COMMUNICATION SCIENCE AND TECHNOLOGY

Third Postgraduate Consortium
International Workshop
IICST 2013

Tomsk State University of Control Systems and Radio Electronics
Ritsumeikan University

Innovations in Information and Communication Science and Technology

Third Postgraduate Consortium
International Workshop
IICST 2013
Tomsk, Russia, September 2–5, 2013
Proceedings

Tomsk
Publishing Office of Tomsk State University of Control Systems
and Radio Electronics
2013

Organization

The 3rd Postgraduate Consortium International Workshop, IICST 2013, was organized by the following members, with support from Tomsk State University of Control Systems and Radio Electronics (TUSUR), Russia, and Ritsumeikan University, Japan.

Workshop Chairs

Victor V. Kryssanov, Ritsumeikan University
Alexander F. Uvarov, TUSUR

Organizing Committee

Chairs

Gennady A. Kobzev, TUSUR
Hitoshi Ogawa, Ritsumeikan University

Members

Kozaburo Hachimura, Ritsumeikan University
Hajime Murao, Kobe University
Frank Rinaldo, Ritsumeikan University
Maria Afanasyeva, TUSUR
Konstantin Nefedev, FEFU
Fumio Hattori, Ritsumeikan University

Program Committee

Chair

Eric W. Cooper, Ritsumeikan University

Tutorials and Organized Session Chair

Ross Walker, Ritsumeikan University

Members

Ikuko Nishikawa, Ritsumeikan University
Igor Goncharenko, 3D Inc.
Koh Kakusho, Kwansei Gakuin University
Uwe Serdült, University of Zurich
Joo-Ho Lee, Ritsumeikan University
Machi Zawidzki, University of Tokyo
Njuki Mureithi, Polytechnique Montreal
Evgeny Shandarov, TUSUR
Yugo Hayashi, University of Tsukuba
Tomasz M. Rutkowski, University of Tsukuba
Mikhail Svinin, Kyushu University
Hiroyuki Shinoda, University of Tokyo
Yoshio Nakatani, Ritsumeikan University
Hideto Ikeda, CyberPro, Ltd.
Evgeny Kuleshov, FEFU
Hajime Murao, Kobe University
John C. Wells, Ritsumeikan University
Hironori Hibino, TRI JSPMI
Yutaka Kanou, Soft Cube Co., Ltd.
Toshiya Kaihara, Kobe University

Industry Advisory Committee

Chair

Vladimir Bykov, VoTa PR Group & Foundry

Sponsoring Institutions

Tomsk State University of Control Systems and Radio Electronics, Russia
Ritsumeikan University, Japan

Table of Contents

Innovations in Information and Communication Science and Technology
Third Postgraduate Consortium International Workshop
E. Cooper, G.A. Kobzev, A.F. Uvarov, and V.V. Kryssanov (Eds.)
IICST 2013: pp. 65-72.

CLUSTERING WORDS WITH SIMILAR SENSE *USING INFORMATION ABOUT THEIR SYNTACTIC DEPENDENCIES*

Klyshinsky[1] E.S., Kochetkova[1] N.A., Logacheva[2] V.K.

[1] *Moscow Institute of Electronics and Mathematics of Higher School of Economics, eklyshinsky@hse.ru*
[2] *Keldysh Institute of Applied Mathematics RAS, Russia*

ABSTRACT

In this article we report some new experiments in the area of words clustering for the Russian language. We introduce a new clustering method that distributes words into classes according to their syntactic relations. We used a large untagged corpus (about 7,2 bln of words) to collect a set of such relations. The corpus was processed using a set of finite state automata that extracts syntactically dependent combinations having explicit structure. These automata were used to process only unambiguous text fragments because of combination of these techniques increases the quality of sampled input data. The modification of group average agglomerative clustering was used to separate words between clusters. The sampled set of clusters was tested using one of the semantic dictionaries of the Russian language. The NMI score calculated in this article is equal to 0.457 and F_1-score is 0.607.

Key words: Word clustering, Syntactic dependencies, Russian language.

1. INTRODUCTION

Natural language can be formally represented as a set of language units and a set of rules of combination of these units. Therefore in order to analyze a text one has to know these rules. That is how the first natural language processing (NLP) applications worked: they used explicit sets of rules created manually by experts. However, such approaches are usually laborious and time-consuming, so state-of-the-art NLP applications try to automate this process: they extract rules (in explicit form or represented as probabilistic distributions) from the instances of language (words, monolingual or parallel texts, audio records etc.).

In (Harris, 1954) Harris uttered a hypothesis that is quite in line with these contemporary tendencies in NLP. He stated that words that are used in similar contexts are likely to have similar meaning. This thesis is fairly controversial, and there are a lot of examples supporting as well as disproving it. Nevertheless we state that relying on this thesis one can get positive results.

In the paper we follow the Harris' hypothesis and perform semantic words clustering using the information on words context, namely the information on their syntactic relations. Following Manning (Manning, 1993) we extract the syntactic relations from untagged text corpora without using full-fledged parsing system. Thus we save time and resources and make our method applicable to under-resourced languages. Moreover, since we use only unambiguous words and syntactic relations, our list of relations is highly precise. Unfortunately, these limitations do not allow for achieving high recall.

For every unambiguous word from the corpus we extract a dictionary of words that are syntactically related to it. On the basis of this dictionary we cluster words. Our quality assessment shows that this clustering agrees with the actual semantic partitions.

This paper is organized as follows. Section 2 contains a brief review of works in words clustering. In section 3 we report our method of syntactic relations extraction and our clustering algorithm. Section 4 outlines the used data and experimental setup. Section 5 contains the results of experiments with the Russian language. In section 6 we discuss the strong points and the drawbacks of our method and point on the ways of further development.

2. RELATED WORK

The clustering of words into semantic classes can be useful in many NLP tasks. When developing or augmenting ontologies and thesauri (Volkova, 2013) word clustering allows to perform operations on clusters instead of single words, which results in faster ontology development. According to our experiments a linguist creating an ontology or thesaurus can increase the speed of development by two or three times if provided with word clusters. Moreover, word clustering can be used to check correctness and consistence of existing ontology. One of the first researches in the field of clustering of words is the work (Pereira *et al.*, 1993).

During parsing of a sentence the information on words co-occurrence or sub-categorization frames can help define correct links between words or refine the word's role in the sentence. The same information can be useful in part-of-speech homonymy resolution. The use of word's co-occurrence information can improve the quality of POS-tagging by 1-3%, increasing it from 95% to around 98%.

Moreover, word clustering based on partial co-occurrence information can help complete the gaps in it and build more proper word subcategorization frames (Manning, 1993; Korhonen *et al.*, 2003; Korhonen *et al.*, 2006). The research (Baroni and Lenci, 2010) provides a thorough review of various applications of word clustering and word co-occurrence information.

Any objects are usually divided into clusters defined by the means of some similarity measure. Therefore objects have to be represented as a set of feature vectors, and a similarity measure for a pair of objects depends on how close are the values of their corresponding features.

For the task of clustering the majority of researchers represent words as lists of their possible syntactic relations, hence two words are close if they can interact with the same words. Baroni (Baroni and Lenci, 2010) distinguish two types of relations between words, or more specifically between concepts denoted by these words:

- Attributional similarity: two words which are synonyms, co-hypernyms or hypernyms share the large number of properties or possible actions;
- Relational similarity: two words do not belong to one class, but are in some relationship, for example, “has-part”. In such a case they are likely to co-occur often in texts.

Notice that synonyms that belong to different styles can have different related words expressing the same meaning. For example, words doctor and healer are pretty similar. Therefore the second word has more chances to be met in science fiction or historical texts rather than in news texts.

Conversely, words that are not related semantically can have similar sets of co-occurrences and sub-categorization frames. For example, words “newspaper” and “police” have a lot of words in common but are not semantically similar. We have found about 300 adjective attributes shared between these words in Russian language. Most of them are related to geographical areas (i.e. names of city or country), government or social institutions, emotional expressions. We also have found that situation is quite similar in English. Therefore we are not always able to cluster words based on their syntactic relations and contexts. However, in most of cases this information is fairly reliable, and similar sub-categorization frames indicate similar meanings.

There are several methods of extraction of word’s context information. The simplest method is a bag-of-words model, used, for example, in (Li and Abe 1998). It is based on the assumption that if two words are close in text they are semantically related. The area of similarity is set to a fixed window of 2 to 5 words (Boleda at al., 2005) or expanded to the whole sentence (Bassiou and Kotropoulos, 2011) and even up to 1200 characters (Schütze, 1992). The bag-of-words approach is favoured by many researchers for its simplicity and the possibility to get different types of relations between words which are quite far in the text. On the other hand, it generates much noise. Thereby in many works simple bag of words is replaced by the results of shallow parsing (Preiss, 2007; Baroni and Lenci, 2010). It offers information of higher quality, but requires more time. Moreover, the output of state-of-the-art parsing systems contains at least 5-10% of erroneous relations, which is quite a high rate. In addition to parsing, some researchers use the information about word’s role in a sentence, which is achieved by semantic analysis (Baroni and Lenci, 2010). However, the latter is often even less reliable than parsing.

Some methods use a trade-off between bag of words and full syntactic information about a sentence. They extract words’ relations using lexical templates. These are regular expressions that allow one to define relations without full parsing. Therefore, in contrast to bag of words the output is not so noisy, and in the same time it does not require so much resources as parsing. The lexical templates afford to achieve high quality of extracted relations (95-98% correct relations), however, the recall is much lower than that of other methods (20-40% of concerned text). Moreover, the part-of-speech homonymy, which exists virtually in any language, often hinders the relation extraction. Nonetheless, we have found means to avoid this problem while processing texts in Russian.

There are works on syntactic relations extraction for different parts of speech: verbs (Schulte im Walde, 2000), nouns (Hindle, 1990), and adjectives (Boleda and Alonso, 2003). The majority of works consider only English, but there are also researches for Russian (Mitrofanova, 2009), German (Bohnet, 2002), Catalan (Boleda *et al.*, 2005), and other languages.

3. CLUSTERING METHOD

3.1 Input Data and Feature Set

The input of our clustering algorithm is the information on co-occurrence, extracted from big untagged corpora. Our paper (Klyshinsky *et al.*, 2011) describes the method to acquire the database of selectional preferences (i.e. database of words co-occurrences, extracted from the corpora, that shows the tendency of a head word to fill slots in its valency pattern with particular words) and the extraction process. We collected these selectional preferences using the chunking system: it used a set of templates to extract noun phrases, verb phrases, and prepositional phrases. We extracted only groups of consecutive unambiguous words, and chose the templates that ensured unambiguous relations. Here are some examples of applied templates:

<s> [NP|PP]$_U$ V – the only prepositional or noun phrase in the beginning of the sentence that precedes the only verb phrase in the sentence;

Prep Adj* Noun – adjectives that have a preposition in position before and noun in position after them.

Here **NP** and **PP** are noun phrase and prepositional phrase, respectively, “*” means one or more occurrences of an instance, subscript “$_U$” means that instance of this type occurs in the sentences only once, and <s> means that considered text starts a sentence. The first template allows us to assume syntactic connection between the noun (prepositional) phrase and the verb, the second one draws connection between the noun and the adjective(s).

These and some other templates were used to extract the following tuples of syntactically related words: <verb, noun>, <verb, preposition, noun>, <noun, adjective>, where the first word is the head of relation and the second and third are the attributes. Notice that in the text these pairs (triplets) can be met in different order (it depends on the used template) and in different forms (nouns and adjectives can be in different case and number, verbs – in different tense, number, and person). After extraction they are all normalized. For example, we can apply the regular expressions **<s> PP PRP V** and **V NP**, where **PRP** is a personal pronoun, to the sentence “For acquisition we use the original method” and acquire the following co-occurrences: “use for acquisition”, “use method”, “method original”. We also calculate the frequency of every tuple.

The described process often cannot be performed because of part-of-speech homonymy. Templates cannot be applied with certainty to phrases with POS-ambiguous words as it increases the chance of extracting wrong relations. Many state-of-the-art POS-taggers perform POS-homonymy resolution, but even the most high-quality taggers are not error-free. Moreover, due to noun and verb inflection in Russian the well-known tagging algorithms like n-gram tagging (Brill, 1995; Zelenkov *et al.*, 2005) or using a decision-tree (Schmid, 1994) perform much worse in Russian and in other synthetic languages much more than in English.

However, we have found a compromise solution that does not make us add errors to the output by using POS-taggers or overhaul the tagging algorithms. Russian language affords us to extract a sufficient number of relations using only non-homonymous words. According to our studies (Klyshinsky and Kochetkova 2012) approximately 75-85% of words in Russian do not have part-of-speech homonymy (see the comparison of homonymy in English and Russian in Table 1). Therefore we can accumulate a big set of syntactically related words applying our syntactic templates only to words without part-of-speech homonymy. Notice that such words do not need to be unambiguous: for example, a word can be either a noun in Nominative case or a noun in Accusative case. We account it unambiguous if all the tagging variants recognize it as a noun.

Table 1. Comparison of homonymy types in Russian and English.

POS tagger output	Russian News		English News
	lenta.ru 2005-2009	**compulenta.ru 2001-2009**	**Reuters 2009**
Unambiguous word	48,28%	46,55%	38,87%
Out-of-vocabulary word	4,38%	4,28%	7,65%
Part-of-speech homonymy	5,26%	6,33%	50,35%
Multiple sets of grammatical parameters	27,68%	28,22%	2,79%
Multiple initial forms of word	4,67%	4,01%	0,32%
Combined homonymy (e.g. POS homonymy + initial forms homonymy)	9,92%	10,61%	0,96%

Our previous experiments showed that extraction of relations for unambiguous words gives us a set where 90-99% (depending on type of relation) relations are correct. However, the reverse side of the high quality is the low recall. In this connection we have to extract relations from huge text corpora (several billions of words).

3.2 Similarity Measure and Feature Vectors

The extracted set of syntactically related words can be formally defined as a set of tuples $\mathbf{B}=\{<w_m, w_s, f>\}$, where w_m is a head of relation, w_s is an attributive word or words (for example, “preposition+noun”), f – tuple’s frequency in the analysed text . Therefore we can construct a dictionary of attributive heads $\mathbf{D}^+$ (direct dictionary). It contains all w_m words from **B**. In addition, we construct a dictionary of attributive word $\mathbf{D}^-$ (reverse dictionary), which contains all attributive words (w_s) from **B**. Let we consider $\mathbf{D}^+$ and $\mathbf{D}^-$ as vectors.

Therefore we define feature vectors $\mathbf{v}^+_a$ and $\mathbf{v}^-_a$ for every word. These vectors are formally defined as follows: $\mathbf{v}^+_a$ is a vector $<v_1, v_2, \ldots, v_n>$, where $n=|\mathbf{D}^-|$ and $v_i = \{f_i$ if **B** has an entry $<a, d^-_i, f_i>$, 0 otherwise$\}$, where d^-_i is i-th element of $\mathbf{D}^-$.

In other words, $\mathbf{v}^+_a$ contains counts of co-occurrences of every attributive words with the headword a in the considered corpus. The dimension of the vector is equal to $|\mathbf{D}^+|$. If an attributive word was not met with the word a, the corresponding position in the vector is set to 0.

The vector $\mathbf{v}^-_a$ is defined analogously on $\mathbf{D}^-$ dictionary. Vectors $\mathbf{v}^+_a$ and $\mathbf{v}^-_a$ are used as a feature vectors. Therefore the similarity of words is defined by cosine similarity measure counted on feature vectors (Manning 2008). In accordance with this, clustering is performed in $|\mathbf{D}^+|$ or $|\mathbf{D}^-|$-dimensional space of features.

According to our experiments, the use of raw co-occurrence frequencies results in multiple errors in calculation of similarity measure. For example, words “Zealand” and “Orleans” will be classified as similar, as they collocate with the word “New” in most cases and are rarely met with any other words. In order to get rid of such errors we replace the word count with its logarithm. It smooths the difference between common and rare words, hence their contribution to the similarity measure justifies.

3.3 Data Standardization and Clustering Method

In order to reduce the error rate it is recommended to standardize the used data by filtering some objects out. Therefore we filtered out all words combinations which frequency is lower than some pre-defined threshold α. This step should reduce the level of the noise in the input word combinations. We also excluded the words that do not have enough dependent words, i.e. the number of non-zero elements in the feature vector of each considered word in the database **D** should exceed some pre-defined threshold β.

The data filtering is performed as follows. Firstly, we collect the set of feature vectors from the database **B**. Let $\mathbf{S}^+ = \{\mathbf{v}^+_a\}$, $a \in \mathbf{D}^+$. Thus the feature vector is defined as $\mathbf{v}^+_a = <v_1, v_2, \ldots, v_n>$, $n=|\mathbf{D}^-| : v_i = \{f_i$ if **B** has an entry $<a, d^-_i, f_i>$ and $f_i \geq \alpha$, 0 otherwise$\}$, where d^-_i is i-th element of $\mathbf{D}^-$. The $\mathbf{v}^+_a \in \mathbf{S}^+$ only if $NZCount(\mathbf{v}^+_a) > \beta$, where $NZCount(\mathbf{v}^+_a)$ function returns the number of non-zero elements of $\mathbf{v}^+_a$. The set of $\mathbf{v}^-_a$ is defined analogously: $\mathbf{S}^- = \{\mathbf{v}^-_a\}$, $a \in \mathbf{D}^-$, $\mathbf{v}^-_a = <v_1, v_2, \ldots, v_n>$, $n=|\mathbf{D}^+| : v_i = \{f_i$ if **B** has an entry $<d^+_i, a, f_i>$ and $f_i \geq \alpha$, 0 otherwise$\}$, where d^+_i is i-th element of $\mathbf{D}^+$. The $\mathbf{v}^-_a \in \mathbf{S}^-$ only if $NZCount(\mathbf{v}^-_a) > \beta$.

The first step of the clustering method is the calculation of the distance matrix **M**, which contains distances between word vectors. We calculate the distances as a cosine similarity between a pair of vectors, therefore each element $\mathbf{m}_{ij}$ of **M** is defined as $\mathbf{m}_{ij} = cos(\mathbf{v}^+_i, \mathbf{v}^+_j)$ or $\mathbf{m}_{ij} = cos(\mathbf{v}^-_i, \mathbf{v}^-_j)$, $i \neq j$.

Our method considers the first n maximal distances or all the distances that are bigger than threshold γ. In order to measure distances between clusters the method uses group average agglomerative clustering approach (Manning 2008). Thus the clustering step is performed as follows. Firstly every word is considered as a separate cluster. At every step the closest pair of clusters is joined into one cluster. The distance from other clusters is calculated as an average distance from the words of joined cluster to the centroid of all other clusters. The algorithm stops when the distance between joined clusters is less than a pre-defined value (note that the lower cosine value means the bigger distance).

The number of clusters depends on maximal allowed distance between the joined objects. In case of a too small distance there will be too many small clusters and some of them should be joined as well. In case of a too big distance the clusters are too big and inconsistent. This problem can be tackled by performing a two-stage clustering: use flat clustering to divide words into clusters and then combine them using the hierarchical classification. However, if the class structure is represented as a binary tree there is a problem of clusters bounds detection. Therefore the proposed clustering method is very sensitive to the selected threshold values.

4. EXPERIMENTAL SETUP AND EVALUATION METHOD USED

For our experiments we used an untagged Russian text corpus contained 7.2 billion words. It consists of fiction books (~6.6 billion words), news articles (~550 million words), and scientific papers (articles, thesis, reports – ~50 million words). Syntactically connected words combinations were collected using a set of simple parsers described above. Each of them was developed using a finite automaton; each of them collects its own kind of syntactic dependencies. For our experiments we used the "Crosslator" tagger (Crosslator, 2013).

The database for Russian language contains more than 23 million different verb phrases (verb + noun or verb + preposition + noun) and about 5.5 million different noun phrases (noun + adjective) Different dependent words connected with one governing word are counted separately here.

To evaluate the flat clustering method following Sun and Korhonen (2011) we used normalized mutual information NMI(A, B) and F$_1$-score defined as follows (Manning *et al.*, 2008):

$$NMI(A,B) = \frac{I(A,B)}{[H(A)-H(B)]/2}\,;I(A,B) = \sum\nolimits_k \sum\nolimits_j \frac{|v_k \cap c_j|}{N} \log \frac{N|v_k \cap c_j|}{|v_k|\cdot|c_j|},$$

where $|v_k \cap c_j|$ is the number of common elements in the cluster v_k and gold-standard class c_j. *H(A)* is the entropy of clusters and *H(B)* is the entropy of a standard. F$_1$-score here is the doubled harmonic mean of precision and recall. We use information entropy defined by Shannon: $H(x) = -\sum\nolimits_{i=1}^{n} p(i)\cdot \log_2 p(i)$, where $p(i)$ – probability to find the object in the cluster: $p(i) = n / N$, n – number of objects in cluster, N – overall amount of objects.

We used the dictionary (Shvedova, 1998) as the standard. We took 97 verbs of movement, obtaining and hiring. We also add 49 extra verbs to add some noise to the test data. Verbs were divided into 25 classes. 14 of them were taken from Shvedova; 9 were naturally classified by such parameters as transitive and non-transitive or perfect and imperfect forms.

5. EXPERIMENTAL RESULTS FOR THE RUSSIAN LANGUAGE

For our experiments with the Russian corpus we used the following thresholds: co-occurrences frequency (α) not less than 5, headwords have at least 7 dependent words (β), algorithm stops when the distance is less than 0.3. In order to make a comparison with an existing ontology we select a small part of collected data.

As it was mentioned above we used 146 verbs taken from the standard or naturally classified. The obtained value of NMI was equal to 0.457, F$_1$-score was equal to 0.607.

In our experiments we were not able to evaluate all the results collected from the whole corpus using the selected standard. The amount of acquired co-occurrences was too big and the obtained lexis was too different from the standard (e.g. because of specific lexis or different logic of clustering). Since we failed to find free ontology for the Russian language that was big enough (about 100 000 words) for such comparison, all the evaluations were performed by an assessor.

Instead of cutting the list by words similarity threshold we limited the list by 7000 pairs that have maximal similarity. The direct dictionary (containing head words for syntactically connected co-occurrences) consists of 4200 different verbs. These verbs were clustered into 1550 clusters. Note that all words in this experiment were clustered so the recall was equal to 1. Precision in this experiment was equal to 0.79, so F$_1$-score is equal to 0.88.

Table 2. Comparison of the experimental results: Recall, Precision, and F$_1$-measure.

Tuple	Recall	Precision	F$_1$-measure
<verb + preposition, noun>	1	0,79	0,88
<noun + adjective>	0,85	0,88	0,85
<noun + preposition, verb>	1	0,85	0,92
<adjective + noun>	0,85	0,96	0,9

We also conducted experiments with clustering of nouns and adjectives. In case of nouns we have had possibility to use both the direct (noun + adjective) and the reverse (verb + preposition + noun) dictionaries as well. All clusters were also checked by the assessor. For the reverse dictionary recall was equal to 1, precision was equal to 0.85, F$_1$-score reached 0.92. In case of using of the direct dictionary we obtain 0.88 precision but 0.85 recall, so F$_1$-score is equal to 0.85. Experiments with adjectives by reverse dictionary shows 0.85 recall, 0.96 precision and 0.9 F$_1$-score. All results are presented in Table 2.

Table 3. Comparison of the experimental results: NMI and F_1-score.

Article	NMI	F_1-score
Sun and Korhonen (2011)	0.378 – 0.573	0.367 – 0.4
Schulte im Walde (2000)	–	0.78
Snider and Diab (2006)	–	0.46
Sass (2007)	–	0.36
This study	0.457	0.607

Results obtained by the proposed method correspond to those obtained in other projects. The Table 3 compares these results (numbers in italics mean that F_1-value was not contained in the original article and was calculated by us using the information from the article). Sun and Korhonen (2011) achieved NMI between 0.378 and 0.573 and F_1-score between 0.367 and 0.4. Schulte im Walde (2000) do not provide F_1-score, but according to the reported precision and recall the F_1-score is approximately equal to 0.78. Snider and Diab (2006) achieved F_1-score as big as 0.46 for Arabic. In Sass (2007), where the F_1-score is neither computed, it is around 0.36. For all considered languages the size of generated clusters is fairly small (2-4 words). The NMI score calculated in this article is equal to 0.457 and F_1-score is 0.607. Note that we consider the experiment with a standard. The larger value of F_1-score can be explained by larger corpus used in experiments. In this paper we do not examine differences between languages and used methods. Note that different researchers use different word similarity measures from classical cosine similarity to slightly complicated LSA.

6. DISCUSSION AND FUTURE WORK

The proposed method has a serious disadvantage: it groups words into a lot of small clusters. However, all the methods described by other authors have the same weak point: the average number of words in cluster is usually 3. Therefore the achieved clusters are too small to use them without any post-editing because there are a lot of clusters that have to be joined. The increase of the size and the quality of the input combinations can improve the result for some languages, but these improvements will be insignificant. Nonetheless, the current method can be useful in some other applications, for example, in automating manual thesauri development. According to our experiments a linguist creating a thesaurus works up to 2-3 times faster if provided with such initial clusters.

We have come to the conclusion that the area of research has some fundamental difficulties that stem from certain features of natural languages:

1. There are some pairs of words with close or same meaning and different selection preferences, i.e. they are unlikely to share any significant number of collocations. Such words will not be recognized as similar by the proposed method. The illustration of this phenomenon is synonyms that pertain to different styles or genres. For example, the words "healer" and "doctor" have the same meaning, but the former is more likely to be met in science fiction or historical texts rather than in news texts, whereas the latter is neutral. The set of adjectives that modify the word "healer" is smaller than that of the word "doctor", which results in small similarity rate.
2. Conversely, words that are not related semantically can have similar sets of co-occurrences and sub-categorization frames. For example, words "newspaper" and "police" share a lot of co-occurrences although they are not semantically close. We have found around 300 adjectives that can modify both of these words in the Russian language. Most of them are related to geographical areas (i.e. names of a city or a country), government or social institutions, emotional expressions. English has quite similar pattern.
3. The vast majority of languages have polysemic words. These are words which have several different meanings and hence several sub-categorization frames. The list of selection preferences of a given word can be divided into sub-clusters depending on the considered meaning of the word. Thus the overall similarity measure for two polysemic words sharing only one sub-cluster is very low. On the other hand, two words sharing many senses is evaluated as similar and will be clustered at one of the first steps of the clustering algorithm. This property of polysemic words is used to extract and cluster their meanings (Yarowsky, 1995; Toldova, 2008).

There are some possibilities to improve the quality of clustering. First of all, we can use more data in order to increase the amount of the processed lexis. We can also employ the fuzzy clustering that can place one word in several clusters. It should enlarge the size of clusters and help to distinguish different words' meanings.

Synthetic languages like Russian can also benefit from the consideration of grammatical categories. Our present model does not consider grammatical categories of words. However, they are usually the only marker of word's syntactical role of dependent words. If both head and dependent words are considered only in their initial form, two words with different sub-categorization frames can be erroneously joined into a one cluster. Therefore we plan to use this information in our future experiments.

However, the overall conclusion is that the result can be improved dramatically only via a completely new approach to word clustering.

7. ACKNOWLEDGEMENT

This project is partially funded by an RFH grant, No. 12-04-00060, and an RFBR grant, No. 11-01-00793.

REFERENCES

Baroni M. and Lenci A. (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics,* vol. 36 (4), pp. 673-721.

Bohnet B., Klatt S. and Wanner L. (2002). A Bootstrapping approach to automatic annotation of functional information to adjectives withan application to German. *In Proceedings of the Workshop on Linguistic Knowledge Acquisition and Representation at the 3rd LREC Conference.*

Brill E. (1995) Unsupervised Learning Of Disambiguation Rules For Part Of Speech Tagging. *On Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, Massachusetts, USA.

Boleda Torrent G. and Alonso i Alemany L. (2003). Clustering Adjectives for Class Acquisition. *In proceedings of the tenth conference on European chapter of the Association for Computational Linguistics.*

Boleda Torrent G., Badia T., Schulte im Walde S. (2005). Morphology vs. Syntax in Adjective Class Acquisition *In proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pp. 77–86.

"Crosslator" POS-tagger (2013). http://clschool.miem.edu.ru/ link "Материалы школы" (in Russian).

Fellbaum C. (1998). *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Harris, Z. (1954). "Distributional structure". *Word* 10 (23), pp. 146–162.

Hindle D. (1990). Noun classification from predi-cate-argument structures. *In Proceedings of ACL-90.*

Kipper K., Korhonen A., Ryant N., and Palmer M. (2006). Extending VerbNet with Novel Classes. *In Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa. Italy.*

Klyshinskii E., Kochetkova N., Litvinov M., and MaksimovV. (2011). A Method for Disambiguation of Part-of-Speech Homonymy Based on Application of Syntactic Compatibility in the Russian Language // *Automatic Documentation and Mathematical Linguistics*, Vol. 45, No. 1. pp. 15-19.

Klyshinsky E., Kochetkova N. (2012). Method of verbal government automatic generation (for Russian). *In Proceedings of the National conference on Artificial Intelligence 2012.* (In Russian)

Korhonen A., Krymolowski Yu., Marx Z. (2003). Clustering Polysemic Subcategorization Frame Distributions Semantically. *In Proceedings of the 41st annual meeting of the Association for Computational Linguistics.*

Korhonen A., Krymolowski Yu., Briscoe T. (2006). A Large Subcategorization Lexicon for Natural Language Processing Applications. *In Proceedings of the 5th international conference on Language Resources and Evaluation.*

Levin B. (1993). *English verb classes and alternations: A preliminary investigation.* Chicago, IL.

Li H. and Abe N. (1998). Word Clustering and Disambiguation Based on Co-occurrence Data. *In Proceedings of COLING-ACL'98*. pp. 749–755

Manning C. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. *In proceedings of the 31st annual meeting on Association for Computational Linguistics.*

Manning C., Raghavan P., and Schtze H. (2008). *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA.

Mitrofanova O. (2009) Automatic Word Clustering in Studying Semantic Structure of Texts. *Advances in Computational Linguistics. Research in Computing Science 41*, 2009, pp. 27-34

Pereira F., Tishby N., Lee N. (1993). Distributional Clustering of English Words. *In Proceedings of ACL-93.*

Preiss J., Briscoe T., and Korhonen A. (2007). A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. *In Proceedings of ACL*, pp. 912–919.

Sass B. (2007). First Attempt to Automatically Generate Hungarian Semantic Verb Classes. *In Proceedings of the 4th Corpus Linguistics conference*, Birmingham.

Schulte im Walde S. (2000). Clustering Verbs Semantically According to their Alternation Behaviour. *In Proceedings of the 18th International Conference on Computational Linguistics.* Saarbrücken, Germany.

Schütze H. (1992) Dimensions of meaning. *In Supercomputing*, pp. 787–796, Minneapolis

Schmid H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK

Shvedova N.Yu. *et al.* (1998). *Russian Semantic Dictionary.* V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences. Moscow, Russia (in Russian).

Snider N. and Diab M. (2006) Unsupervised Induction of Modern Standard Arabic Verb Classes Using Syntactic Frames and LSA. *In Proceedings of ACL*, pp. 795–802.

Sokal R.R. and Rohlf F.J. (1962). The comparison of dendrograms by objective methods. Taxon

Sun L. and Korhonen A. (2011). Hierarchical Verb Clustering Using Graph Factorization. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Edinburgh, UK.

Toldova S.Yu., Kustova G.I., and Lashevskaja O.N. (2008) Semantic filters for word sense disambiguation in the Russian National Corpus: verbs. *In Proceedings of International Workshop Dialogue'2008.* Vol. 7 (14). Moscow, pp. 522-529.

Yang M.-S. (1993) A Survey of Fuzzy Clustering. In *A survey of fuzzy clustering, Mathematical and Computer Modelling* 18(11), pp. 1-16.

Volkova G. (2013). Creating "ontology of everything". Problems of classification and their solution. *In Proceedings of the "New Information Technologies in Automated Systems Development"*, pp. 293-300. Moscow, Russia (In Russian).

Yarowsky, D. (1995) Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. Cambridge, MA, pp. 189–196.

Zelenkov J., Segalovich I., Titov V. (2005) Probabilistic model for morphological disambiguation based on normalising substitutions and adjacent words positions. *In Proceedings of the International Conference Dialog'2005*. Zvenigorod (In Russian).

Научное издание

Инновации в информационных и коммуникационных науках и технологиях

2–5 сентября 2013 г., Томск, Россия

Материалы конференции

Томск, Томский государственный университет систем управления
и радиоэлектроники, 2013

На английском языке

Scientific publication

Innovations in Information and Communication Sciences and Technologies

September 2-5, 2013, Tomsk, Russia

Proceedings

Tomsk, Tomsk State University of Control Systems
and Radio Electronics, 2013

In English

Passed for printing 16.08.2013. Format 60x84/8. Conventional printed sheet 34,64.
Pressrun 120 copies. Order 782.

Tomsk State University of Control Systems and Radio Electronics
40 Lenina Prospect, Tomsk, Russia 634050
+7.3822.510804 Phone
+7.3822.701500 Fax
www.tusur.ru