Section II.
Main program of the Conference

АВТОМАТИЧЕСКОЕ ОБНАРУЖЕНИЕ КВАЗИСИНОНИМОВ В НОВОСТНЫХ КЛАСТЕРАХ

А. Алексеев (a.a.alekseevv@gmail.com)

Н. Лукашевич (louk_nat@mail.ru)

Московский Государственный Университет, Москва, Россия

В данной работе рассматривается метод извлечения квазисинонимов — вариантов наименования одной и той же сущности в новостном кластере. Метод основан на тематической структуре новостного кластера и использует как сравнение разного рода контекстов употребления выражений, так и сопоставление употребления выражений в одних и тех же и соседних предложениях.

Ключевые слова: квазисинонимы, кластеры, новостные кластеры, контекст, метод, метод обнаружения, обнаружение.

AUTOMATIC DETECTION OF NEAR-SYNONYMS IN NEWS CLUSTERS

A. Alekseev (a.a.alekseevv@gmail.com)

N. Loukachevitch (louk_nat@mail.ru)

Lomonosov Moscow State University, Moscow,
Russian Federation

The paper presents a method for extraction of alternative names of a concept or a named entity mentioned in a news cluster. The method is based on the structural organization of news clusters and exploits comparison of various contexts of words. Word contexts are used as basis for multiword expression extraction and detection of alternative names. As a result of cluster processing we obtain groups of near-synonyms, in which the central synonym of each group is determined.

Key words: near-synonims, clusters, news clusters, context, method, detection method, detection

1. Introduction

An important step in news processing is thematic clustering of news articles describing the same event. Such news clusters are the basic units of information presentation in news services.

After a news cluster is formed, it undergoes various kinds of automatic processing:

- — Duplicates are removed from the cluster. Duplicate is a message that almost completely repeats the content of an initial document,
- — A cluster is categorized to a thematic category,
- — A summary of a cluster is created, usually containing the sentences from different documents of the cluster (multi-document summary) etc.

The formation of a cluster can represent a serious problem. It is especially difficult to form clusters correctly for complex hierarchical events having some duration in time and distributed geographic location (world championships, elections) (Dobrov, Pavlov, 2010).

A part of news cluster forming and processing problems is due to the fact that in cluster documents, the same concepts or entities may be named differently. Lexical chain approaches could partly overcome this problem using thesaurus information (Li et. al., 2007; Loukachevitch, Dobrov, 2009). However in a pre-created resource, it is impossible to fix all possible alternatives for entities naming in various clusters. For example, the U.S. air base in Kyrgyzstan may be called in documents of the same news cluster as *Manas base, Manas airbase, Manas, base at Manas International Airport, U.S. base, U.S. air base* and etc.

The problem of alternative names for named entities is partly solved by coreference resolution techniques (*Russian President Dmitry Medvedev, President Medvedev, Dmitry Medvedev*) (Ermakov, 2007; Ng, 2005), but the variability of entity names in news clusters refers not only to concrete entities but also to concepts.

In this paper we consider a method for extraction of alternative names of a concept or a named entity mentioned in a news cluster. The method is based on the structural organization of news clusters and exploits comparison of various contexts of words. The word contexts are used as basis for multiword expression extraction and alternative names detection. At the end of cluster processing we obtain groups of near-synonyms, in which the main synonym of a group is determined. Such synonym groups include both single words and multiword expressions.

2. Principles of cluster processing

Processing of cluster texts is based on the structure of coherent texts, which have such properties as the topical structure and cohesion.

Van Dijk (Van Dijk, 1985) describes the topical structure of a text, the macrostructure, as a hierarchical structure in a sense that the theme of a whole text can be identified and summed up to a single proposition. The theme of the whole text can usually be described in terms of less general themes which in turn can be characterized in terms of even more specific themes. Every sentence of a text corresponds to a subtheme of the text.

The macrostructure of a connected text defines its global coherence: “Without such a global coherence, there would be no overall control upon the local connections and continuations” (Van Dijk, 1985). Sentences must be connected appropriately according to the given local coherence criteria, but the sequence would go simply astray without some constraint on what it should be about globally.

Cohesion, that is surface connectivity between text sentences, is often expressed through anaphoric references (i. e. pronouns) or by means of lexical or semantic repetitions. Lexical cohesion is modeled on the basis of lexical chains (Hirst, St-Onge, 1998).

The proposition of the main theme, that is interaction between theme participants, should be represented in specific text sentences, which should refine and elaborate the main theme. This means that if a text is devoted to description of relations between thematic elements $C_1 \ldots C_n$, then references to these participants should be met in different roles to the same verb in text sentences.

Thus if even very semantically close entities C_1 and C_2 often co-occur in the same sentences of a text, it means that the text is devoted to consideration of relations between these entities and they represent different elements of the text theme (Hasan, 1984; Loukachevitch, 2009). At the same time, if two lexical expressions C_1 and C_2 are rarely met in the same sentences but occur very frequently in neighbor sentences then we can suppose that they are elements of lexical cohesion, and there is a semantic relation between them.

A news cluster is not a coherent text but cluster documents are devoted to the same theme. Therefore statistical features of the topical structure are considerably enhanced in a thematic cluster, and on such a basis we try to extract unknown information from a cluster.

3. Stages of cluster processing

Cluster processing consists of three main stages. At the first stage noun and adjective contexts are accumulated. The second stage is devoted to multiword expression recognition. At the third stage the search of near-synonyms is performed.

In next sections we consider processing stages in more detail. As an example we use the news cluster, which is devoted to Kyrgyzstan and the United States agreement denunciation over U.S. air base located at the Manas International Airport (19.02.2009). This news cluster contains 195 news documents and is assembled on the basis of the algorithm described in (Dobrov, Pavlov, 2010).

3.1. Extraction of word contexts

Sentences are divided into segments between punctuation marks. Contexts of word W include nouns and adjectives situated in the same sentence segments as W. The following types of contexts are extracted:

— Neighboring words: neighboring adjectives or nouns situated directly to the right or left from W (*Near*),

— Across verb words: adjectives and nouns occurring in sentence segments with a verb, and the verb is located between W and these adjectives or nouns (*AcrossVerb*),

— Not near words: adjectives and nouns that are not separated with a verb from W and are not direct neighbors to W (*NotNear*).

In addition, adjective and noun words that occur in neighboring sentences are memorized (Ns). For this context extraction only sentence fragments from the beginning up to a segment with a verb are taken into consideration. It allows us to extract the most significant words from neighboring sentences.

To illustrate how these contexts can help in extraction of near-synonyms we ran the following experiment.

Documents of the example cluster were matched with RuThes thesaurus entries (Loukachevitch, 2011); pairs of synonyms and directly related expressions were extracted (*USA — American, Kyrgyzstan — Kyrgyz Republic, base — airbase* etc.). We took pairs of such expressions with the frequencies more than half of the number of documents in the cluster (98). Then we calculated co-occurrence of the expressions in the same sentences *(Near+NotNear+AcrossVerb)* and in neighbor sentences *(Ns).* For thesaurus-related expressions the ratio between the values was:

(1) (Near+NotNear+AcrossVerb) / Ns= 0,56

If to take all other (not-related) pairs of thesaurus expressions found in the example cluster (with the same restriction on frequencies) and to calculate the same values and the ratio between them then we obtain ***2.09***. This confirms our idea that near-synonyms tend to occur more often in neighbor sentences than in the same sentences of a document.

3.2. Extraction of multiword expressions

We consider recognition of multiword expressions as a necessary step before near-synonym extraction. An important basis for multiword expression recognition is the frequency of word sequences (Witten et. al., 1999). However, a news cluster is a structure where various word sequences are repeated a lot of times. We supposed that the main criterion for multiword expression extraction from clusters is the significant excess in co-occurrence frequency of neighbor words in comparison with their separate occurrence frequency in segments of sentences (see (2), cf. Dobrov et. al., 2003):

(2) Near > 2 * (AcrossVerb + NotNear)

In addition, the restrictions on frequencies of potential component words are imposed.

Search for candidate pairs is performed in order of the value “*Near — (AcrossVerb + NotNear)*“ reducing. In case that a suitable pair has been found, its component words are joined together into a single object and all contextual relationships are recalculated. The procedure starts again and repeats until at least one join is performed.

As a result, such expressions as *Parliament of Kyrgyzstan, the U.S. military, denunciation of agreement with the U.S., Kyrgyz President Kurmanbek Bakiyev* are extracted from the example cluster.

3.3. Detection of near-synonyms

At the third stage, search for near-synonyms is produced. For assuming a semantic relationship between expressions U_1 and U_2, the following factors are used:

— U_1 and U_2 have formal resemblance (for example, words with the same beginning),

— U_1 and U_2 occur more often in neighboring sentences than within segments of the same sentence,

— U_1 and U_2 have similar contexts based on Near, AcrossVerb, NotNear and Ns features, which are determined by calculating scalar products of corresponding vectors (NearScalProd, AVerbScalProd, NotNearScalProd, NsentScalProd),

— U_1 and U_2 should be enough frequent in a cluster to be evident statistically.

Note that if comparison of word contexts is a well known procedure for synonym detection and taxonomy construction (Yang, Callan, 2009), but generation of contexts from neighboring sentences has not been described in the literature.

Near-synonyms detection consists of several steps. A different set of criteria is applied at each step. The lookup is performed in order of frequency decreasing: for every expression U_1 all expressions U_2 having a lower frequency than U_1, are considered. If all conditions are satisfied, then less frequent expression U_2 is postulated as a synonym of U_1 expression, all U_2 contexts are transferred to U_1 contexts, the expressions U_1 and U_2 become joined together. As a result the sets of near-synonyms (synonym groups) are produced, i. e. linguistic expressions that are equivalent with respect to the content of the cluster.

We assume that U_1 and U_2 expressions, when they are enclosed in such a synonym group, are closely related in sense, or their referents in current cluster are closely related to each other, so that U_2 does not represent separate thematic significance with respect to U_1. For example, such words as *parliament* and *parliamentarian* have a close semantic relationship between them in general context, but they are not synonyms. But within a particular cluster, e. g., in which decision-making process in a parliament is discussed, these words may be classified as near-synonyms.

At the first step (3.1) semantic similarity between expressions consisting of similar words is sought, e. g. *Kyrgyzstan — Kyrgyz, Parliament of Kyrgyzstan — Kyrgyz Parliament*. We used simple similarity measure — the same beginning of words.

To connect words with the same beginning in synonym groups, the following conditions are required: the co-occurrence frequency in neighboring sentences is significantly higher than co-occurrence frequency in the same sentences (3, 4) (see section 3.1); both expressions should have sufficient frequencies in the cluster. The procedure is iterative:

(3) $\mathrm{Ns} > 2 * (\mathrm{AcrossVerb} + \mathrm{Near} + \mathrm{NotNear})$

(4) $\mathrm{Ns} > 1$

If expressions are rarely located in neighboring sentences (Ns <2), then the scalar product similarity of contexts is required:

(5) $\mathrm{NearScalProd} + \mathrm{NotNearScalProd} + \mathrm{AVerbScalProd} + \mathrm{NSentScalProd} > 0.4$

At the second step (3.2) semantic similarity between expressions, one of which is included into another, is sought, for instance, *Parliament — Parliament of Kyrgyzstan, airbase — Manas airbase*. The meaning of this step lies in the fact that a cluster might not mention any other parliaments, except of the Kyrgyz Parliament, i. e. in both cases the same object is mentioned. Similarity of neighbor contexts is required here:

(6) $\mathrm{NearScalProd} > 0.1$

At the third step (3.3) we are looking for semantic similarity between the expressions with equal length and including at least one the same word, for example, *Manas Base — Manas Airbase, the U.S. military — the U.S. side* (7). High frequency of co-occurrence in neighboring sentences is required (8, 9):

(7) $\mathrm{NS} > 2 * (\mathrm{AcrossVerb} + \mathrm{Near} + \mathrm{NotNear})$

(8) $\mathrm{NS} > 1$

Finally, at the last step (3.4) semantic similarity between arbitrary linguistic expressions, mentioned in cluster documents, is searched, e. g. *USA — American, Kyrgyzstan — Bishkek*.

An assumption on semantic similarity between arbitrary expressions requires the maximum number of conditions: high frequency of co-occurrence in neighboring sentences (9, 10); restrictions on occurrence frequencies of candidates, context similarity:

(9) $\mathrm{NS} > 2 * (\mathrm{AcrossVerb} + \mathrm{Near} + \mathrm{NotNear})$

(10) $\mathrm{NS} > 0.1 * \mathrm{MaxAcrossVerb}$

The following synonym groups were automatically assembled for the example cluster as a result of described stages (the main synonym of a group, which was automatically determined, is highlighted with bold font):

— ***Manas base:*** *base, Manas Air Base, Air Base, Manas;*
— ***USA:*** *American, America;*
— ***Kyrgyzstan:*** *Kirghizia, Kyrgyz, Kyrgyz-American, Bishkek;*
— ***Parliament of Kyrgyzstan:*** *Kyrgyz parliament, parliament, parliamentary, parliamentarian;*
— ***Manas International Airport:*** *airport, Manas airport;*
— ***Bill:*** *law, legislation, legislative, legal* and etc.

4. Evaluation of method

To test the introduced method we took 10 news clusters on various topics with more than 40 documents in each cluster.

Two measures of quality were tested for multiword expression extraction. Firstly, we evaluated the percentage of syntactically correct groups among all extracted expressions. Secondly, we have attracted a professional linguist and asked her to select the most significant multiword expressions (5–10) for each cluster, and to arrange them in descending order of importance.

So for the example cluster, the following expressions were considered significant by the linguist:

— *Manas Airbase,*
— *Parliament of Kyrgyzstan,*
— *Manas base,*
— *Kyrgyz Parliament,*
— *Denunciation of agreement,*
— *Government's decision.*

Note that such an evaluation task differs from evaluation of automatic keyword extraction from texts (Su Nam Kim et. al., 2010), when experts are asked to identify the most important thematic words and phrases of a text. In our case we tested exactly multiword expression extraction. In addition, a list created by the linguist could contain repetitions (*Parliament of Kyrgyzstan — Kyrgyz Parliament*).

364 multiword expressions were automatically extracted from test clusters, 312 (87.9 %) of which were correct syntactic groups. With account of phrase frequencies, correct syntactic expressions achieved 91.4 % precision. The linguist chose 70 most important multiword expressions for clusters and 72.6 % of them were automatically extracted by the system.

We tested extracted synonym groups evaluating semantic relatedness of every synonym in a group to its main synonym. Every occurrence of supposed synonyms was tested. If more than a half of all occurrences of such a synonym in a cluster were related to the main synonym in the group, the synonymic relation was considered as correct.

Table 1 contains information about the quality of generated synonym groups calculated in number of expressions and in their frequencies.

Table 1. Test results for automatic detection of synonym groups in news clusters

Step	Number of joins	Total join frequency	Percent of correct joins	Percent of correct joins by frequency
3.1. The same beginning expressions joining	155	4 383	87.9 %	91.4 %
3.2. Embedded expressions joining	99	9 131	91.4%	92.9 %
3.3. Intersecting expressions joining	8	677	85.7 %	80.8 %
3.4. Arbitrary expressions joining	38	4 822	62.5 %	62.4 %

To assess the contribution of co-occurrence in neighboring sentences, we conducted detailed testing of the same beginning expression joining (step 3.1) for the example cluster (Table 2). Table 2 shows that adding Ns factor, as it is done in step 3.1, improves precision and recall of near-synonym recognition.

Table 2. Test results for the different methods of the same beginning synonym joining

Method	Number of joined expressions	Total joining frequency	Correct joining frequency	Precision by frequency (%)	Recall byfrequency (%)
Expressions with the same beginning (BasicLine)	383	2 266	1 472	65 %	100 %
Expressions with the same beginning + scalar products (threshold 0.1)	38	996	834	83.7 %	56.7 %
Expressions with the same beginning + scalar products (threshold 0.4)	36	976	814	83.4 %	55.3 %
Step 3.1 conditions	36	965	873	**90.5 %**	**59.3 %**

Conclusion

In this paper we have described two experiments on news clusters: multiword expression extraction and near-synonyms detection. In addition to known methods of contexts comparison, we exploited co-occurrence frequency in neighboring sentences for synonym detection. We conducted the testing procedure for the introduced method.

In future we are going to use extracted near-synonyms in such operations as cluster boundaries correction, automatic summarization, novelty detection, formation of subclusters and etc. We also intend to study methods of combination automatically extracted near-synonyms and thesaurus relations.

References

1. *Dii k van T.* 1985. Semantic Discourse Analysis. Handbook of Discourse Analysis : 103–136.
2. *Dobrov B., Loukachevitch N., Syromyatnikov S.* 2003. Automatic Detection of Text Entries for Information Retrieval Thesaurus. Proceedings of the fifth Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies” : 201–210.
3. *Dobrov B., Pavlov A.* 2010. Basic Line for News Clusterization Methods Evaluation. Proceedings of the fifth Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies”.
4. *Dobrov B., Loukachevitch N., Shternov S.* 2005. News Processing Based on Large Linguistic Resource. Internet Mathematics, available at: http://download.yandex.ru/company/grant/2005/10_Loukachevitch_103030.pdf
5. *Dobrov B., Loukachevitch N.* 2009. Summarization of News Clusters Based on Thematic Representation. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2009” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2009”) : 299–305.
6. *Ermakov A.* 2007. Automatical Extraction of Facts from Texts of Personal Files: Experience in Anaphora Resolution. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2007” (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference “Dialog 2007”).
7. *Hasan R.* 1984. Coherence and Cohesive Harmony. Understanding Reading Comprehension :181–219.
8. *Hirst G., St-Onge D.* 1998. Lexical Chains as Representation of Context for the Detection and Correction Malapropisms. WordNet: An Electronic Lexical Database and Some of its Applications.
9. *Li J., Sun L., Kit C., Webster J.* 2007. A Query-Focused Multi-Document Summarizer Based on Lexical Chains. Proc. of the Document Understanding Conference DUC-2007.
10. *Loukachevitch N.* 2009. Multigraph Representation for Lexical Chaining. Proc. of SENSE workshop :67–76.

11. *Loukachevitch N.* 2011. Thesauri for Information Retrieval Tasks.
12. *Ng V.* 2005. Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. Proc. of ACL-2005.
13. *Su Nam Kim, Medelyan O., Min-Yen Kan, Baldwin T.* 2010. Automatic Keyphrase Extraction from Scientific Articles. Proc. of the 5-th International Workshop on Semantic Evaluation, ACL -2010: 21–26.
14. *Witten I., Paynter G., Frank E., Gutwin C., Newill-Manning C.* 1999. KEA.: Practical Automatic Keyphrase Extraction. Proc. of the fourth ACM Conference on Digital Libraries.
15. *Yang H., Callan J.* 2009. A Metric-Based Framefork for Automatic Taxonomy Induction. Proc. of ACL-2009.