Multigraph Representation for Lexical Chaining

Natalia Loukachevitch
Research Computing Center of M.V. Lomonosov Moscow State University
Moscow, Russia
louk@mail.cir.ru

Abstract

In this paper lexical cohesion modeling is considered. We argue that to model lexical cohesion in connected texts it is not enough to find related words in a text. It is important to take into account relations between entities participating in the described situations. Consideration of this factor gives the possibility to develop more flexible lexical chaining algorithms and to construct lexical chains more corresponding to the discourse structure. The multigraph representation seems to be appropriate as a basis for lexical cohesion modeling.

1 Introduction

One property of coherent texts is presence of multiple lexical repetitions and closely related words in the texts. This phenomenon is called 'lexical cohesion' [4] and usually modeled by means of lexical chains – sets of related words revealed on the basis of thesaurus relations. A lexical chain is a chain of words in which the criterion for inclusion of a word is some kind of cohesive relationship to a word that is already in the chain [13].

For example, in text (*) we can see repetitions (*war crimes),* the full name of a corporation and its abbreviation *BBC,* use of derivative words such as *Ossetia, Ossetian*, part-whole relations as between *South Ossetia* and *Tskhinvali*[1]*:*

> *(*) The British Broadcasting Corporation is the first foreign news agency to be granted unrestricted access to the breakaway Georgian republica of South Ossetia since Georgian forces attacked the capital Tskhinvali. Journalists working for the BBC have unearthed evidence of Georgian war crimes against South Ossetian civilians. The indiscriminate use of force is a clear and serious violation of the Geneva Convention and can constitute a war crime.*

So there are several evident lexical chains as

1) British Broadcasting Corporation, news agency, journalists, BBC,

2) South Ossetia, Tskhinvali, South Ossetian

3) Georgian, Georgian, Georgian

4) war crimes, Geneva Convention, war crime

[1] The text is taken from World Socialist Web Site (http://www.wsws.org/articles/2008/nov2008/geor-n10.shtml) and it is fully based on the BBC reportage located at http://news.bbc.co.uk/1/hi/world/7692751.stm. Using this text example we would not like to discuss any political positions. We need this example to demonstrate the dynamic nature of lexical chains because their construction should depend not only on the static thesaurus knowledge but on the text content and structure.

Automatic construction of lexical chains is considered as an important step to construction of the discourse structure and better understanding of the text content. The first lexical chaining algorithm based on Roget's thesaurus was proposed in [13]. Next approaches usually utilize lexical relations from WordNet [1, 7, 16]. Graph representations of thesaurus relations as a basis for lexical chaining [1, 12, 16] are often used.

Initially it was suggested that lexical chains are intuitively clear for text readers. Lately it was shown that lexical chaining by humans is a very subjective procedure [6, 8, 14]. Two readers can propose different lexical chains even for a small text. So in text (*) it is not clear if lexical chains 2) and 3) are separate chains or they should be joined in a single chain, because at the moment of text creation South Ossetia was officially a part of Georgia, therefore South Ossetia and Georgia were very related entities.

In this paper we will describe the main techniques of automatic construction of lexical chains. We will consider experiments showing distinctions in lexical chain construction between several annotators. We argue that the subjectivity of manual construction of lexical chains is due to the fact that an important factor is not considered and the more appropriate form of description of lexical cohesion in a given text is not a graph but a multigraph with two types of relations between vertices.

2 Methods of Lexical Chain Construction

Most techniques to lexical chaining are based on lexical relations described in WordNet [1, 7].

Hirst and St-Onge [7] divide lexical cohesion relations from WordNet into three categories: extra-strong (repetitions), strong (synonyms and symmetric relations) and medium-strong relations. Every next relation is weaker than previous one. Medium-strong relations include paths of the WordNet conceptual structure with maximum 5 links and can have different weights depending on path length.

The main stages in the proposed construction of lexical chains are as follows:

- the construction of lexical chains begins from the first words of a text;
- to insert the next word, its relations with members of existing lexical chains are checked;
- if there are such relations with any element of a chain then the new word is inserted in the chain. Only one lexical chain can be chosen: among several possible lexical chains a lexical chain with maximal weight of a relation with a current word is chosen.
- If a current word is ambiguous then the choice of a lexical chain determines the choice of a sense.
- For strong relations and medium-strong relations there are restrictions on distance between a current word and existing lexical chains.

Hirst and St.Onge [7] indicate the problems of the lexical chaining process such as extra cohesive relations and missing relations. The problems arise from several sources: (1) limitations in the set of relations in WordNet, or a missing connection; especially lack of situational relations (*school - child care*, *physician - hospital*); (2) inconsistency in the proximity in links in WordNet (*stew* and *steak* were not considered as related because distance of 6 synsets; *public* and *professionals* are considered as related - distance 4 synsets); (3) incorrect or incomplete disambiguation.

Barzilay and Elhadad, 1997 [1] pointed out that preceding lexical chains do not give enough information for correct disambiguation. Therefore they proposed to take all possible alternatives for word senses in a text and try to assign them to existing lexical chains. Barzilay and Elhadad define the best interpretation as the one with the most connections (edges in a graph). They define the score of interpretation as a sum of its chain scores, determined by the number and weight of the relation between chain members. When the number of possible interpretations is larger than a certain threshold then the weak interpretations are pruned.

To add situational relations to lexical chaining process Stokes et.al. [16] propose to use statistical associative relations of words in a text corpus.

O.Medelyan [12] considers relations described in an information-retrieval thesaurus as situational relations. She defines a lexical chain as a graph $G = (V, E)$ with nodes $v_i \in V$ being terms and edges $(v_i, v_j, w_{ij}) \in E$ restricting semantic relations between them, where w_{ij} is a weight expressing the strength of the relation. She proposes to build lexical chain candidates for the whole text, and the resultant graph is divided into lexical chains with the condition of the minimum path length between nodes $m < 4$. An algorithm for graph clustering divides the sparsely connected graph into dense lexical chains.

Most algorithms of lexical chaining assign a current word only to a single lexical chain. We found only one approach [8] where the lexical chaining algorithm enables to include current word into several lexical chains. In this case the overgeneration problem arises when too many lexical chains are generated [8].

3 Subjectivity of Lexical Chains

Lately considerable subjectivity of lexical chaining in experiments with human annotators was revealed. Hirst and Morris [6] introduce an example text (**) and show that even for such a short text authors of the paper had different opinions on available lexical chains:

> (**) *How can we figure out what a text means. One could argue that the meaning is in the mind of the reader, but some people think that the meaning lies within the text itself.*

So one author thinks that there exist two lexical chains. One chain is "understanding" chain including such words as *figure out, means, meaning, mind, think, meaning,* another chain is "text" chain, including words *text, reader, text.* Another author also distinguishes two lexical chains but words *means, meaning* were assigned to chain "text".

In [6] an experiment in manual lexical chaining is described. A study was conducted with five participants as readers of a general-interest article from the Reader's Digest on the topic of movie actors. Subjects were instructed to read the article and mark the word groups they perceive, using a different color of pencil for each different group. Subject's groups were compared in pairwise manner: for each pair of participants number of words they agreed was divided by the total number of words they used. Averaged over all pairs of participants, the agreement was 63%.

Hollingsworth and Teufel [8] describe an experiment on comparison of lexical chains created by different annotators for a scientific paper on computational linguistics. The task was to collect sets of related terms mentioned in the paper. A term can comprise a single word or a combination of words, all taken directly from the text.

As a result of the experiment considerable subjectivity of manual lexical chaining was demonstrated. The first annotator built 12 lexical chains, the second annotator constructed 22 lexical chains. Coincidence of main elements of lexical chains (the most frequent terms in the chain) appeared only in 4 lexical chains.

In this experiment all annotators assigned at least one term to several lexical chains. That is the principle of assignment of a term to a single lexical chain utilizing in automatic lexical chaining procedures seems to be too restrictive.

4 Cohesive Harmony, Discourse Structure and Lexical Chaining

Hasan [5] introduces the concept of *cohesive harmony*, which presents an attempt to formalize internal and external structure of sentences in texts. Cohesive harmony is based on cohesion chains and semantic relations between members of the chains. Semantic intrasentence relations are similar to case relations of Fillmore [3] or conceptual relations in conceptual graphs [15] such as *agent, object, instrument* and so on.

The actual rule for chain formation is that elements of a chain can be joined together if (at least) two instances of the same conceptual relation exist between them. Hasan explains that "the source of unity … resides in the fact that similar 'things' are said about similar/same 'entities', 'events' etc." [4, p.212]. Texts with more chains participating in cohesive harmony, and fewer chains left isolated, were consistently judged as more coherent [6].

From this consideration we can make the following conclusion

> *(***) if some entities participate in some situations or events in different roles more than once then these entities should not be represented as members of the same lexical chain.*

So in text (*) Georgia is several times mentioned as an agent of the attack, and South Ossetia is presented as an object of the attack. Therefore Georgia and South Ossetia should not be joined into the same lexical chain.

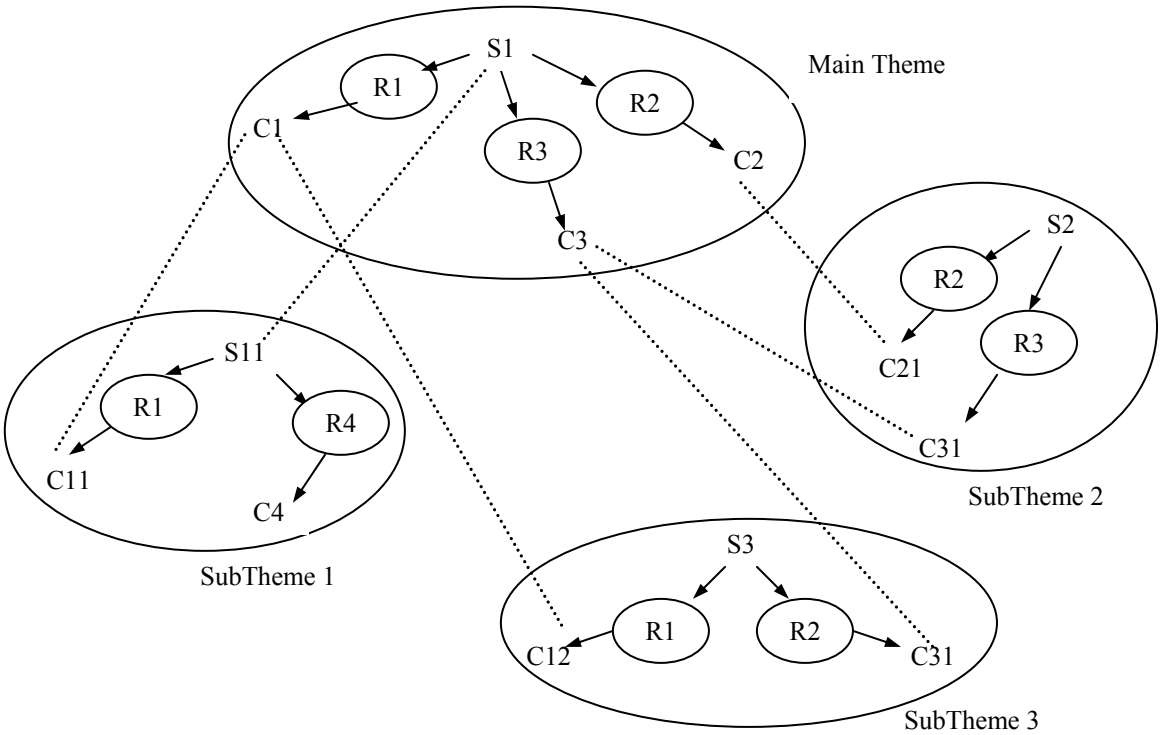

Figure 1: Main theme and its elaboration in subthemes.

The similar conclusion can be obtained by other means, on theoretical grounds of Van Dijk theory [2]. Van Dijk 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

macroproposition. 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, and so on (Figure 1: Si means a situation discussed in the main theme or in a subtheme of a text, Ci is an entity, participating in the situation, Ri is a relation between the situation and a participant). 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. 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 [2, p.115-116].

So lexical chains should also have some restrictions on their development. But in proposed techniques there are no any constraints.

In our opinion this restriction consists in the specific function of lexical chains in the topical structure. Lexical chains are responsible not only for local connections between sentences. They provide references between levels of the hierarchical structure of the text content.

To refer to the main theme of the whole text a subtheme has to include a main concept (concept of the text macroproposition) or its related concept; in sentences of a text such references look like lexical cohesion relations. Besides sentences in the text have to elaborate relations between entities of macroproposition. It means that many sentences have to include more that one entity from the macroproposition. Therefore the more two entities are met in the same sentences of a text, the more probable that they are representatives of different issues of the macroproposition and it means that they should be assigned to different lexical chains to present the text macrostructure correctly.

So again we can make a conclusion similar to (***):

> *(****) The more often words co-occur in the same clauses (simple sentences) of a text, the less they should be included to the same lexical chains.*

Co-occurrence of words in the same clauses can be treated as the generalized conceptual relation *R (X,Y)* [15].

Our consideration means that in lexical chaining experiments with human annotators the task for annotators should restrict similar terms extraction [6, 8] with the rule that if two entities interact with each other in a situation described in a text then they should not joined into the same lexical chain.

Recall that the proximity inconsistency problem of WordNet relations described in [7] (see section 2) then it should be stressed that this problem is due not only to the specific linguistic resource. Lexical chaining algorithms need to be more flexible and take into account the word co-occurrence factor.

5 Use of Multigraphs for Lexical Chaining Algorithms

Let us return to text (**) and its lexical chains (see Section 3). It will be remembered that two readers of the text disagreed on the allocation of words *means, meaning* to lexical chain *figure out, mind, think* or lexical chain *text, reader, text.*

We can analyze this text using information about co-occurrences of words in the text. In such a small text words *means, meaning* were mentioned three times in the same sentences with words *text, reader:*

*what **a text means***
*the **meaning** is in the mind of the **reader***
*the **meaning** lies within the **text** itself*

This indicates that the text (**) is devoted to consideration of relation "text – its meaning". *Text* and *Meaning* present different elements of the topical structure and therefore words *text* and *meaning* should belong to different lexical chains to correspond to the discourse structure of the text.

At the same time words *means, meaning* should not also be assigned to another lexical chain *figure out, think,* because these verbs govern clauses including words *means, meaning.* So relation between *figure out, think* and *meaning* is an important issue of the text fragment.

***figure out** what a text **means**…*
***think** that the **meaning** lies within the text itself.*

In our opinion words *means, meaning, meaning* should not be included in both lexical chains and form a separate lexical chain.

So lexical chains for text (**) should be as follows:

1) *text, reader, text.*
2) *figure out, think*
3) *means, meaning, meaning*

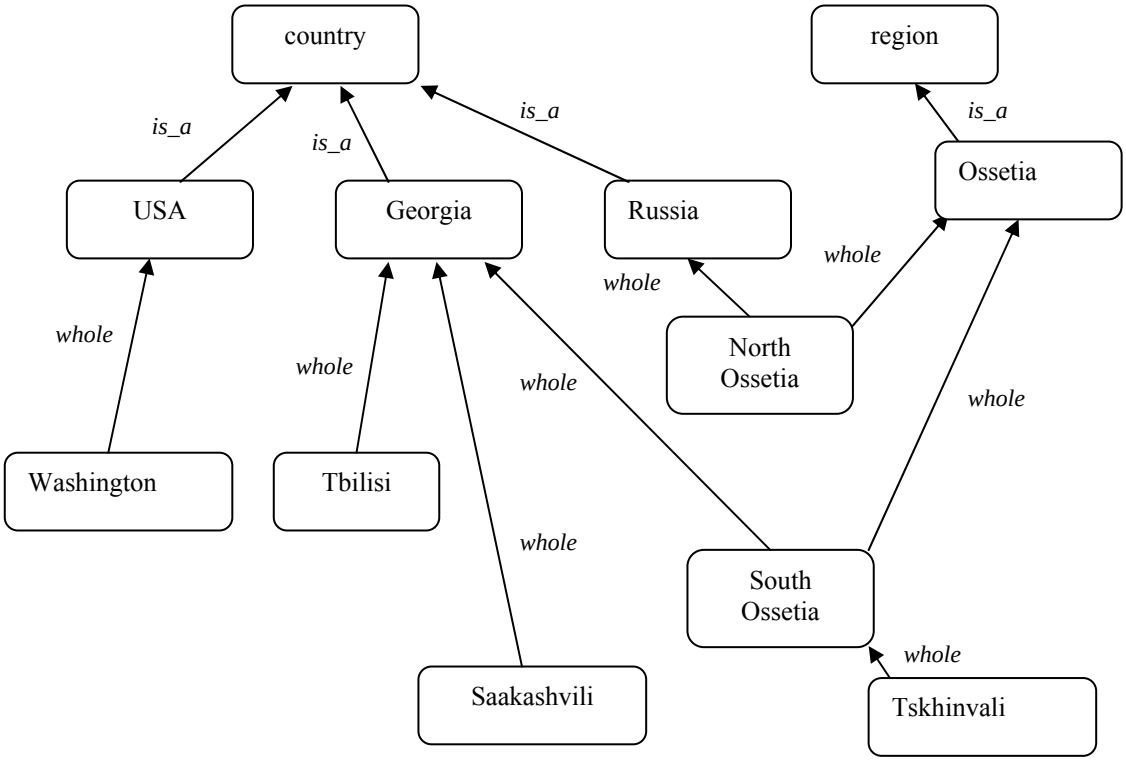

Figure 2: Fragment of the conceptual net for text (*****)

Now we would like to address a fuller version of text (*):

> *(*****) The British Broadcasting Corporation is the first foreign news agency to be granted unrestricted access to the breakaway **Georgian** republica of **South Ossetia** since **Georgian** forces attacked the capital **Tskhinvali**. Journalists working for the BBC have unearthed evidence of **Georgian** war crimes against **South Ossetian** civilians. The indiscriminate use of force is a clear and serious violation of the Geneva Convention and can constitute a war crime.*

*Consequently many people wrongly believe that **Russia** precipitated the conflict by invading **Georgia**. The stubborn fact remains that it was **Georgian** military forces, no doubt following consultation with **American** military "advisers", who bombarded **South Ossetia's** small town capital **Tskhinvali.***

*After consulting with **Washington,** the **Saakashvili** administration in **Tbilisi** attempted to take back **the region** whilst the world's attention was focused upon the opening ceremony of the Beijing Olympics. At approximately 23.30 local time on August 7, the **Georgian** military mounted an exceptionally heavy artillery attack on **Tskhinvali.***

***Moscow** responded rapidly by advancing through the Roki Tunnel, which links **South Ossetia** with **North Ossetia** in **Russia**, crushing **Georgian** forces within a couple of days. Since the outbreak of hostilities, accusations of war crimes have flown between the two sides. **Saakashvili** employed an **American** PR company to amplify his claims of **Russian** atrocities.*

In the text several countries, cities and regions are mentioned. In the conceptual net (see Figure 2) corresponding entities are situated very close to each other. If to use lexical chaining rules described in ([7], see section 2), all highlighted words tend to be included into the same lexical chain. But in fact there are four distinct participants (Russia, South Ossetia, Georgia and USA) interacting with each other and this issue has to be presented using four different lexical chains.

Looking at the text we can see the following co-occurrence in the same clauses of the text:

Countries and Regions	South Ossetia	Georgia (without South Ossetia)	Russia	USA
South Ossetia	-	6	3	1
Georgia (without South Ossetia)	6	-	3	3
Russia	3	3	-	1
USA	1	3	1	-

In Table 1 we presented not initial concepts but manually constructed lexical chains and the picture of interactions is practically evident:

***Georgia**: Georgia, Georgian, Saakashvili, Tbilisi*

***Russia:** Russia, Russian, Moscow, North Ossetia*

***South Ossetia**: South Ossetia, Tskhinvali, Roki Tunnel*

***USA:** American, Washington*

But if to automatically process the text then a conceptual graph has to be generated from the text. Vertices of the graph are not words but concepts because representation of the text content has to be conceptual not lexical. Concepts can be presented in a text as different word forms, derivative words, synonyms.

The graph belongs to multigraphs with two types of edges between vertices. One type of edge, *Rconc*, presents conceptual relations between entities from a thesaurus or an ontology, another type of edge, *Rclauses*, depicts co-occurrences of entities in the same clauses of a text (Figure 3).

Both types of relations and vertices are labeled with frequencies. Labels of vertices are frequencies of entities in the text. Labels of co-occurrence relations are frequencies of corresponding co-occurrences. Labels of conceptual relations are frequencies of co-occurrences within several sentences but not in the same clauses of the text. So conceptual multigraph *MG* of thematic representation can be defined as 6-tuple MG = (*V*, *fv*, *Rconc*, *frconc*, *Rclauses*, *frclause*).

To partition the multigraph to subgraphs that represent different interacting entities in a text we suppose to carry out the following steps:

1. At the first step the most frequent vertex (concept *C0*) is chosen. We begin from the most frequent concepts because they tend to be more important and better reveal their behaviour in the text.

It should be stressed, that the most frequent word of a lexical chain is often considered as a representative of the lexical chain [1, 8].

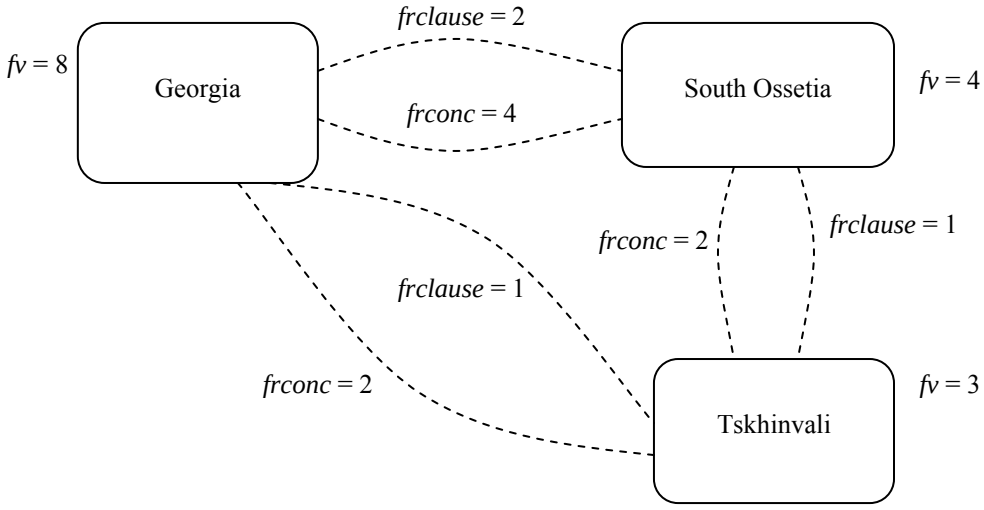

Figure 3: Multigraph with two types of edges

2. For next frequent concept *Ci* we check if it can be joined to an existing subgraph. *Ci* can be included in an existing subgraph with the main vertex *C0* if *Ci* is located on an allowable conceptual path from C0 and the following conditions are fulfilled:

2a) frconc (C0, Ci) > frclause(C0, Ci)
or
2b) frclause(C0, Ci)=1

These conditions provide that relation (C0,Ci) is really used for establishing lexical cohesion in the text. For several concepts with the same frequences the text order can be used.

3. Frequencies of vertices (and corresponding edges) joined to the same subgraph are summed up.

4. A vertex *Ci* can belong to two subgraphs *G0* and *G1* if

4a) Ci is not the main vertex of these subgraphs,
or
4b) Ci may be the main vertex of subgraph G1 if fr(Ci)>frconc(C0,C1), where C0 is main vertex of G0.

In our text (*****) the most frequent concept is concept *GEORGIA* (*fv* = 8). Next concept *is SOUTH OSSETIA* (*fv* = 4). Edges between *GEORGIA* and *SOUTH OSSETIA* have frequencies *frconc* = 2 and *frclause* = 2. So concept *SOUTH OSSETIA* begins its own subgraph.

Next concept is *TSKHINVALI* (*fv* = 3). *TSKHINVALI* vertex can potentially be joined to both subgraphs. But edges between *GEORGIA* and *TSKHINVALI* have frequencies frconc=1 and *frclause* = 2. Edges between *SOUTH OSSETIA* and *TSKHINVALI* have frequencies frconc=2 and frclause=1. So *TSKHINVALI* is joined to *SOUTH OSSETIA* subgraph.

Next concept is *RUSSIA* (*fv* = 3). At this stage there is no allowable path to main concepts of existing graphs. So vertex *RUSSIA* begins a new subgraph. Concept *USA (American – fv* = 2) also begins a new graph.

Next vertex is *SAAKASHVILI* vertex (fv=2). It can be joined to *GEORGIA* subgraph (*frconc* = 2 and *frclause* = 0) and so on.

Thus we obtain four main participants of the situation described in the example text.

In [9] we have already described an algorithm of construction specific types of lexical chains – thematic nodes. But in that approach co-occurrence of concepts was used only for selection of the most important thematic nodes for the text content. We used the thematic representation of documents as a basis for conceptual indexing [10], automatic text categorization [11] and summarization [9]. Now we suppose to utilize co-occurrences of concepts for more correct and flexible lexical chaining.

Conclusion

In this paper we showed that to model lexical cohesion in connected texts it is not enough to find related words in a text. It is important to take into account relations between entities participating in the described situations. The simple way to do this is to take into consideration co-occurrences of words in the same sentences of the text because the more often words co-occur in the same clauses (simple sentences) of a text, the more probable that they relate to different interacting entities.

The factor of co-occurrence allows us to do lexical chaining algorithms more flexible and more corresponding to the discourse structure. The multigraph representation seems to be appropriate as a basis for lexical cohesion modeling.

The factor of conceptual relations between entities should be also considered in lexical chaining experiments with human annotators to decrease its subjectivity.

Acknowledgements

Partial support for this research is provided by the Russian Foundation for Fundamental Research through grant # 09-06-00390-a.

References

[1] Barzilay R. & Elhadad M. (1998). Using Lexical Chains for Text Summarization. - ACL/ EACL Workshop Intelligent Scalable Text Summarization. Madrid.

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

[3] Fillmore, Charles J. (1968). The Case for Case. In: Bach and Harms (Ed.): *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston, 1-88.

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

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

[6] Hirst G. & Morris J. (2005). The subjectivity of Lexical Cohesion in Text. In James C. Chanahan, Yan Qu, and Janyce Wiebe, editors, *Computing attitude and affect in text*. Springer, Dodrecht, The Netherlands. p. 41–48.

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

[8] Hollingsworth W. & Teufel. S. (2005). Human Annotation of Lexical Chains: Coverage and Agreement Measures. In: Workshop proceedings ``ELECTRA: Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications", SIGIR 2005, Salvador, Brazil.

[9] Loukachevitch N. & Dobrov B. (2000). Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems - Machne Translation Review, - N 11, p. 10-20.

[10] Loukachevitch Natalia V. & Dobrov Boris V. (2002). Evaluation of Thesaurus on Sociopolitical Life as Information Retrieval Tool. In: Proceedings of Third International Conference on Language Resources and Evaluation (LREC2002) / M.Gonzalez Rodriguez, C. Paz Suarez Araujo (Eds.) - Vol.1 - 2002, Gran Canaria, Spain - p.115-121.

[11] Loukachevitch N.V. & Dobrov B.V. (2003). Knowledge-Based Text Categorization of Legislative Documents // Proceedings of 7th Conference on Computational Lexicography and Text Research (COMPLEX 2003) / Ed. F.Kiefer, J.Pajzs - Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest. - pp.57-66.

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

[13] Morris J.& Hirst G.(1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of the Text. Computational Linguistics, 17(1): pp. 21-45.

[14] Morris, J., Beghtol, C. & Hirst, G. (2003). Term relationships and their contribution to text semantics and information literacy through lexical cohesion. In: Proceedings 31st Annual Conference of the Canadian Association for Information Science, Halifax, Canada.

[15] Sowa, John (1984). Conceptual Structures^ Information Processing in Mind and Machine. Addison-Wesley.

[16] Stokes N., Hatch P. & Carthy J. (2000). Lexical semantic relatedness and online news event detection. In the proceedings of the Annual 23rd ACM SIGIR Conference on Research and Development (SIGIR-00), pp.324-325.