Use of Multiple Features for Extracting Topics from News Clusters

© Alekseev A., Loukachevitch N.

Lomonosov Moscow State University
a.a.alekseevv@gmail.com, louk_nat@mail.ru

Abstract

In this paper we consider a method for extraction of sets of semantically similar language expressions representing different participants of the text story – thematic nodes. 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 thematic node construction. We evaluate our method on the multi-document summarization task.

1 Introduction

As it is widely known, a natural language text contains a lot of sense-related words and expressions such as synonyms, antonyms, hyponyms, hypernyms and others. The presence of such expressions in texts contradicts to the standard representation of texts as a bag of words.

The modeling of semantic relatedness between words in a text can be based on linguistic or statistical approaches. A linguistic approach considers this phenomenon as a property of natural language texts called lexical cohesion and represents it as lexical chains of semantically-related words [22]. This representation is based on existing linguistic resources such as WordNet [20] or user-generated resources as Wikipedia [26]. The evident problem of such an approach is the possible absence of necessary knowledge in a utilized resource or the irrelevance of described information to the text contents.

Well-known statistical approaches such as Latent Dirichlet Allocation are often called "topic models" (LDA) [6, 16]. Topic Models [2] are based on the idea that documents are mixtures of topics where a topic is a probability distribution over words. In such models two probability distributions are usually considered:

- Topics-VS-Documents distribution,
- Words-VS-Topics distribution.

Extraction of such topics is based on the iterative application of statistical methods (for example, Gibbs Sampling) and the co-occurrence of words in the same documents of a collection. The statistical inference of topics does not consider the existing lexical relations between words or the internal structure of texts.

In this paper we will consider a deterministic approach to extraction of thematically-related chains of words and expressions (thematic nodes) based on various factors including:

- thesaurus information,
- spelling resemblance,
- several types of context similarity,
- discourse-based similarity. To extract such a similarity we utilize some assumptions on the structure of natural language texts.

We will demonstrate our approach on the news cluster summarization task. News clusters can contain semantically-related words as within a single cluster document as in different documents of the cluster, which can lead to problems in news clusterization and further summarization of the cluster. 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.

This paper is organized as follows: after related works are surveyed in section 2, we discuss in section 3 a theoretical basis of the proposed algorithm, in particular coherent text-structure model. Detailed algorithm description is provided in section 4. In section 5, we describe the evaluation procedure and present the results. Section 6 concludes the paper.

2 Related Approaches

A context-based analysis is the most popular and widely used way to detect semantically related expressions [15]. Intuitively, words that can be used in the same context have a high chance to have similar meanings. Based on this, many methods have been proposed focusing on two aspects – what type of context to consider and what similarity measure to employ [28].

Dang et al. [9] proposes two simple methods addressing the quality of contexts for near-synonym extraction. Different types of contexts have different

Proceedings of the Spring Young Researcher's Colloquium On Database and Information Systems SYRCoDIS, Moscow, Russia, 2012

synonymy contribution. For example, consider two context words *"carries"* and *"points"* in the following sentences:

(1) *"He carries a gun in the bag"*
(2) *"He carries a pistol in the bag"*
(3) *"He points his gun at us"*
(4) *"He points his pistol at us".*

Obviously, *"points"* is a better context than *"carries"* for determining that *"gun"* and *"pistol"* should be near-synonyms. Therefore, not only a context word co-occurrence frequency with a term is important, but also how many different terms co-occur with this context word. This characteristic reflects the quality of a context word. The paper offers two formulas for the context quality estimation.

There are some more sophisticated approaches for context extraction and exploiting. An unsupervised learning algorithm for identification of paraphrases from a corpus of multiple English translations of the same source text is proposed in [4]. Part-of-speech templates of neighbouring words are considered as contexts in this work. An iterative algorithm starts from the same words and phrases extraction in multiple translations. These are "good" examples of paraphrasing. Afterwards, context templates are extracted for found coincidences. These are "good" context templates. All other context templates are considered as "bad" examples. A step of the algorithm lies in the generation of new "good" examples of paraphrasing on the basis of extracted "good" templates and so on. A set of paraphrases is produced as a result of the described procedure.

The problem of alternative names for named entities is partly solved by co-reference resolution techniques (*Russian President Dmitry Medvedev*, *President Medvedev*, *Dmitry Medvedev*) [15, 25]. In Entity Detection and Tracking Evaluations, mainly such entities as organizations, persons and locations are detected and provided with co-referential relations [13]. 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 such as *ecology* or *economic problems*.

News clusters as sources of various paraphrases are studied in several works. In [3] the authors describe the procedure of corpus construction for paraphrase extraction in the terrorist domain. The study in [14] 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 similar expressions of various types from a news cluster and to use them to improve the processing of the same news cluster.

3 Text-Structure Model

3.1 Thematic Structure and Thematic Nodes

The 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 natural language text defines its global coherence: "Without such a global coherence, there would be no overall control upon the local connections and continuations. 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" [10].

Thus, a natural language text should have the main theme. In the hierarchical thematic structure of the document the main theme should be elaborated, specified with subthemes corresponding to specific sentences. Because of the global connectivity of the thematic structure, a considerable number of subtheme participants should be related to main participants of the main theme (fig. 1). So we suppose that numerous lexical relations in a text should refer to the participants of the main theme. We call such a node of links to more important thematic element – *thematic node*.

Interactions between subtheme participants described in specific sentences should be also related to the main theme of the document. From here we conclude that if two entities C_1 and C_2 often co-occur in the same sentences of a text, it means that the text is devoted to the consideration of relations between these entities and they represent different elements of the text theme [19, 23]. At the same time, if two lexical expressions C_1 and C_2 are rarely met in the same sentences, but co-occur very frequently in neighbour sentences then we can suppose that they are elements of a lexical chain, and there exists a semantic relation between them.

So we think that an important step to reveal the thematic structure of a document is to reconstruct thematic nodes. In comparison with LDA topics, thematic nodes do not comprise words co-occurring in the same documents or the same sentences - each thematic node is supposed to collect words and expressions corresponding to a separate participant of the situation described in the text.

If to compare with standard lexical chaining techniques [20], which try to construct chains of semantically related expressions in texts, thematic node elements are supposed to be related to its main element (center of the thematic node), and if two related

expressions (for example, *doctor* and *patient*) co-occur

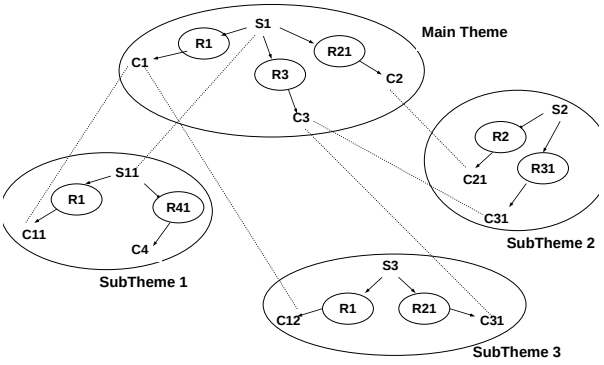

Fig.1. The hierarchy of themes in a natural language text and links between levels of the hierarchical structure where S1, S11, S2, S3 are predicates describing a situation, C1… C4 are entities participating in the described situations, Ri are roles of entities.

in the same sentences of the text, it means that their relations represent the focus of the text contents, these expressions are related to different participants of the text topics and should be assigned to different thematic nodes. And on the contrary, if two expressions rarely co-occur in the same sentences, but frequently co-occur in neighbouring sentences, then they have to be considered as the elements of the same thematic node.

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.

3.2 Hypothesis Validation

To check our idea that semantically related expressions are more often met in neighbouring 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 Socio-political thesaurus [23] and thesaurus-based 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 neighbouring sentences F_{sent}. Table 1 shows the results of our experiment.

Table 1: Frequency ratio of related expressions within segments of sentences and neighbouring 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 1 we can see that the most closely-related expressions (synonyms, derivates) are much more frequent in neighbouring sentences than in the same clauses of sentences. Further, the more the distance in a sense between expressions is the more the ratio Fs_{egm}/F_{sent} is until the stabilization near the value equal to 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 components of noun groups, which can play own roles in a news cluster. Therefore, the first step in detection of thematic nodes should be extraction of multiword expressions.

4 Thematic Nodes Construction

Thus our aim in cluster processing is to reveal the main participants of the situation described in a cluster by means of thematic nodes extraction. We believe that such information about a cluster should improve performance of such operations as cluster refining and cluster summarization. The construction of thematic nodes is based on different types of similarities between expressions. Besides, the necessary condition of inclusion of two expressions in the same thematic nodes is their high co-occurrence frequency in neighbouring sentences (see section 3) in comparison with the same sentence co-occurrence frequency.

The 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 thematic nodes are constructed.

In next sections we consider processing stages in more detail. As an example we use a news cluster 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 clusterization algorithm described in [11].

4.1 Word Contexts Extraction

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

- Neighbouring words: neighbouring 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 neighbours to W (NotNear).

In addition, adjective and noun words that co-occur in neighbouring sentences are memorized (NS). For NS context extraction, only sentence fragments from the beginning up to a segment with a verb in a personal form are taken into consideration. It allows us to extract the most significant words from neighbouring sentences.Each context type has a numeric value equal to its frequency for each candidate pair. For example, if a candidate pair of objects A and B occurred 3 times directly near in an analysed news cluster, it means, that this candidate pair would have Near value equal to 3.

Along with the described context types, we exploited classical n-gramm contexts. We call such contexts – strict contexts: two words to the left and two words to the right in the fixed order around the word W. For example, if we extract strict contexts of word *"processing"*, then in the sentence *"Cluster processing consists of three main stages"* we will yield the string context: (**, cluster, W, consist, of*), where * means a context element missing in the beginnings and endings of sentences. Thereon strict contexts for all the words are gathered and two candidate words can be compared by the number of identical strict contexts.

4.2 Extraction of Multiword Expressions

We consider recognition of multiword expressions as a necessary step before thematic nodes construction. An important basis for multiword expression recognition is the frequency of word sequences [27]. 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 a co-occurrence frequency of neighbour words in comparison with their separate occurrence frequency in segments of sentences (1):

$$Near > 2 * (AcrossVerb + NotNear) \quad (1)$$

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

The search for candidate pairs is performed in order of the *"Near – 2*(AcrossVerb + NotNear)"* value decrease. If a suitable pair has been found, its component words are joined together into a single phrase 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.

Two measures of quality were tested for multiword expression extraction in our previous work [1]. Firstly, the share of syntactically correct groups among all extracted expressions was evaluated. Secondly, a professional linguist was invited to select the most significant multiword expressions (5-10) for each cluster, and arrange them in the descending order of importance. The proposed algorithm for multiword expression extraction showed 91.4% precision and 72.6% recall result, which is enough for further thematic node construction.

4.3 Similarity Features

A set of the five main similarity features is used for determining of semantically related expressions and the following thematic nodes construction. Some of these features are based on context information, extracted directly from the news cluster under consideration. Others reflect formal resemblance information and information from pre-defined resources. Each similarity feature contributes some points to the overall score of a candidate pair. The scoring algorithm would be described in the next section in more details.

Context-dependent features:

The neighbouring sentence feature (**NSF**). This feature is based on the discourse model described in section 3 and reflects the expected co-occurrence of thematic node elements in the same and neighbouring sentences.

NSF feature is calculated on the basis of AcrossVerb, Near, NotNear and NS context features and their average distribution in the cluster. NSF feature estimates the excess of neighbouring sentence counts in comparison to across-verb, near and not-near contexts and the following value is the basis of this feature:

$$C = NS - 2 * (AcrossVerb + Near + NotNear) \quad (2)$$

The general formula for NSF feature score contribution has the next form:

$$\mathrm{NSF} = Min\left[1, \frac{C}{Avg(C)}\right] \quad (3)$$

where AVG(C) is an average value of C among positive values in the whole cluster.

NSF feature is also our regulatory feature. It means that a candidate pair could not be included in the same thematic node if NSF feature has a negative value (see section 3.2). A candidate pair with negative NSF feature value has no overall score and is not considered by the algorithm. It is worth noting that such a feature has not been utilized before for such tasks as near-synonym detection, topic extraction or lexical chaining before.

Strict context feature (**SC**). This feature is based on the comparison of fixed order contexts. The more identical templates a pair shares the more its similarity is. Contexts with missing information (or not full 4-gramm contexts) have a less weight than full contexts.

Strict contexts are scored using the following weighting formula: each word in a context n-gramm has the weight 0.25. For instance, the n-gramm (*, *, *consist, of*) would have the weight 0.5 and (*news, cluster, consist, of*) would have the weight equals 1.0, which is the maximum weight for a full context n-gramm.

SC feature has a Boolean value: 0 or 1. The maximum value of 1 is assigned if SC feature has the value not less than 2. It means, that for satisfying of SC feature a candidate pair has to share not less than 2 identical context templates (with context template weights taken into account).

Cosine similarity feature (**S**calar **P**roduct **S**imilarity, **SPS**). Each type of context information, described in 4.1, represents a vector of frequencies assigned to each word. Dimensions in this vector reflect co-occurrence frequencies of a word under consideration with all other words, mentioned in the news cluster. When the context vectors are calculated, they can be compared with classic cosine similarity metric reflecting the similarity of two expressions. SPS feature can be considered as more smoothed and flexible than SC feature, because they both were designed to analyse the sentence context.

SPS feature score has a decimal value and directly related to cosine similarity metrics of the corresponding vectors. It is calculated as the sum of cosine similarity values for all context vectors (AcrossVerb, Near and NotNear vectors) and restricted by 1 from above.

Context-independent features:

Formal resemblance feature (**B**eginning **S**imilarity, **BS**). Formal resemblance is a natural way for similarity detection. We exploited a simple formal resemblance metric – the same beginning of words. This feature allows recognition of such pairs as *Kyrgyzstan - Kyrgyz, Parliament of Kyrgyzstan - Kyrgyz Parliament* and etc.

The overall score contribution of BS feature has a Boolean value. So, this feature could add one point to the whole score of a candidate pair.

External resource similarity feature (**T**hesaurus **S**imilarity, **TS**). There are a lot of existing and pre-defined resources, which could give additional information about relations between words and phrases. Such information can be used in thematic nodes construction and make their recognition more reliable. Moreover, it is known that some types of relations between words and phrases are widely used for the text connectivity (for example, such relations as synonymy, hyponym-hypernym, meronym-holonym). To compute TS feature we utilized information from Russian language thesaurus RuThes [23]. Only directly established thesaurus relations (without any inference) were considered (synonyms, hyponyms, hypernyms, meronyms (parts), holonyms (wholes)), but what types of relations are the most appropriate for this purpose would be studied in our further work.

TS feature has also a Boolean value - 0 or 1.

The values of the similarity features for each pair of expressions are summed up to their overall similarity score. Each feature can add from 0 up to 1 point to the overall score. So, the score value is a decimal number located between 0 (minimum similarity) and 5 (maximum similarity).

Additionally, we use **SPlus** feature reflecting similarity through a third-party object or so called "plus one similarity". If an expression A is similar to an expression C and an expression B is similar to C then we postulate that SPlus condition is satisfied for A and B expressions. By "similar" in this case we understand fulfilment of the following two conditions:

- C >= 0 (from (2) formula)
- One of the features BS, TS or SC has

Pairs \ Features	Context-independent		Context-dependent				SPlus	SCORE
	BS	TS	NSF	SC1	SC2	SPS		
"Kyrgyzstan" – "Kyrgyz"	1.00	1.00	0.07	+	1.00	1.00	+	**4.07**
"Airbase" – "Manas Airbase"	1.00	0.00	1.00	+	1.00	1.00	+	**4.00**
"Kyrgyz parliament" – "Parliament of Kyrgyzstan"	1.00	0.00	0.79	+	1.00	1.00	+	**3.79**
"Manas Airbase" – "Manas base"	1.00	0.00	0.71	+	1.00	0.94	+	**3.65**
"Airbase" – "Base"	0.00	0.00	1.00	+	1.00	1.00	+	**3.50**

Table 2: Example of candidate pairs ranking (top 5 pairs at the first iteration)

TRUE value.

So, if a candidate pair has no BS and TS features satisfied, but SPlus condition has TRUE value then an additional 0.5 point is added to the overall similarity score. At last the pairs of similar expressions are ordered in the similarity score decrease, constructing the similarity ranking.

An instance of such ranking for the example cluster is provided in Table 2 (several pairs from the top of the list before the first iteration of the algorithm).

4.4 Algorithm

The algorithm constructs thematic nodes from the most similar pairs of expressions.

The supposed structure of the thematic nodes is as follows:

- a textual expression can belong to one or two thematic nodes; double links to thematic nodes provide the possibility to represent different aspects of the expression or its lexical ambiguity;
- every thematic node has its main element – the thematic center, which belongs only to one thematic node. The thematic center is the most frequent expression among thematic node elements.

The thematic node construction consists of the following steps:

- The pair of expressions with the maximal similarity score is taken;
- The most frequent element of the pair absorbs the second element with all its occurrences and contexts and becomes the representative of the pair, that is the thematic center of a new thematic node;
- The second participant of the pair can further be joined in a similar manner to another thematic node;

Merging of thematic nodes consisting from several elements is fulfilled in the same way as single expressions. The center of a more frequent thematic node becomes the center of a new, merged thematic node.

On the whole, each iteration of the algorithm consists of three main steps:

1. Candidate pairs ranking
2. Top-ranked pair selection
3. Joining procedure

The iterative process proceeds until the top-ranked pair score would be less than a pre-defined threshold. For example, a thematic node with the main expression *Manas base* is constructed as follows (top-ranked pairs at various iterations are presented; a more frequent object is the first object in a pair):

Iteration 2: ***airbase*** <- *Manas airbase*

Iteration 4: ***Manas base*** <- *Manas airbase*

Iteration 5: (***Manas base***, *Manas airbase)* <- *base*

Iteration 6: (***Manas base,*** *Manas airbase, base)* <- *Manas*

Iteration 7: (***Manas base***, *Manas airbase, base, Manas* <- ***(airbase,*** *Manas airbase)*

Iteration 41: (***Manas base***, *Manas airbase, base, Manas, airbase)* <- *base closure*

Iteration 51: (***Manas base***, *Manas airbase, base, Manas, airbase, base closure)* <- *base withdrawal*

The following thematic nodes were obtained as a result of the described algorithm for the example cluster. We present top ten the most frequent thematic nodes ordered by frequency without any correction, the thematic centers are highlighted by the bold font (translation from Russian):

Manas base: *Manas airbase, base, Manas, airbase, base closure, base withdrawal*

Kyrgyzstan: *Kyrgyz, Parliament of Kyrgyzstan, Kyrgyz parliament, Kyrgyz president Kurmanbek Bakiyev, Bishkek, President of Kyrgyzstan*

USA: *United States, American (noun), American (adj), Pentagon, American military*

Deputy: *Paliament deputy, Legislator, Parlamentarian, Parliament, Parliamentary Committee*

Soldiers: *Military contingent, troops, Military base, American military, Martial contingent, Military-transport aircraft*

Country: *Territory of country, Russia, State, Russian, Territory of the republic*

Manas airport: *Manas international airport, Manas, Airport*

Withdrawal: *Airbase withdrawal, Closure*

Decision: *Government decision*

USA agreement denunciation: *Agreement denunciation, Denunciation, Contract denunciation, Agreement termination, Contract termination*

5 Summary-Based Evaluation

The main purpose of the proposed method was to improve the overall performance of various news cluster automatic processing tasks.

We selected the multi-document summarization task as the basis of evaluation; we suppose that constructed thematic nodes can allow avoiding undesirable repetitions in summary sentences and improve the quality of generated summaries.

In general, summarization is a task of creating a brief summary of text or a set of related texts. News cluster summarization is widely used in news services, such as Google.News, Yandex.News, Rambler.News etc. These services collect information from multiple news sources, divide this information into thematic categories (news clusters), process this information and afterwards present short texts describing a specific event to end-users of the service.

A lot of summarization algorithms have been developed. Some of them are comparatively simple and based on frequency features only [7]. Others exploit additional semantic information from the pre-defined resources [12] and use more sophisticated algorithms for sentence ranking and selection [21]. And there is set of modern algorithms, which employ probabilistic

language models (such as Latent Dirichlet Allocation, LDA) for summary creating and present state-of-the-art results along with other approaches [17]. Such summarization approaches are usually based on hierarchical Latent Dirichlet Allocation model (hLDA) [5], which is built on sentence level (not on document level as in classic LDA). It allows capturing of expected topic distributions in given sentences directly from the model. It is important, because we are considering extractive summarization approaches and sentence is an atomic unit in this case. Besides, news clusters could contain a relatively small number of documents, which may limit the variability of topics if they are evaluated on the document level.

Celikyilmaz et al. [8] propose to construct a hierarchical tree structure of candidate sentences. Each sentence is represented by a path in the tree, and each path can be shared by many sentences. The assumption is that sentences sharing the same path should be more similar to each other because they share the same topics. The tree-based sentence scoring and ranking algorithm is also provided in this work.

For our evaluation, we selected one of the most well-known summarization algorithms – Maximal Marginal Relevance (MMR) [7]. We substitute initial words in the cluster sentences with corresponding thematic nodes and suppose that this generalization operation can improve generated summaries.

5.1 MMR Method

Maximal Marginal Relevance Multi-Document summarization is a classic purely extractive summarization method, which is based on Maximal Marginal Relevance concept proposed for information retrieval [7]. In the original version it is a query-oriented summarization algorithm, but there is a variant of MMR for general summarization too.

MMR criterion means that the best sentence for a summary has to be maximally similar to the user query (or the whole text in case of general summarization) and maximally different from the already selected sentences of the summary.

The summary is constructed incrementally from a list of ranked sentences; the sentence which maximizes MMR is chosen at each iteration:

$$MMR = \arg\max_{s \in S} \left[\lambda \cdot Sim_1(s, Q) - (1 - \lambda) \cdot \max_{s_j \in E} Sim_2(s, s_j) \right]$$

where S is the set of candidates sentences and E is the set of selected sentences; λ represents an interpolation coefficient between sentence relevance and non-redundancy; Sim_1 is the similarity metric used in document retrieval and relevance ranking between sentences (documents) and a query; and Sim_2 can be the same as Sim_1 or a different metric. In our work we used the classic cosine similarity metric as Sim_1 and Sim_2.

5.2 Pyramid Evaluation Method

Evaluation of automatically generated summaries is a very complicated procedure. The summary evaluation involves different aspects, the main of which are the summary content and coherence. In 2005 an algorithm for summary content evaluation – the Pyramid method was proposed [18]. The algorithm was successfully used in the large-scale evaluation of competitive summarization systems [25] and in our work we also utilized this method.

The method is based on extraction of all "information nuggets" from expert (manually created by experts) summaries, or Summary Content Units (SCUs). SCU describes some fact, which expert summaries take in. Therefore, an automatic summary has to reflect this fact too. Here is an example of summary content unit and its occurrences in different documents of news cluster from [18]:

> **SCU:** *Mini-submarine trapped underwater*
> **contr1:** *mini-submarine... became trapped... on the sea floor*
> **contr2:** *a small... submarine... snagged... at a depth of 625 feet*
> **contr3**: *mini-submarine was trapped... below the surface*
> **contr4**: *A small... submarine... was trapped on the seabed*

The number of contributors (contr) is equal to the weight of the SCU, i.e. an SCU from four contributors has a weight of 4, an SCU from 3 contributors has the weight of 3 and etc. It means that an essential step in summary evaluation by the Pyramid method is creation of several expert summaries (4 summaries at DUC/TAC conference) and manual selection of content units from them. Each SCU after this process obtains a weight, which equals the number of expert summaries, where this SCU occurred.

So, all found summary content units form a pyramid. The upper levels are usually occupied by a comparatively small amount of the most significant summary content units. A lot of less important information units are placed at the lower levels of the pyramid. SCUs pyramid construction is a preliminary step in the summary evaluation. When this step is done, each automatic summary could be assessed for the presence of SCUs from the formed pyramid and the final summary score could be calculated on the basis of the following formula:

$$Sum\_Score = \frac{[Found\_SCU\_Weight]}{[Sum\_SCU\_Weight]} \quad (2)$$

where Found_SCU_Weight – the total weight of all SCUs, presented in a given automatic summary, Sum_SCU_Weight – the total weight of all SCUs, determined for the current cluster. Accordingly, the final Pyramid score for an automatic summary is its total SCU weight divided by the maximum SCU weight available to a summary of average length (where the

average length is determined by the mean SCU count of the expert summaries for this topic). This way of assessment reflects the coverage of expert SCUs by an automatic summary with taking into account SCU different weights.

The Pyramid method for summary content evaluation allows measuring the information coverage by automatic summary, regardless the synonyms and paraphrases used in news cluster documents.

5.3 Evaluation Procedure

To evaluate our approach we apply MMR summarization method to different input data. The original version of MMR method considers an input text (or texts) as a bag-of-words. No information about multiword expressions and related expressions is exploited. Our idea was to add such information to the input data and to examine the results of MMR algorithm. Accordingly, four variants of the input data structure were investigated:

1. Simple bag-of-words model with no additional information. It is a classic input data for MMR method. This version is considered as a baseline.
2. Bag-of-words model with information about multiword expressions added. All consecutive words included to a multiword expression, are considered as a single word. This model is important for evaluation of the influence of multiword expressions on the overall performance.
3. Bag-of-words model with thesaurus information added. Thesaurus-based thematic nodes were described in [12]. Elements of the same thematic node are substituted with their thematic center and considered as the same input word.
4. Bag-of-words model with cluster-based thematic nodes information (assembled by the proposed algorithm) added. Thematic node elements are also considered as the same input word (thematic center of this thematic node) with the weight propotional to the similarity score between a given element and the thematic center.

To test the MMR method with different input data models we took 10 news clusters on various topics. The Pyramid evaluation procedure was performed. For this purpose two-four expert summaries were created for each news cluster by professional linguists. Summary Content Units were manually extracted from these summaries. On the whole, 129 SCUs were extracted. Each SCU has the weight equal to the number of expert summaries, where it occurred. Thus, SCU pyramid was assembled. Afterwards, each automatic summary was manually assessed for SCUs references and the score of the examined summary was calculated (see previous section).

5.4 Results

Table 3 shows evaluation results of MMR method with various input data.

We can see that the adding of multiword expressions to the simple bag-of-words model decreases the overall performance of the MMR algorithm. This is due to the appearance of low-frequent multiword expressions and therefore the increase of diversity in input data. Low results were achieved by the MMR method on only thesaurus-based input data, which possibly can be explained by the differences between information described in the thesaurus and the real cluster structure.

The best results were obtained on cluster-based thematic nodes, taking into account multiword phrases and similarity between words and expressions

Input data model	Score
Simple bag-of-words	57,8%
Bag-of-words with multiword expressions	53,1%
Bag-of words with thesaurus information	52,6%
Bag-of-words with cluster-based thematic nodes	**59,8%**

Table 3: MMR method evaluation results for various input data models

6 Conclusion

In this paper we have proposed to use the discourse structure of natural language texts to extract sets of semantically similar expressions representing different participants of the text story – thematic nodes. We described two experiments on news clusters: multiword expression extraction and cluster-based thematic node construction. In addition to known methods of context comparison, we exploited the co-occurrence frequency in neighboring sentences to detect the semantic similarity of language expressions. We also combined several heterogeneous features for the thematic node construction:

- Formal resemblance features
- Information from the pre-defined resource (Russian language thesaurus RuThes [23])
- Context-based features

The evaluation of the introduced method showed that the cluster-based thematic nodes can improve the overall performance of the multi-document summarization algorithm.

In future we are going to use cluster-based thematic nodes for various operations as cluster refining, novelty detection, sub-clustering and etc.

References

[1] Alexeev A., Loukachevitch N. Automatic detection of near-synonyms in news clusters. *In: Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference Dialog`2011, pp. 32-40* (2011)

[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] Barzilay R., Lee L.: Learning to Paraphrase: an Unsupervised Approach Using Multiple Sequence Alignment. *In: Proceedings of HLT/NACCL-2003* (2003)
[4] Barzilay R., McKeown K.: Extracting Paraphrases from a Parallel Corpus. *In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (2001)
[5] Blei D., Griffiths T., Jordan M., Tenenbaum J. Hierarchical topic models and the nested chinese restaurant process. *In: Neural Information Processing Systems (NIPS) (*2003)
[6] Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. *In: Journal of Machine Learning Research, 3:993-1022* (2003)
[7] Carbonell J., Goldstein J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. *In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 335-336* (1998)
[8] Celikyilmaz A., Hakkani-Tur D. A Hybrid Hierarchical Model for Multi-Document Summarization. *In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pp. 815–824* (2010)
[9] Dang V., Xue X., Croft B. Context-based Quasi-Synonym Extraction. *CIIR Technical Report* (2009)
[10] Dijk van T.: Semantic Discourse Analysis. *In: Teun A. van Dijk, (Ed.), Handbook of Dis-course Analysis, vol. 2., pp. 103-136, London: Academic Press* (1985)
[11] Dobrov B., Pavlov A.: Basic line for news clusterization methods evaluation. *In: Proceedings of the 5-th Russian Conference RCDL-2010* (2010) (in Russian)
[12] 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)
[13] 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)
[14] Dolan B., Quirk Ch., Brockett Ch.: Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. *In: Proceedings of COLING-2004* (2004)
[15] 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)
[16] Griffiths T., Steyvers M. Finding scientific topics. *In: Proceedings of the National Academy of Sciences of the United States of America, Vol. 101, No. Suppl 1. (6 April 2004), pp. 5228-5235* (2004)
[17] Haghighi A., Vanderwende L.: Exploring Content Models for Multi-Document Summarization. *In: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, Boulder, Colorado, pp. 362–370* (2009)
[18] Harnly A., Nenkova A., Passonneau R., Ram-bow O.: Automation of summary evaluation by the pyramid method. *In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'2005), Borovets, Bulgaria* (2005)
[19] Hasan R.: Coherence and Cohesive harmony. *J. Flood, Understanding reading comprehension, Newark, DE: IRA, pp. 181-219* (1984)
[20] 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)
[21] 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)
[22] Loukachevitch N.: Multigraph representation for lexical chaining. *In: Proceedings of SENSE workshop, pp. 67-76* (2009)
[23] 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)
[24] Ng V.: Machine learning for co-reference resolution: from local classification to global ranking. *In: Proceedings of ACL-2005* (2005)
[25] Passonneau R.J., Nenkova A., McKeown K.R., Sigelman S.: Applying the pyramid method in DUC 2005. *In: Proceedings of the Document Understanding Conferences (DUC'2005), Vancouver, Canada* (2005)
[26] Turdakov D., Lizorkin D. HMM Expanded to Multiple Interleaved Chains as a Model for Word Sense Disambiguation. *In: Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computations, pp. 549–559* (2009)
[27] 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)
[28] Yang H., Callan J.: A metric-based framework for automatic taxonomy induction. *In: Proceedings of ACL-2009* (2009)