Loukachevitch Natalia V., Dobrov Boris V.

Center for Information Research
339, Scientific Research Computer Center of Moscow State University
Vorobyevy Gory, Moscow, Russia
fax: 7-095-9382136
louk@mail.cir.ru, dobroff@mail.cir.ru

Thesaurus as a Tool for Automatic Detection of Lexical Cohesion in Texts

Keywords: thesaurus, lexical cohesion, conceptual relations

Abstract

We describe structure of Thesaurus on Sociopolitical Life which was specially created as a linguistic resource for automatic text processing. For several years we intensively used the Thesaurus for various applications of automatic text processing such as conceptual indexing, text categorization and automatic text summarization. Our technology of automatic text processing was based on such a property of a text as cohesion. We considered capability of the automatic process to detect lexical cohesion relations as an important step to revealing text structure. In the paper we will present evaluation of the Thesaurus as a tool for automatic detection of lexical cohesion relations in texts.

Introduction

One of main properties of a connected text is cohesion (Halliday, Hasan, 1976). Cohesion involves relations between words that connect different parts of the text. Lexical cohesion is the most frequent type of cohesion. It can be expressed by repetitions, synonyms and hyponyms or by words connected with other semantic relations such as whole -part, situation participant, object - property and so on.

For example, in a fragment of a text from Text Retrieval Conference text collection (Vorhees, Harman 1995) words *government, cabinet* and noun group *deputy minister* establish one chain of lexical cohesion in the text; noun group *draft law* and word *legislative* of the text form other pair of lexical cohesion relations. Such relations connect sentences of the text without visible markers

> *A package of six **draft laws** for 1995 that the **governmen**t submitted to the State Duma became the subject of discussion in the "Interaction" reform club. **Deputy Minister of Finance** Sergey Aleksashenko, who represents the **cabinet'**s point of view, said that "the main task of tax reform is to reinforce the **legislative** and income base of the budget."*

Cohesion relations connect not only sentences of a text, but also the main theme and sub-themes of a text between each other (van Dejk and Kintsch, 1983). Therefore detection of cohesion relations in texts can become an important step in the way of recognition of the main theme and the discourse structure of a text in automatic text processing.

Lexical cohesion relations different from repetitions are implicit in texts. An author uses them because regards them as a part of knowledge of a reader. Therefore automatic detection of lexical cohesion relations should be based on preliminary lexical and encyclopedic knowledge, described as a linguistic resource in machine-readable form. Usually it is necessary not only to use direct relations indicated in a linguistic resource but also to propose a model of inference of possible cohesive relations using different paths of relations in the linguistic resource. For example, how can a linguistic resource explain a cohesive relation between *government* and *deputy minister,* does it exist because *a minister* is a member of a *government*, and *deputy minister* is a subordinate of a *minister,* or because *government* and *deputy minister* are representatives of executive power, or both?

Morris and Hirst,1991 proposed specification of cohesive relations based on Roget's Thesaurus, but not implemented the corresponding automatic procedure. Hirst and St.Onge 1997; Barzilay and Elhadad,1997 constructed lexical cohesion relations based on WordNet (Miller et.al. 1990). New linguistic resources are developed for different applications of automatic text processing and their capability to be used for automatic detection of lexical cohesion in texts can be considered as an important characteristics of quality of description of concepts and/or words in them.

The paper presents structure and conceptual relations in of the Thesaurus on Sociopolitical Life (Loukachevitch, Salii, Dobrov 1999) and evaluation of the Thesaurus as a tool for automatic detection of lexical cohesion in texts.

The Thesaurus was constructed as a linguistic resource for automatic text processing such as conceptual indexing, text categorization (Loukachevitch, 1997), information filtering and automatic text summarization (Loukachevitch, 1998). It was created using semi-automatic techniques during automatic processing of more than 200 Mb of Russian economical and political documents. This allowed to collect extensive sets of synonymic terms and collocations. Relations between concepts in the Thesaurus were intended for description of semantically and thematically related concepts for each certain concept of the domain.

Now the Thesaurus contains 16.5 thousand concepts, more than 62 thousand relations between concepts, more than 34 thousand terms and proper names. The Thesaurus was intensively tested during semi-automatic procedures and was used in different applications of automatic text processing of Russian texts.

Conceptual Relations in the Thesaurus

The Thesaurus is a coherent hierarchical net of concepts. Synonyms of a concept in the Thesaurus can be nouns, adjectives, multiword noun groups, verbs or verb groups.

Conceptual relations in the Thesaurus serve for solution of three different problems:

- for every concept determination of a set of concepts that can be used in automatic expansion of a query, containing a given concept;

- identification of semantically and thematically related concepts in a text as a basis for recognition of the main theme and subthemes of a text;

- term disambiguation.

There are three basic conceptual relations in the Thesaurus: *hyperonymy-hyponymy* (IS-A) relation, WHOLE-PART relation and ASSOCIATION.

To solve three tasks associated with conceptual relations we determine sets of concepts related to a given concept - such a set is called conceptual neighbourhood. Construction of the conceptual neighbourhood for every concept of the Thesaurus is based on such properties of conceptual relations as transitiveness, inheritance, symmetry.

Usually in conceptual systems transitivity of hyperonymy (ISA) relations is considered. Use of transitivity of WHOLE-PART relation is more difficult since a meronym (PART) can have many holonyms (WHOLE).

To increase a basis of transitivity in the Thesaurus we involved an additional class of transitive relations that we considered as an extended whole-part relation. Besides, we recognized relations with restricted transitivity - such relations are marked with special markers. For example, relations to multiple holonyms are marked with such a marker. There are two markers A (aspectual) for relations, presenting various aspects of a concept, and V (variant) for alternatives.

Properties of relations are as follows:

1. Transitivity of ISA relations (hyperonymy- hyponymy relations).
2. Transitivity of WHOLE-PART relations
3. Restricted transitivity of ISA relations with marks A, V and WHOLE-PART relations with marks A,V:

WHOLE (A/V) + WHOLE = WHOLE (A/V)

but: WHOLE (A/V) + WHOLE (A/V) is not transitive

4. Inheritance of WHOLEs, PARTs and ASSOCIATIONs to subtypes of a concept

5. Restricted inheritance of WHOLEs and ASSOCIATIONs to PARTs of a concept

The whole set of related concepts for a given concept (conceptual neighbourhood) is determined according properties of transitivity and inheritance. If a path between two concepts of the Thesaurus can be reduced to one relation using relation properties then these two concepts are in conceptual neighbourhood of each other. For example, concept *TAX SYSTEM* has 20 direct relations with other concepts of the Thesaurus, but in fact according to the properties of inheritance and transitivity it is related to more than 100 ones.

Experiment on Evaluation of the Thesaurus

In our approach we supposed that all concepts of the Thesaurus related to a given concept can be in cohesive relations with it in texts. Therefore if we determine all Thesaurus related concepts for a given text, we can find a basis of lexical cohesion in this text. If we compare lists of related concepts received from the Thesaurus for a given text and real lexical cohesion relations in this text we can obtain evaluations of quality of Thesaurus descriptions.

The whole text seems to be an intermixture of implicit knowledge and explicit information. To avoid comprehensive analysis of every text and have possibility to study various texts, we decided to test lexical cohesion relations only for the most important concepts of a text corresponding to the main theme of the text. We will call these concepts ‘macroconcepts’. Thus, we could test how information, described in the Thesaurus, supported exposition of the main themes of various texts.

For every text we tried to choose three or four macroconcepts characterizing the main theme of the text in the best way. We chose them mainly from the title, the first paragraph of the text or took the most frequent concepts manually.

At the second stage we fulfilled the following automatic procedure:

- texts were automatically compared with the Thesaurus terms on the basis of morphological analysis. List of Thesaurus concepts found in a text was created;

- terms were disambiguated on the basis of conceptual neighbourhoods of concepts corresponding to different meanings of terms;

- for every chosen macroconcept the list of possibly related concepts from the whole text was created. These lists were created using Thesaurus conceptual relations between concepts of the text and properties of relations. So, for the example text the list of concepts, related to concept *GOVERNMENT,* was as follows: *STATE POWER, MINISTER OF FINANCE, DEPUTY MINISTER.* Term c*abinet* is one of synonyms to concept *GOVERNMENT.*

- during manual reading we tested if every element of the list really served for establishing cohesive relations with the initial macroconcept. In this process we could compute precision and recall of the automatic process of detection of lexical cohesion relations in texts.

In our evaluation we tried to distinguish conceptual relations that were necessary for correct interpretation of a given text from relations that were also true but were not exploited in the text structure. Here there were several cases:

1) a relation was explicitly indicated in a sentence of a text. In this case absence of the corresponding concept in a list of related concepts was not considered as a miss. But if the Thesaurus supported this relation, it was evaluated as a hit.

2) a relation between a concept C and other concept from upper levels of the Thesaurus hierarchy (for example, a hypernym of C) is considered as an extra relation if in a text in all usages this hypernym is not related to C (for example, is used as a reference to other hyponyms different from C).

3) some features or PARTs are inherited by C from upper levels of Thesaurus hierarchy. If in a text they were used in noun compounds with C, we did not considered such relations as misses even if the Thesaurus did no support them. For example, if C is a *CAR*, and in a text word expression *a door of a car* was mentioned, we did not consider the relation “door is a part of a car” as necessary . But if it were supported by the Thesaurus, it would be considered as a hit.

All initial macroconcepts were different, that is, if in a new text there was a macroconcept considered before for another text, we did not test its lexical cohesion again. Also we did not considered lists of related concepts less than 3 elements.

After analysis of 73 lists of related concepts, serving for organizing lexical cohesion relations in 25 texts of sociopolitical domain, our results are as follows: precision -89 %, recall - 71%.

Examples

In the paper we will present an example of an English text from text collection of Text Retrieval Conference with examples of lexical chains, their recall and precision on the basis of English translation of terms of our Thesaurus. We will compare results obtained for this text with results that can be obtained using WordNet.

Also we will describe main types of missing or extra concepts in lists of related concepts.

Conclusions

We described structure of Thesaurus on Sociopolitical Life which was specially created as a linguistic resource for automatic text processing. For several years we intensively used the Thesaurus for various applications of automatic text processing such as conceptual indexing, text categorization and automatic text summarization. We evaluated the Thesaurus as a tool for automatic detection of lexical cohesion relations in texts.

Bibliography

Barzilay R., Elhadad M. 1997. Using Lexical Chains for Text Summarization. - ACL/EACL Workshop Intelligent Scalable Text Summarization.- Madrid.

van Dijk T.A., Kintsch W. 1983. Strategies of Discourse Comprehension. New York. Academic Press, 1983.

Halliday M., Hasan R. 1976. Cohesion in English. Longman, London.

Hirst G., St-Onge D. 1997. 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.

Loukachevitch N.1997. Knowledge Representation for Multilingual Text Categorization // AAAI Symposium on Cross-Language Text and Speech Retrieval, AAAI Technical Report, p. 133-142.

Loukachevitch N. 1998. Text Summarization Based on Thematic Representation of Texts. In Procedings of the *AAAI'98 Spring Symposium on Intelligent Text Summarization*.

Loukachevitch Natalia V., Salii Alla D., Dobrov Boris V.1999. Thesaurus for Automatic Indexing: Structure, Developement, Use. In Proceedings of International Congress “Terminology and Knowledge Engineering”. p. 343-355.

Miller G., Beckwith R., Fellbaum C., Gross D. and Miller K. 1990. Five papers on WordNet. CSL Report 43. Cognitive Science Laboratory, Princeton University.

Morris J., Hirst G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of a text. Computational Linguistics,17 (1), 21-48.

Voorhees E., Harman D.1997. Overview of the Fifth Text REtrieval Conference (TREC-5), in Information Technology: The Fifth Text REtrieval Conference (TREC-5), NIST SP 500-238, National Institute of Standards and Technology. - pp. 1-28.

Winston M. Chaffin R. Herman D. (1987): A Taxonomy of Part-Whole Relations. - Cognitive Science 11, 417-444.