# Automatic Entity Detection Based on News Cluster Structure

Aleksey Alekseev, Natalia Loukachevitch

Research Computing Center of
Lomonosov Moscow State University, Russia

a.a.alekseevv@gmail.com, louk_nat@mail.ru

**Abstract.** 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 main entity detection. At the end of cluster processing we obtain groups of near-synonyms, in which the main synonym of a group is determined.

**Keywords.** Entity Detection, Lexical Cohesion, News Clusters.

## 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) [1], [2].

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 [3], [4]. However in a pre-created resource, it is impossible to fix all variants for entity 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*) [5], [6]. In Entity Detection and Tracking Evaluations, mainly such entities as organizations, persons and locations are detected and provided with coreferential relations [7]. But main entities of a cluster can be events such as *air base closure* and *air base withdrawal.* Besides, the variability of entity names in news clusters refers not only to concrete entities but also to concepts, which can also be main discussed entities such as ecology or economic problems.

News clusters as sources of various paraphrases are studied in several works. In [8] the authors describe the procedure of corpus construction for paraphrase extraction in the terrorist domain. The study in [9] is devoted to creation of a corpus of similar sentences from news clusters as a source for further paraphrase analysis. These studies are aimed to obtain general knowledge about a domain or linguistic means of paraphrasing, but it is also important to extract near-synonyms or coreferential expressions of various types from a news cluster and to use them to improve the processing of the same news cluster or a corresponding theme.

In this paper we consider a method for extraction of main entities from a news cluster including named entities, activities and concepts. 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 a basis for multiword expression extraction and main entities detection. At the end of cluster processing we obtain main entities of a news cluster and their mention expressions presented as a group of near-synonyms, in which the main synonym of a group is determined. Such synonym groups include both single words and multiword expressions. In this paper we study only simple features generated from a news cluster without attraction of additional semantic and other types of information as a basic line for future research. The experiments were carried out for Russian news flows.

## 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 [10] 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” [10]. 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 [11].

The proposition of the main theme, that is an 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 [12], [13]. 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.

To check our idea that near-synonyms can be more often met in neighbour sentences than in the same sentences we have carried out the following experiment. More than 20 large news clusters have been matched with terms of Sociopolitical thesaurus [14] and thesaurus-based potential near-synonyms have been detected. Such types of near-synonyms include (these examples are translations from Russian, in Russian the ambiguity of expressions is absent):

- nouns – thesaurus synonyms (*Kyrgyzstan – Kirghizia*),
- adjective – noun derivates **(***Kyrgyzstan – Kyrgyz*),
- hypernym and hyponym nouns(*deputy – representative*),
- hypernym–hyponym noun - adjective (*national – Russia*),
- part-whole relations between nouns (*parliament – parliamentarian*),
- part-whole relations for adjective and noun (*American – Washington*),

For each cluster we considered all these pairs of expressions with a frequency filter: the frequencies of the expressions in a cluster should be more than a quarter of the number of documents in the corresponding cluster. For these pairs we computed the ratio between their co-occurrence in the same sentence clauses $F_{segm}$ and in neighbour sentences $F_{sent}$. Table 1 shows the results of our experiment.

**Table 1.** Frequency ratio of related expressions within segments of sentences and neighbour sentences

| Type of relation | $F_{segm}/F_{sent}$ ratio | Number of pairs |
|---|---|---|
| Synonymic Nouns | 0.309 | 31 |
| Noun-adjective derivation | 0.491 | 53 |
| Hyponym – Hypernym (nouns) | 1.130 | 88 |
| Hyponym – Hypernym (noun – adjective) | 1.471 | 28 |
| Meronym- holonym (nouns) | 0.779 | 58 |
| Meronym- holonym (noun – adjectives) | 1.580 | 29 |
| Other | 1.440 | 21483 |

From the table we can see that the most closely-related expressions (synonyms, derivates) are much more frequent in neighbour sentences than in the same clauses of the same sentences. Further, the more the distance in a sense between expressions is the more the ratio $F_{segm}/F_{sent}$ is until stabilization near the value equal 1.5.

We can also see that noun-noun and noun-adjective pairs have different values of the ratio. We suppose that in many cases adjectives are elements of noun groups, which can play own roles in a news cluster. Therefore the first step in detection of main entities should be extraction of multiword expressions denoting main entities of the 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 on 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 [1].

### 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 (*Across-Verb*),
- 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 co-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.

### 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 [15]. 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 (1):

$$\text{Near} > 2 * (\text{AcrossVerb} + \text{NotNear}) \tag{1}$$

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. If 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* were 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 exploited:

- $U_1$ and $U_2$ have formal resemblance (for example, words with the same beginning),
- $U_1$ and $U_2$ co-occur more often in neighboring sentences than within segments of the same sentences; here we use results of the experiment described in section 2;
- $U_1$ 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 present main entities.

Note that if the comparison of word contexts is a well known procedure for synonym detection and taxonomy construction [16], but the 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 (2, 3) (see section 2); both expressions should have sufficient frequencies in the cluster. The procedure is iterative:

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

$$\mathrm{Ns} > 1 \qquad (3)$$

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

$$\mathrm{NearScalProd} + \mathrm{NotNearScalProd} + \mathrm{AVerbScalProd} + \mathrm{NSentScalProd} > 0.4 \qquad (4)$$

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:

$$\mathrm{NearScalProd} > 0.1 \qquad (5)$$

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*. High frequency of co-occurrence in neighboring sentences is required (6, 7):

$$NS > 2 * (AcrossVerb + Near + NotNear) \tag{6}$$

$$NS > 1 \tag{7}$$

Finally, at 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 (8, 9); restrictions on occurrence frequencies of candidates, context similarity:

$$NS > 2 * (AcrossVerb + Near + NotNear) \tag{8}$$

$$NS > 0.1 * MaxAcrossVerb \tag{9}$$

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 [17], 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 semantic 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 by 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 2 contains information about the quality of generated synonym groups calculated in number of expressions and in their frequencies.

**Table 2.** 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 | 155 | 4383 | 87.9% | 91.4% |
| 3.2. Embedded expressions | 99 | 9131 | 91.4% | 92.9% |
| 3.3. Intersecting expressions | 8 | 677 | 85.7% | 80.8% |
| 3.4. Arbitrary expressions | 38 | 4822 | 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 3). Table 3 shows that Ns factor adding, as it is done in step 3.1, improves precision and recall of near-synonym recognition. The proposed method has not the absolutely best F-measure value, but the precision less than 80% is inadmissible for the near-synonym detection task. Therefore, the BasicLine should not be considered as the best approach.

**Table 3**. Test results for different methods of detection of near-synonyms with the same beginning

| **Method** | Number of joined expressions | Total joining frequency | Correct joining frequency | Precision by frequency (%) | Recall by frequency (%) | F-measure (%) |
|---|---|---|---|---|---|---|
| Expressions with the same beginning (BasicLine) | 383 | 2266 | 1472 | 65% | 100% | 78.8% |
| Expressions with the same beginning + scalar products (threshold 0.1) | 38 | 996 | 834 | 83.7% | 56.7% | 67.6% |
| Expressions with the same beginning + scalar products (threshold 0.4) | 36 | 976 | 814 | 83.4% | 55.3% | 66.5% |
| Step 3.1 conditions | 36 | 965 | 873 | **90.5%** | **59.3%** | **71.7%** |

## 5 Conclusion

In this paper we have described two experiments on news clusters: multiword expression extraction and detection of near-synonyms presenting the same main entity of a news cluster. In addition to known methods of context comparison, we exploited co-occurrence frequency in neighboring sentences for near-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 sub-clusters and etc. We also intend to study methods of combination automatically extracted near-synonyms, methods of coreference resolution and thesaurus relations.

## 6 References

1. Dobrov, B., Pavlov, A.: Basic line for news clusterization methods evaluation. In: Proceedings of the 5-th Russian Conference RCDL-2010 (2010) (in Russian)
2. Allan, J.: Introduction to Topic Detection and Tracking. In: Topic detection and tracking, Kluwer Academic Publishers Norwell, MA, USA,. pp. 1-16 (2002)

3. Li, J., Sun, L., Kit, C., Webster, J.: A Query-Focused Multi-Document Summarizer Based on Lexical Chains. In: Proceedings of the Document Understanding Conference DUC-2007 (2007)
4. Dobrov, B., Loukachevitch, N.: Summarization of News Clusters Based on Thematic Representation. In: Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference Dialog`2009, pp. 299-305 (2009) (In Russian)
5. Duame, H., Marcu, D.: A large Scale Exploration of Global Features for a Joint Entity Detection and Tracking Model. In: Proceedings of Human Language Conference and Conference on Empirical Methods in Natural Language Processing, pp. 97-104 (2005)
6. Ng, V.: Machine learning for coreference resolution: from local classification to global ranking. In: Proceedings of ACL-2005 (2005)
7. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weishedel, R.: The Automatic Content Extraction (ACE): Task, Data, Evaluation. In: Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004 (2004)
8. Barzilay, R., Lee, L.: Learning to Paraphrase: an Unsupervised Approach Using Multiple Sequence Alignment. In: Proceedings of HLT/NACCL-2003 (2003)
9. Dolan, B., Quirk, Ch., Brockett, Ch.: Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In: Proceedings of COLING-2004 (2004)
10. Dijk, van T.: Semantic Discourse Analysis. In: Teun A. van Dijk, (Ed.), Handbook of Discourse Analysis, vol. 2., pp. 103-136, London: Academic Press (1985)
11. Hirst, G., St-Onge, D.: Lexical Chains as representation of context for the detection and correction malapropisms. In: WordNet: An electronic lexical database and some of its applications / C. Fellbaum, editor. Cambrige, MA: The MIT Press (1998)
12. Hasan, R.: Coherence and Cohesive harmony. J. Flood, Understanding reading comprehension, Newark, DE: IRA, pp. 181-219 (1984)
13. Loukachevitch, N.: Multigraph representation for lexical chaining. In: Proceedings of SENSE workshop, pp. 67-76 (2009)
14. Loukachevitch, N., Dobrov, B.: Evaluation of Thesaurus on Sociopolitical Life as Information Retrieval Tool. In: M.Gonzalez Rodriguez, C. Paz Suarez Araujo (Eds.), Proceedings of Third International Conference on Language Resources and Evaluation (LREC2002), Vol.1, pp.115-121 (2002)
15. Witten, I., Paynter, G., Frank, E., Gutwin, C., Newill-Manning, C.: KEA: practical automatic keyphrase extraction. In: Proceedings of the fourth ACM conference on Digital Libraries (1999)
16. Yang, H., Callan, J.: A metric-based framework for automatic taxonomy induction. In: Proceedings of ACL-2009 (2009)
17. Su Nam Kim, Medelyan, O., Min-Yen Kan, Baldwin, T.: Automatic Keyphrase Extraction from Scientific Articles. In: Proceedings of the 5-th International Workshop on Semantic Evaluation, ACL -2010, pp. 21-26 (2010)