Summarizing News Clusters on the Basis of Thematic Chains

Loukachevitch Natalia, Alekseev Aleksey
Lomonosov Moscow State University
Leninskie Gory, Moscow, 119991, Russian Federation
E-mail: louk_nat@mail.ru, a.a.alekseevv@gmail.com

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 chains. 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 extracting multiword expressions and constructing thematic chains. The main difference of thematic chains in comparison with lexical chains is the basic principle of their construction: thematic chains are intended to model different participants (concrete or abstract) of the situation described in the analyzed texts, what means that elements of the same thematic chain cannot often co-occur in the same sentences of the texts under consideration. We evaluate our method on the multi-document summarization task.

Keywords: lexical cohesion, text coherence, thesaurus

1. Introduction

Automatic text summarization is one of important techniques for document processing implemented in various information systems. Multi-document summarization of news flows is utilized in many online news services, which cluster news articles devoted to the same event, and create summaries for such a news cluster. A lot of multi-document summarization methods of news documents were proposed and evaluated (Nenkova, McKeown, 2012).

A summary should present the main topics of a news cluster. Such a topic can be expressed with the variety of semantically related words. Therefore one of well-known techniques for text summarization exploits specific constructions – lexical chains, which unite such semantic groups of words, and are usually built on the basis of predefined lexical resources (usually WordNet) (Barzilay, Elhadad, 1998; Brunn et al., 2001; Reeve et al., 2006; Ye et al., 2007). Other resources used as a basis for lexical chains include Roget's thesaurus (Jarmasz, Szpakowicz, 2013), AGROVOC thesaurus (Medelyan, 2007).

However, any hand-crafted resource is insufficient to describe the rapidly changing world and language. Therefore the performance of lexical chain techniques is seriously restricted with the resource coverage (Nenkova, McKeown, 2012). To overcome this restriction Stokes et al. (Stokes et al., 2004) proposed to use statistical associative relations of words in a text corpus, which should enrich constructed lexical chains with additional information. Besides, some robust, resource-independent methods such as latent semantic analysis or Bayesian topic models (Blei, Ng, 2003; Griffiths, Steyvers, 2004) were applied to automatic summarization (Celikyilmaz, Hakkani-Tur, 2010; Li et al., 2012).

In this paper we propose the notion of so-called *thematic chains,* construction of which is based on several sources:

- a hand-made lexical resource,
- extraction of unknown collocations from texts under consideration,
- exploitation of several similarity types (predefined and context-based) between language expressions for forming thematic chains.

The main difference of thematic chains in comparison with lexical chains is the basic principle of their construction: thematic chains are intended to model different participants (concrete or abstract) of the situation described in the analyzed texts, what means that elements of the same thematic chain cannot often co-occur in the same sentences of the texts under consideration.

Besides, the thematic chains are able to include words and phrases extracted on the fly. Word sense disambiguation is not required but the ambiguity of lexical units described in a lexical resource is taken into account.

In this paper we present our approach to thematic chain construction, describe used sources of information and the algorithm combining information on several types of similarities between words (expressions) to create thematic chains. We experiment with Russian news clusters and as a thesaurus we use large Russian thesaurus RuThes, which was lately published online (http://labinform.ru/ruthes/index.htm).

The structure of the paper is as follows. In Section 2 we review lexical chain approaches. In section 3 we explain the notion of the thematic chain and basic principles for its construction. In section 4 thematic chain construction algorithm is presented. In section 5 we describe the evaluation of our approach on the basis of generic summarization of Russian news clusters.

2. Related Work

A lexical chain is a set of words (usually nouns or noun groups) gathered on the basis of predefined semantic relations (usually, repetitions, synonymy, hyponymy, meronymy) (Halliday, Hasan, 1976). The first Wordnet-based algorithm of lexical chain construction was described in (Hirst, St-Onge, 1998). The authors developed a

greedy algorithm that moves through a text from the beginning and links a current word to one of the existing lexical chains. Selection of an appropriate chain is based on the strength of a relation between the word and an element of the lexical chain. It was noted that the greediness of the algorithm can often lead to the choice of incorrect sense of an ambiguous word.

Barzilay, Elhadad (Barzilay, Elhahdad, 1998) created a less greedy algorithm for identifying lexical chains and used them for text summarization. Their algorithm first segments the text, then for each sense of the noun in the segment, it attempts to insert the senses into all existing chains in every possible way. To decrease the number of variants the algorithm estimates the strength of existent chains and eliminate the weakest variants if the number of chains exceeds a threshold. At last, the so-called "strong chains" (more than two standard deviations above the mean in length) constructed for the whole text are used to generate a summary.

To overcome a bottleneck of WordNet coverage in works (Stokes et al., 2004; Doran et al., 2004) it was supposed to use additional information:

- statistical associative links between words;
- lexical chains for proper noun phrases.

To obtain statistical associations authors extracted word pairs in four-noun text windows within sentences. Associative relations between words (such as *actor - director*) were considered as the weakest type of relations and applied if other relations to existing lexical chains were not revealed. Such relations were established between word forms, not between synsets. To construct lexical chains between proper noun phrases, several types of relations were introduced: full match, partial fullword match, and partial word match (Doran et al., 2004).

(Li et al. 2007) study the use of lexical chains for query-based multidocument summarization. Construction of lexical chains begins from the most frequent synsets (the frequent half from the whole number of revealed synsets). Each such a synset generates its own lexical chain based on the pre-defined set of relations. At the second stage lexical chains with coinciding words are merged. To generate a query-based summary the weight of a sentence is partially based on the sum of weights of lexical chains mentioned in this sentence (Li et al. 2007).

Another approach to linguistically motivated structuring of topical words in text clusters is described in (Harabagiu, Lacatusu, 2005; Harabagiu, Lacatusu, 2010). They proposed to reveal in texts so called 'topic themes', which captures different aspects of the text topic. A topic theme is based on predicate-argument structures and consists of (Harabagiu, Lacatusu, 2010):

- the common predicate,
- the set of semantically consisted arguments,
- the arguments related to time, location, manner etc.

To extract predicates and attributes a semantic parser trained on PropBank annotations is utilized. Themes with the same predicate (or its paraphrases) and semantically consisted arguments are clustered. Every theme is associated with sentences where it was mentioned (fig. 1).

As a result in every cluster many themes are revealed, so Pinochet Trial cluster has 853 themes. Supervised binary classification is used to select from them the most important ones. Between selected themes lexical cohesion and discourse coherence relations (Marcu, Echihabi 2002) are established. Each type of discourse relations is recognized by a separate classifier. The proposed model demonstrated an improvement in ROUGE and Pyramid measures Harabagiu, Lacatusu, 2010).

We can see that in this approach a lot of techniques were applied including semantic role labeling, coreference resolution, paraphrase detection, a supervised technique for important theme detection (which requires additional manual efforts and possibly provides additional tuning to data), cohesion and coherence techniques. Every utilized technique has drawbacks and mistakes, the contribution of every single stage is unknown, and for many languages, including Russian, most similar tools are absent.

In the theoretical study Hasan (1984) introduces the concept of *cohesive harmony* that presents an attempt to formalize the internal and external structure of sentences in texts. Cohesive harmony is based on cohesion chains and semantic relations between members of the chains such as *agent, object, instrument* and so on.

Predicate: ARRESTED (PLACED UNDER ARREST) cluster size:38 frequency:15)
Arg0: (S_3,12,13) British police (S_1,4,4) they (S_2,11,11) they
Arg1: (S_1,7,12) former Chilean dictator Gen. Augusto Pinochet
(S_3,0,2) Pinochet, 82 (S_2,14,14) Pinochet
Arg2: (S_2,15,43) on allegations that he murdered an unidentified number of Spaniards in Chile between Sept. 11, 1973, the year he seized power, and Dec. 31, 1983
(S_1,13,23) on allegations of murdering Spanish citizens during his years in power
ArgM-LOC: (S_3,8,9) in London
ArgM-TMP: (S_3,10,10) Friday
ArgM-ADV: (S_2,0,5) Responding to a Spanish extradition warrant

Figure 1. A theme from Pinochet Trial cluster (Harabagiu, Lacatusu, 2010)

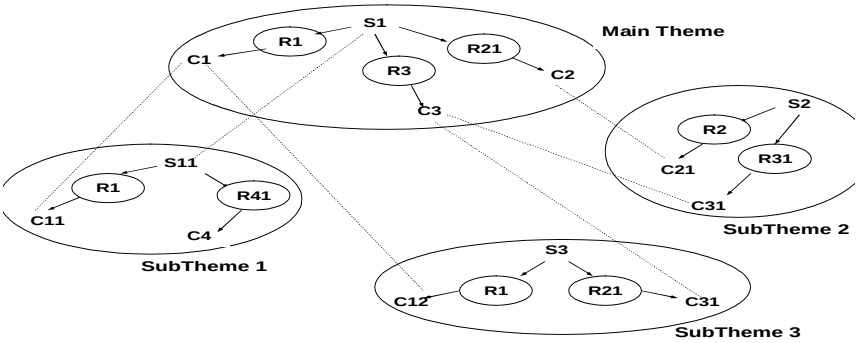

Figure 2. The hierarchy of themes in a natural language text and links between levels of the hierarchical structure where *S1, S11, S2, S3* are predicates describing a situation, *C1… C4* are entities participating in the described situations, *Ri* are roles of entities.

Hasan explains that "the source of unity … resides in the fact that similar 'things' are said about similar/same 'entities', 'events' etc" (Hasan, 1984). Experiments described in (Hirst, Morris, 2005) showed that texts with more participating in cohesive harmony, and fewer chains left isolated, were consistently judged as more coherent.

3. Main Idea of our Approach

However, above-mentioned work (Harabagiu, Lacatusu, 2010) has demonstrated the high complexity of automatic following semantic structures through the text.

Before explaining the main idea of our approach to lexical grouping in a text or a text cluster let us consider an example – the newspaper article from Washington Post about relations between United Kingdom and Scotland published 13 February, 2014. We show here the beginning of the article:

U.K. to Scotland: Walk away, lose the pound

*The **British** government warned **Scotland** on Thursday that if it votes to leave the **United Kingdom**, it would not be able to keep the **British** currency, the venerable pound sterling.*

*If **Scotland** walks away from the **U.K.**, it walks away from the **U.K.** pound," **British** Chancellor of the Exchequer George Osborne said in a speech in **Edinburgh** on Thursday, upping the ante in the battle over **Scottish** independence.*

*Commentators called the speech one of the most important developments in the fight for **Scotland's** future, with clear battle lines being drawn by the **British** government, which until now has resisted discussing the terms of a possible breakup.*

*The pro-independence **Scottish** National Party (SNP) accused the **British** government of bullying and bluffing...*

The first thing we would like to note that according to the classic algorithm of lexical chaining (Hirst, St-Onge, 1998) one of prominent lexical chains in this text should comprise the mentions of United Kingdom and Scotland:

U.K, Scotland, British, Scotland, United Kingdom...

However, in this text knowledge about relations between these two entities is not utilized for lexical cohesion, the text discusses the interaction between them. The United Kingdom and Scotland are different participants of the discussed situation. To present the text contents correctly, their mentions in the text should generate two different lexical chains. Thus, from this consideration we can conclude that traditional lexical chain techniques (Hirst, St-Onge, 1995), which prefer close co-occurrence of sense-related words, should take into account additional information allowing the differentiation between establishing the cohesion relation or the co-argument relation considering semantically related words.

If to try to apply to this text the approach accounting for predicate-argument structures as described in ((Harabagiu, Lacatusu, 2005; Harabagiu, Lacatusu, 2010) - see previous section) then it is possible to see how various the mentioned predicates are: *warn, vote, leave, keep, walk away* etc. It is clear that establishing correct relations between them is quite a difficult task requiring a lot of resources.

However, gathering main interacting arguments of these predicates seems to be much easier. The arguments correspond to participants of the situation described in the text and their extraction can be based on modeling the global coherence of the text.

4. Thematic Analysis of News Clusters: Theoretical Basis

Van Dijk (Dijk, 1985) describes the thematic 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 (the main theme) is usually described in terms of less general themes, which in turn are 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 appropri-

ately according to the given local coherence criteria, but the sequence would go simply astray without some constraint on what it should be about globally" (Dijk, 1985).
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. 2). So we suppose that numerous lexical cohesion relations in a text should refer to the participants of the main theme (Loukachevitch, 2009). We call such a node of links to more important thematic element – *thematic node*.
In addition, we suppose that:

1. interactions of participants are discussed in specific sentences, therefore the more words (expressions) are mentioned in the same sentences of a text, the more the possibility of their correspondence to different participants of the described situation is;
2. every participant can be mentioned in a text by means of several words or expressions: we suppose that there is the most frequent (basic) naming of a participant, therefore this group of related words and expressions – *thematic chain* - is constructed in the form of a *thematic node*: the main expression and related expressions.

So we think that an important step to reveal the thematic structure of a document is to reconstruct thematic chains having the node-like internal structure.
In comparison with LDA topics, thematic chains do not comprise words co-occurring in the same documents or the same sentences - each thematic chain 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 (Hirst, St-Onge, 1998), which try to construct chains of semantically related expressions in texts, thematic chain elements are supposed to be related to its main element (center of the thematic chain), and if two related expressions (for example, *doctor* and *patient*) co-occur in the same sentences of the text, it means that their relations represent the focus of the text contents, they are related to different participants of the text theme, and they should be assigned to different thematic chains. And on the contrary, if two expressions rarely co-occur in the same sentences, but frequently co-occur in neighbouring sentences, then they may be considered as the elements of the same thematic chain.
A news cluster is not a coherent text but cluster documents are devoted to the same theme, describing the same event or situation. Therefore, statistical features of the thematic structure are considerably enhanced in a cluster, of related news, and on such a basis we try to extract unknown information from a cluster.

5. Proposed Algorithm for Thematic Analysis of News Clusters

Our principal aim in news cluster processing is to reveal the main participants of the situation described in a cluster by means of constructing thematic chains. This procedure is based on several types of similarities between expressions. In addition, the necessary condition for inclusion of two expressions in the same thematic chain is their high co-occurrence frequency in neighboring sentences in comparison with the same sentence co-occurrence frequency.
The cluster processing consists of four main stages. At the first stage word context statistics is accumulated. Multiword expressions, which can also denote a participant, are extracted at the second stage. At the third stage similarity measures between mentioned language expressions are calculated, and these similarities are utilized for constructing thematic chains at the last, fourth stage.

5.1 Extracting Word Contexts

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 extracting NS contexts 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 fragments from neighbouring sentences. Each context type obtains 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 exploit 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 the 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.

5.2. Extraction of Multiword Expressions

We consider recognition of multiword expressions as a necessary step before constructing thematic chains. Recognizing multiword expressions is usually based on accounting for frequencies of word sequences. However, a

news cluster is a structure where various word sequences are repeated a lot of times. We suppose that the main criterion for extracting multiword expressions 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:

$$Near > 2 \cdot (AcrossVerb + NotNear)$$

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 news cluster about U.S. military base in *Kyrgyzstan.*

Two measures of quality were applied for multiword expression extraction. 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 arranging them in the descending order of importance. The proposed algorithm for extracting multiword expressions showed 91.4% precision and 72.6% recall result, which is enough for further constructing thematic chains.

5.3 Similarity Features

The set of six main similarity features is used for determining semantically related expressions and for the following constructing thematic chains. Some of these features are based on context information, extracted directly from the news cluster under consideration. Other features reflect the formal resemblance of expressions and information from pre-defined resources. Each similarity feature contributes some points to the overall similarity score of a candidate pair.

Context-dependent features include*:*

***Neighbouring sentence feature* (NSF).** This feature is based on the described discourse model and reflects the difference between the co-occurrence of thematic chain elements in the same and neighboring sentences. NSF feature is also a regulatory feature. It means that a candidate pair could not be included in the same thematic chain if NSF feature is less than a predefined threshold.

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 next value is the basis of this feature:

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

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

$$NSF = \min\left(1, \frac{C}{Avg\ (C)}\right)$$

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

***Strict context feature* (SC).** The SC feature is based on the comparison of fixed order contexts of two words. The more identical templates a candidate pair shares the more its similarity is. Currently, strict order contexts are constructed as 4-gramms: 2 content words to the left and 2 content words to the right from a target expression within a sentence. SC similarity score for two expressions t_i, t_j is counted as the relative value of the number of the same strict contexts for t_i, t_j to the maximal number strict contexts revealed in the current text cluster.

Cosine similarity feature (**S**calar **P**roduct **S**imilarity, **SPS**) represents the cosine similarity between sentence contexts of specific words or expressions. Word vectors are constructed from content words between punctuation marks.

Context-independent features comprise:

Formal resemblance feature (**B**eginning **S**imilarity, **BS**) between words and expressions based on the same beginning of words. Words with the same 5-letter beginning or prefix plus the same letter are considered as similar.

BS weights of phrases t_i, t_j are counted from their component similarity (function words are excluded) and based on modified Dice measure, adapted to comparison of relatively short sequences of words (currently, k=3):

$$\underset{t_i, t_j \in T}{weight_{BS}}(t_i, t_j) = \begin{cases} \dfrac{n_{word}(t_i \cap t_j) + k}{n_{word}(t_i \cup t_j) + k}, \text{ if } n_{word}(t_i \cap t_j) > 0 \\ 0, \text{ npu } n_{word}(t_i \cap t_j) = 0 \end{cases}$$

where $n_{word}\ (t_i \cap\ t_j)$ is the number of similar words in phrases t_i, t_j ,

$n_{word}\ (t_i \cup\ t_j) = n_{word}(t_i) + n_{word}\ (t_j) - n_{word}\ (t_i \cap\ t_j)$.

Thesaurus similarity feature (**T**hesaurus **S**imilarity, **TS**) reflects the semantic distance between expressions based on a pre-defined thesaurus. We use RuThes thesaurus of Russian language (Loukachevitch, Dobrov, 2014. The publicly available version of RuThes contains around 100 thousand Russian words and expressions. If compared to WordNet-style resources RuThes is organized as a united semantic net where different parts of speech can be text entries of the same concepts. Ambiguous words in RuThes are described similar to WordNet-style resources through attachment to several concepts.

TS feature can be computed only if the both expressions are text entries of the thesaurus. Currently, TS feature linearly depends on the minimal path length between thesaurus concepts which the text entries t_i, t_j are assigned to:

$$TS = 1 - 0.2 \cdot N_{rel}$$

where N_{rel} is the length of the minimal path between concepts.

Features / Pairs	Context-independent		Context-dependent			SCORE
	BS	TS	NSF	SC	SPS	
президент России – президент РФ (President of Russia - RF President)	0.66	1.00	0.00	0.50	0.68	**2.84**
инвестгруппа – инвестиционная группа (investgroup - investing group)	0.80	1.00	0.40	0.00	0.63	**2.83**
ГМК Норильский никель – Норильский никель (GMK Norilsk Nickel - Norilsk Nickel)	0.82	1.00	0.40	0.00	0.21	**2.44**
Российская Федерация – Россия (Russian Federation - Russia)	0.80	1.00	0.00	0.00	0.51	**2.31**
отставка – отставка с должности resignation - resignation from the post	0.80	1.00	0.40	0.00	0.00	**2.20**

Table 1. Top-ranked pairs of similar words and phrases for ALROSA text

If a word is described as ambiguous in RuThes, then its TS values with other expressions in the news cluster aresimply decreased with applying the special parameter *k (k<1)*, that is the whole processing is performed without real word sense disambiguation.

Embedded objects similarity feature (EOS). This feature plays its role when thematic chains having several elements are compared. This feature is equal to 1, when two thematic chains contain the same word or phrase as their members (it is possible that an expression can be a member of two thematic chains - see section 5.4).

5.4 Constructing Thematic Chains

The supposed structure of the thematic chains is as follows:

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

The algorithm begins to construct thematic chains from the most similar pairs of expressions and consists of the following steps:

1. The candidate pair of expressions with the maximal similarity score is taken;
2. 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 chain;
3. The second participant of the pair can further be joined in a similar manner to another thematic chain.

The iterative process proceeds until the top-ranked pair score reaches a pre-defined threshold (see Table 1). The examples of obtained thematic chains from a news cluster devoted to resignation of ALROSA diamond mining company president Alexander Nichiporuk are as follows (translated from Russian):

- **company**, *company's share, share, company owner, company merge, controlling percentage of, block of shares, owner, ownership;*

- ***post,*** *removal from post, office, dismissal from office, departure from office, dismissal etc.*

5.5 Analysis of Obtained Thematic Chains

Looking at the first example chain we can see that such an entity word as *company* is linked together with persons (*owner*), relations (*ownership*), processes (*company merge),* securities *(share).* In fact, in the news cluster the ownership problems of ALROSA are discussed as the main factor of ALROSA President resignation. In the article discussing the same issue in English (http://www.mineweb.com/mineweb/content/en/mineweb-diamonds-and-gems?oid=22306&sn=detail) we can see the following distribution of similar concepts (expressed with other words):

The Alrosa sources have told Mineweb Minister Kudrin wants to step down from the chairmanship, because he is tired of fighting Shtirov over the federal ***shareholding takeover of the company****. ..., Shtirov is still resisting the final acquisition by the government in Moscow of* ***50% plus one share.*** *... Last autumn, after attacking Shtirov's resistance to the federal* ***takeover****, Vybornov sided with Shtirov in order to campaign for the removal of chief executive Nichiporuk...The current federal government* ***stake*** *in Alrosa, held by the government directly and by state-controlled Vneshtorgbank, is 48%. By complex schemes of trusteeship, small* ***blocs of shares*** *are controlled by leadings figures like Shtirov..*

So in the text we can see the ownership thematic chain going through sentences and linking then together. The ownership aspect is considered as a separate issue together with personal participants of the described situation (*Kudrin, Shtirov, Vybornov, Nichiporuk).* The presented chain looks like a traditional lexical chain (Barzilay, Elhadad, 1998; Hirst, St-Onge, 1998).

Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-S	ROUGE-SU	Pyramids
MMR + Chains	**0.625**	**0.416**	**0.602**	**0.355**	**0.366**	**0.645**
MMR	0,576	0.381	0.555	0.298	0.310	0.617
SumBasic + Chains	0.522	0.229	0,493	0.243	0.255	0.602
SumBasic	0.518	0.247	0,498	0.231	0.243	0.575

Table 2. Rouge and Pyramid scores for summaries

But in comparison with those approaches we control the distribution of the chain elements in different sentences of a document and, additionally, take to account the context similarity of words and expressions.

6. Experiments

For evaluation of our approach in news cluster summarization we took two well-known summarization algorithms Maximal Marginal Relevance (MMR) (Carbonel Goldstein, 1998) and SumBasic (Vanderwende et al., 2007).

Maximal Marginal Relevance approach (Carbonel Goldstein, 1998), was used in many later approaches (Nenkova, McKeown, 2012; Radev et al., 2004).

MMR algorithm was proposed for query-based summarization but later it has been adopted for generic document summarization. It is based on greedy selection of sentences when at each step the algorithm picks up the sentence that maximally similar to the whole document (a news cluster in our case) and minimally similar to the sentences already included in the summary.

SumBasic was developed to employ the idea of intensive using frequencies for summarization. Each sentence in the input obtains the weight equal to the average probability of the content words in the sentence, calculated on the basis of the whole input for summarization. Then SumBasic picks the best scoring sentence in a greedy fashion, after that the probability of each word that appears in the chosen sentence is recalculated to a smaller value. Such an iterative process proceeds until the desired summary length is achieved.

We substituted initial words in the sentences with corresponding thematic chains and then applied the chosen summarization techniques. Every thematic chain (*tc*) has the weight equal to the sum of frequencies of its elements:

$$weight\ (tc) = \sum_{tc_{elem\_i} \in tc} freq\ (tc_{elem\_i})$$

The weight of the main element of the chain is equal to the whole weight of the chain. The chain elements have weights proportional to their similarity to the main element.

After the transfer from single word weights to weights based on constructed thematic chains it is possible to apply mentioned summarization methods: MMR and SumBasic.

For the evaluation expert summaries were prepared. On the basis of expert summaries we evaluated the automatic summaries using ROUGE metrics (Lin, 2004) and Pyramid method (Harnly et al., 2005). ROUGE package includes several variants of metrics, for most of them it was shown that they significantly correlate with human judgments in various conditions (Lin, 2004). In Rankel et al. (2013) it was also demonstrated that the combination of several ROUGE metrics corresponds to human judgments even more.

The Pyramid method is a formalized procedure, which gives the possibility to evaluate the automatic summary coverage of the news cluster main facts. The method is based on extraction of all "information nuggets" from expert summaries, or Summary Content Units (SCUs) (Harnly et al., 2005). Each SCU obtains the weight equal to the number of expert summaries, where this SCU occurred. Finally, each automatic summary could be assessed for the presence of extracted SCUs.

So, in Table 2 we can see that introduction of chains to MMR method considerably improves the performance of the automatic summarizer in terms of several ROUGE metrics, for SumBasic the improvements are demonstrated in three ROUGE measures. For Pyramid metric both variants based on thematic chains are better than initial variants.

Conclusion

In this paper we present our approach to construction of thematic chains – structures similar to lexical chains. In contrast to existing techniques for lexical chaining, our approach supposes well-defined roles for the thematic chains – they present interacting participants of the situation described in the text. We incorporate several sources of information for including of words and expressions in the same thematic chains. And one of the most important features we use is the co-occurrence of the expressions in the same sentences of the texts, because we suppose that the more frequently two expressions are mentioned in the same sentences of the text, the more probable that they correspond to different participants of the described situation.

We evaluated our approach in news cluster summarization using the modification of well-known multidocument summarization methods such as MMR and SumBasic. As a result, MMR methods were significantly improved for all ROUGE metrics, and SumBasic was improved for the most ROUGE metrics. Both variants based on the thematic chains improved their performance in Pyramid scores.

The approach can be applied to various languages and can be based on various lexical resources (for example, such as wordnets). We are going to test the proposed approach for English documents. The technique can also be used in other NLP applications such as text categorization, semantic duplicate identification etc.

Acknowledgements

This work is partially supported by Russian Foundation for Basic Research grant N 14-07-00383.

References

Barzilay, R., Elhadad, M. (1999). Text summarizations with lexical chains. In Inderjeet Mani and Mark Maybury (Eds.), *Advances in Automatic Text Summarization*, MIT Press.

Blei, D., Ng, A., Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993-1022.

Brunn, M., Chali, Y., Pinchak, C. (2001). Text Summarization Using Lexical Chains. In *Proceedings of the Document Understanding Conference,* pp. 135-140.

Carbonell, J., Goldstein, J. (1998). 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.

Celikyilmaz, A., Hakkani-Tur, D. (2010). A hybrid hierarchical model for multi-document summarization. In *Proceedings of Coling-2010.*

Dijk, van T. (1985). Semantic Discourse Analysis. In Teun A. van Dijk, (Ed.), *Handbook of Discourse Analysis*, vol. 2., London: Academic Press, pp. 103-136.

Griffiths, T., Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. Suppl 1, pp. 5228-5235.

Halliday, M., Hasan, R. (1976). *Cohesion in English*. - Longman, London.

Hasan, R. (1984). Coherence and Cohesive harmony. In J. Flood, (Ed.), *Understanding reading comprehension*,. Newark, DE: IRA, pp. 181-219.

Harabagiu, S., Lacatusu, F. (2005). Topic themes for multi-document summarization. In *Proceedings of the 28th international ACM SIGIR conference*, pp. 202-209.

Harabagiu, S., Lacatusu, F. (2010). Using topic themes for multi-document summarization. *ACM Transactions on Information Systems (TOIS)*, 28(3), 13.

Harnly, A., Nenkova, A., Passonneau, R., Ram-bow O. (2005). Automation of summary evaluation by the pyramid method. In *Proceedings of the International Conference on RANLP-2005.*

Hirst, G., St-Onge, D. (1998). Lexical Chains as representation of context for the detection and correction malapropisms. In C. Fellbaum, (Ed.), WordNet: *An electronic lexical database and some of its applications*. Cambrige, MA: The MIT Press.

Hirst, G., Morris, J. (2005). The Subjectivity of Lexical Cohesion in Text. In James C. Chanahan, Yan Qu, and Janyce Wiebe, (Eds), *Computing attitude and affect in tex*t. Springer, Dodrecht, The Netherlands. pp. 41–48.

Jarmasz, M., Szpakowicz, S. (2012). Not As Easy As It Seems: Automating the Construction of Lexical Chains Using Roget's Thesaurus. arXiv preprint arXiv:1204.0257, 2012.

Li, J., Sun, L., Kit, C., Webster, J. (2007). A Query-Focused Multi-Document Summarizer Based on Lexical Chains. In *Proceedings of DUC-2007*.

Li, J., Li, S., Wang, X., Tian, Ye, Chang, B. (2012). Update Summarization Using a Multi-level Hierarchical Dirichlet Process Model. In *Proceedings of Coling-2012*, pp. 1603-1618.

Lin, C.-Y. (2004). ROUGE: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization* (ACL'2004). Barcelona, Spain, pp. 74-81.

Lin, C. Y., Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of theColing-2000,* V.1, pp. 495-501.

Loukachevitch, N., Dobrov, B (2014). RuThes Linguistic Ontology vs. Russian Wordnets. In *Proceedings of Global Wordnet Conference GWC-2014*.

Loukachevitch, N. (2009). Multigraph representation for lexical chaining. In *Proceedings of SENSE workshop*, pp. 67-76.

Marcu, D., Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL-2002.*

Medelyan, O. (2007). Computing Lexical Chains with Graph Clustering. In *Proceedings of the ACL-2007 Student Research Workshop*, pp. 85-90.

Nenkova, A., McKeown K. (2012). A survey of text summarization techniques. *Mining Text Data*, Springer 2012, pp. 43-76.

Radev, D., Jing, H., Sty, M., Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing and Management*, 40, pp. 919–938.

Rankel, P., Conroy, M., Dang, H., Nenkova, A. (2013). A decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art. In *Procedings of ACL-2013*, pp.131-136.

Reeve, L., Han, H., Brooks, A. (2006). BioChain: Using Lexical Chaining for Biomedical Text Summarization In *Proceedings of the ACM Symposium on Applied Computing*, pp.180-184.

Silber, G., McCoy, K. (2003). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 29 (1).

Stokes, N., Carthy, J., Smeaton, A.F. (2004). SeLeCT: A lexical Cohesion based News Story Segmentation System. *Journal of AI communications*, 17(1), pp. 3-12.

Vanderwende, L., Suzuki, H., Brockett, C., Nenkova, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Management Journal*, Volume 43 Issue 6, November, pp. 1606-1618.