WordNet for Lexical Cohesion Analysis

Elke Teich[1] and Peter Fankhauser[2]

[1] Department of English Linguistics
Darmstadt University of Technology, Germany
Email: E.Teich@mx.uni-saarland.de
[2] Fraunhofer IPSI, Darmstadt, Germany
Email: fankhaus@ipsi.fraunhofer.de

Abstract. This paper describes an approach to the analysis of lexical cohesion using WordNet. The approach automatically annotates texts with potential cohesive ties, and supports various thesaurus based and text based search facilities as well as different views on the annotated texts. The purpose is to be able to investigate large amounts of text in order to get a clearer idea to what extent semantic relations are actually used to make texts lexically cohesive and which patterns of lexical cohesion can be detected.

1 Introduction

Using a thesaurus to annotate text with lexical cohesive ties is not a new idea. The original proposal is due to [1], who manually annotated a set of sample texts employing Roget's Thesaurus. With the development of WordNet [2,3], several proposals for automizing this process have been made. For the purpose of detecting central text chunks which can be used for summarization (e.g. [4,5]), this seems to work reasonably well. But how well does an automatic process perform in terms of linguistic-descriptive accuracy? It is well known from the linguistic literature that any two words between which there exists a semantic relation may or may not attract each other so as to form a cohesive tie (cf. [6]). So when do we interpret a semantic relation between two or more words instantiated in text as cohesive or not? First, certain parts-of-speech may be more likely to contract lexical cohesive ties than others, e.g., nouns may be more likely to participate in substantive cohesive ties than verbs. Another motivation may be the type of vocabulary: special purpose vocabulary may be more likely to contract cohesive ties than general vocabulary. Another possible hypothesis is that cohesive patterns differ due to the type of text (register, genre). While repetition generally appears to be the dominant means to establish lexical cohesion, the relative frequency of more complex relations, such as hyponymy or meronymy may depend on the type of text.

In order to investigate such issues, large amounts of data annotated for lexical cohesion are needed. Manual analyses are very time-consuming and may not reach a satisfactory intersubjective agreement. Completely automatic analysis may introduce significant noise due to ambiguity [4]. Thus, in this paper we follow an approach in the middle ground. We use the sense-tagged version of the Brown Corpus, where nouns, verbs, adjectives, and adverbs are manually disambiguated w.r.t. WordNet, and use WordNet to annotate the corpus with potential lexical cohesive ties. (Section 2.1). We also describe facilities for filtering candidate ties and for generating different views on the annotated text. (Section 2.2). We discuss the results on an exemplary basis, comparing the automatic annotation with a manual annotation (Section 3). We conclude with a summary and issues for future work (Section 4).

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 326–331.

2 Lexical Cohesion Using WordNet

Lexical cohesion is commonly viewed as the central device for making texts hang together experientially, defining the aboutness of a text (field of discourse) (cf. [6, chapter 6]. Along with reference, ellipsis/substitution and conjunctive relations, lexical cohesion is said to formally realize the semantic coherence of texts, where lexical cohesion typically makes the most substantive contribution (according to [7], around fifty percent of a text's cohesive ties are lexical).

The simplest type of lexical cohesion is *repetition*, either simple string repetition or repetition by means of inflectional and derivational variants of the word contracting a cohesive tie. The more complex types of lexical cohesion rely on the systemic semantic relations between words, which are organized in terms of *sense relations* (cf. [6, 278–282]). Any occurrence of repetition or of relatedness by sense relation can potentially form a cohesive tie.

2.1 Determining Potential Cohesive Ties

Most of the standard sense relations are provided by WordNet, thus it can form a suitable basis for automatic analysis of lexical cohesion. As the corpus, we use the Semantic Concordance Version of the Brown Corpus, which comprises 352 texts (out of 500)[3] Each text is segmented into paragraphs, sentences, and words, which are lemmatized and part-of-speech (*pos*) tagged. For 185 texts, nouns, verbs, adjectives, and adverbs are in addition sense-tagged with respect to WordNet 1.6, i.e., with few exceptions, they can be unambiguously mapped to a synset in WordNet. For the other 167 texts, only verbs are sense-tagged.

Using these mappings, we determine potential cohesive ties as follows. For every sense-tagged word we compute its semantic neighborhood in WordNet, and for each synset in the semantic neighborhood we determine the first subsequent word that maps to the synset.

For the semantic neighborhood we take into account and distinguish between most of the available kinds of relations in WordNet: *synonyms*, *hyponyms*, *hypernyms*, *cohypernyms*, *cohyponyms*, *meronyms*, *holonyms*, *comeronyms*, *coholonyms*, *antonyms*, the *pos* specific relations *alsoSee*, *similarTo* for adjectives, *entails* and *causes* for verbs, and the (rather scarce) relations across parts-of-speech *attribute*, *participleOf*, and *pertainym*. Where appropriate, the relations are defined transitively, for example, hypernyms comprise all direct and indirect hypernyms, and meronyms comprise all direct, indirect, and inherited meronyms. In addition, we also take into account *lexical repetition* (same *pos* and lemma, but not necessarily same synset), and *proper noun repetition*.

For each potential cohesive tie we determine in addition the number of intervening sentences, the distance of the participating words from a root in the WordNet hypernymy hierarchy, as a very rough measure of their specificity, and the length and branching factor of the underlying semantic relation, as very rough measures of its strength.

[3] For the purpose of using off-the-shelf XML processing technology (XPath, XSLT, and an XML database), we have transformed the available SGML version of the corpus to XML; likewise we have transformed the WordNet 1.6 format to XML. Moreover, because some of the texts are compiled from multiple sources, we have enriched them with the bibliographic and segmenting information available from the ICAME version of the corpus. For details see [8].

Due to the excessive computation of transitive closures for the semantic neighborhood of each (distinct) word, this initial step is fairly demanding computationally. In the current implementation (which can be optimized), processing a text with about two thousand words takes about 15 minutes on an average PC. Thus we perform this annotation offline.

2.2 Constraints and Views on Lexical Cohesion

The automatically determined ties typically do not all contribute to lexical cohesion. Which ties are considered cohesive depends, among other things, on the type of text and on the purpose of the cohesion analysis. To facilitate a manual post analysis of the ties we can filter them by simple constraints on *pos*, the specificity of the participating synsets, the kind, distance, and branching factor of the underlying relation, and the number of intervening sentences. This allows, for example, to exclude very generic verbs such as "be", and to focus the analysis on particular relations, such as lexical repetition without synonymy (rare), or hypernyms only.

The remaining ties are combined to lexical chains in two passes. In the first pass, all transitively related (forward-)ties are combined. The resulting chains are not necessarily disjoint, because there may be words w_1, w_2, w_3, where w_1 and w_2 are tied to w_3, but w_1 is not tied to w_2. This results in a fairly complex data structure, which is difficult to reason about and to visualize. Thus in a second pass, all chains that share at least one word are combined into a single chain. The resulting chains are disjoint w.r.t. words, and may optionally be further combined if they meet in a specified number of sentences.

To further analyze lexical cohesion, we have realized three views. In the *text view*, each lexical chain is highlighted with an individual color, in such a way that colors of chains starting in succession are close. This view can give a quick grasp on the overall topic flow in the text to the extent it is represented by lexical cohesion. The *chain view* presents chains as a table with one row for each sentence, and a column for each chain ordered by the number of its words. This view also reflects the topical organization fairly well by grouping the dominant chains closely. Finally, the *tie view* displays for each word all its (direct) cohesive ties together with their properties (kind, distance, etc.). This view is mainly useful for checking the automatically determined ties in detail. In addition, all views provide hyperlinks to the thesaurus for each word in a chain to explore its semantic context. Moreover, some statistics, such as the number of sentence linking to and linked from a sentence, and the relative percentage of ties contributing to a chain are given.

Because filtering ties and combining them to chains can essentially be performed in two (linear) passes over the text, and the chains are rather small (between 2 and 200 words), producing these views takes about 2 seconds for the texts at hand, and thus can be performed online.

3 Discussion of Results

This section discusses the results of the automatic analysis on a sample basis, comparing the automatic analysis with a manual analysis of the first 20 sentences of three texts from the "learned" section of the Brown corpus (texts j32, j33 and j34). For the automatic analysis only nouns and verbs which are at least three relations away from a hypernym root, and adjectives

which are tied to a noun or a verb have been included. Following [9] for the manual analysis, whenever a choice had to be made on which type of relation to base the establishment of a tie, priority was given to repetition.

Table 1 shows the results for the automatic analysis, Table 2 gives the results for the manual analysis (strongest chains only). The chains are represented by the anchor word for simple repetition, and a subset of the participating words for the other types of relations. The number of words and the number of sentences are given in parentheses.

Table 1. Major lexical chains – automatic analysis

j32	j33	j34
form/stem/word (18;11) information/list/ spelling (17;13) dictionary/entry (17;11) text (11;9) store (2;2) storage (2;2)	sentence/subject/ word/... (26;14) stress (14;11)	tone/tonal (8;5) tone system/ consonant system (7;5) linguist (5;5) linguistics (2;2) field (6;6)

Table 2. Major lexical chains – manual analysis

j32	j33	j34
dictionary (16;11) form (14;8) information (11;9) text (11;9) store/storage (4;4)	stress (14;11) complement/predicator/ subject (13;9) sentence (4;4)	linguist/linguistics (13;7) tone/tonal (11;5) field (6,6)

As can be seen from the tables, there is a basic agreement between the automatic and the manual analysis in terms of the strongest chains (e.g., in j33, *stress* builds one of the major chains in both analyses). However, some ties are missed in the automatic analysis, e.g., *store/storage* in j32 or *linguist/linguistics* in j34. Also, there are some differences in the internal make-up of the established chains. For example, in j32, *dictionary* and *information* build major chains in both analyses, but the *information* chain includes a few questionable ties in the automatic analysis, e.g., *list* as an indirect hyponym. Also, the chain around *form* includes *word* and *stem* in the automatic analysis, which would be fine, but then words like *prefix*, *suffix*, *ending* at later stages of the text are not included. The chain built around *complement/predicator/subject* in j33 is separate from the *sentence* chain in the manual analysis, but in the automatic analysis the two are arranged in one chain due to meronymy, thus resulting in the strongest chain for this text in the automatic analysis.

The mismatches arise due to the following reasons. (1) *Missing relations*. Only some relations across parts-of-speech and derivational relations are accounted for in WordNet,

e.g., *linguistic/linguistics* but not *linguist/linguistics* or *store/storage*[4]. (2) *Spurious relations*. Without constraints on the length and/or branching factor of a transitive relation rather questionable ties are established, e.g. *alphabetic character* as a rather remote member of *list*. (3) *Sense proliferation*. In some instances the sense-tagging appears to be overly specific, e.g., in j34, *explanation* as 'a statement that explains' vs. 'a thought that makes sth. comprehensible'. Using synonymy without repetition these senses do not form a tie. On the other hand, with repetition, some questionable ties are established, e.g., for *linguistic* as 'linguistic' vs. 'lingual'. (4) *Compound terms*. In some instances the manual analysis did not agree with the automatic analysis w.r.t. compound terms. E.g. *tonal language* is sense-tagged as a compound term and thus not included in the chain around *tone/tonal*.

Generally, the unconstrained automatic annotation is too greedy, i.e., too many relations are interpreted as ties. Unsatisfactory precision is not so much of a problem, however, because the annotation can be made more restrictive, e.g., by including a list of stop words and/or by determining the appropriate maximal branching and maximal distance for each text to be analyzed. More serious is the problem of not getting all the relevant links, i.e., unsatisfactory recall, usually due to missing relations in the thesaurus. To a certain extent this can be overcome by combining chains that meet in some minimal number of sentences, as a very specific form of collocation.

4 Summary and Envoi

We have presented a method of analyzing lexical cohesion automatically. Even if the results of the automatic analysis do not match one-to-one with a manual cohesion analysis, the automatic analysis is not that far off. Some problems are inherent, others can be remedied (cf. Section 3). Even with an imperfect analysis result, we get a valuable source for the linguistic investigation of lexical cohesion. Knowing that not all words that are semantically related contract cohesive ties, we can set out to determine factors that constrain the deployment of sense relations for achieving cohesion comparing the automatic annotation with a manual analysis. Also, we can give tentative answers to the questions posed in Section 1. Looking at part-of-speech, we can confirm that the strongest chains are established along nouns, and the strongest chains are established along the special purpose vocabulary rather than the general vocabulary[5]. Moreover, although repetition (and synonymy) is the most-used cohesive device, the frequency of other relations taken together (in particular hyponymy and meronymy) about matches that of repetition for the texts at hand.

Future linguistic investigations will be dedicated to questions of this kind on a more principled basis. In order to get a more precise idea of the reliability of the automatic analysis, we are carrying out manual analyses on a principled selection of texts from the corpus (e.g., larger samples covering all registers in the corpus) and compare the results with those of the automatic analysis. Moreover, we plan to investigate cohesion patterns, such as the relative frequency of repetition vs. other types of relations, for the different registers. Ultimately, what we are after are cross-linguistic comparisons, including the comparison of translations with original texts in the same language as the target language [10].

[4] The version we have worked with is WordNet 1.6. WordNet 2.0 has been extended so as to handle such relations more comprehensively.

[5] This also holds when less specific words are taken into account by the automatic analysis.

References

1. Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics **17** (1991) 21–48.
2. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J.: Introduction to WordNet: An on-line lexical database. Journal of Lexicography **3** (1990) 235–244.
3. Fellbaum, C., ed.: WordNet: An electronic lexical database. MIT Press, Cambridge (1998).
4. Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. In: Proceedings of ISTS 97, ACL, Madrid, Spain (1997).
5. Silber, H. G., McCoy, K. F.: Efficient text summarization using lexical chains. In: Proceedings of Intelligent User Interfaces 2000. (2000).
6. Halliday, M., Hasan, R.: Cohesion in English. Longman, London (1976).
7. Hasan, R.: Coherence and cohesive harmony. In Flood, J., ed.: Understanding Reading Comprehension. International Reading Association, Delaware (1984) 181–219.
8. Fankhauser, P., Klement, T.: XML for data warehousing – chances and challenges. In: Proceedings of DaWaK 03, LNCS 2737, Prague, CR, Springer (2003) 1–3.
9. Hoey, M.: Patterns of lexis in text. Oxford University Press, Oxford (1991).
10. Teich, E.: Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts. de Gruyter, Berlin and New York (2003).