The Subjectivity of Lexical Cohesion in Text

Jane Morris
Faculty of Information Studies
University of Toronto
Toronto, Ontario, Canada M5S 3G6
morris@fis.utoronto.ca

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada M5S 3G4
gh@cs.toronto.edu

Abstract

A reader's perception of even an "objective" text is to some degree subjective. We present the results of a pilot study in which we looked at the degree of subjectivity in readers' perceptions of lexical semantic relations, which are the building blocks of the lexical chains used in many applications in natural language processing. An example is presented in which the subjectivity reflects the reader's attitude.

Introduction

How much of a reader's understanding of a text is idiosyncratic and how much is common to that of most other readers of the same text of a similar age and education? What is the degree of individual difference or subjectivity in text understanding? The answers to these questions are likely to vary with text type. In this paper, the focus will be on general-interest articles (from *Reader's Digest*), and on readers' perceptions and interpretations of lexical cohesive relations in the text. Perceptions of these relations contribute to a reader's perception of the structure of the text.

There are two fundamentally different approaches to text structure: Some methods, such as Rhetorical Structure Theory (Mann and Thompson, 1988), aim to identify predefined structures in a text. Other methods are *associationist*; they focus on building up text-specific structures, for example through the creation of ad-hoc categories such as those proposed by Barsalou (1989) or groups of related words within the text such as *lexical chains* (Halliday and Hasan, 1976; Morris and Hirst, 1991). There is much to be gained by accepting the contributions of each approach, and in discovering how they interact. In a sense, the work of Morris and Hirst attempted this by relating associationist lexical chains to the predefined intentional structure of discourse that was proposed by Grosz and Sidner (1986). However, that particular model of discourse structure was itself rather associationist in that the "intentional structure" of a text is quite ad hoc and text-specific.

The present work is an examination of the degree of subjectivity of two aspects of the *lexical cohesion* (Halliday and Hasan, 1976) perceived by readers of text: the word groups (lexical chains) that are formed and the *lexical semantic relations* that are perceived between the words. We know of no prior research on readers' perceptions of lexical cohesion or the associated lexical semantic relations in text. Furthermore, most of the research on lexical semantic relations has not been done in the context of text. Instead, most researchers have just looked at word pairs and the four "*classical*" lexical relations: synonymy, antonymy, hyponymy, and meronymy (Fellbaum, 1998; Cruse, 1986; Halliday and Hasan, 1989). The classical relations themselves form predetermined structures consisting of hierarchies that have been studied and widely applied since Aristotle. The *non-classical* relations (all of the rest) have tended to remain unnamed and unstructured, as in the relations implicit in *Roget's Thesaurus*, in the "associative" relations or Related Terms used in Library and Information Science (Neelameghan, 2001; Milstead, 2001), in the "associative" relations widely assumed in psychology, and in the relations between members of Lakoff's (1987) non-classical categories.

Consider, for example, this (constructed) text: "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." In what ways do readers see the relations in this text? One reader reports two lexical chains or word groups: 'understanding', which contains the words *figure out*, *means*, *meaning*, *mind*, *think*, *meaning*, and 'text', which contains the words *text*, *reader*, *text*. In the 'understanding' word group, related word pairs and the non-classical relations that this reader reports are these: *figure out*, *means*: *means* is the likely result of the action *figure out*; *mind*, *figure out*; *mind* is where the *figure out* action happens; *think*, *meaning*: *meaning* is a result of the action of *think*ing. The reader's description of the word group is 'words to do with human understanding'.[1]

We have carried out a study of the degree of subjectivity of the word groups and lexical semantic relations perceived by readers of a text. The results will be presented below.

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

[1] These are the chains and relations that were reported by reader 'JM'. Another reader, 'GH', also reported two chains, but grouped *means* and *meaning* with *text* and *reader*.

Theoretical background

The linguistic study of the contribution made by inter-sentence groups of related words to text understanding started with the concept of lexical cohesion (Halliday and Hasan, 1976) and has been extended by Hasan (1984; Halliday and Hasan, 1989) to include the concept of *cohesive harmony*. Cohesive harmony adds lexico-grammatical structure to word groups (lexical chains) by first dividing them into two types — *identify-of-reference chains*, which combine reference and lexical cohesion, and *similarity chains* (using only classical relations) — and then by linking these chains together into a more tightly-knit unit with grammatical intra-sentence relations similar to the case relations of Fillmore (1968), such as agent–verb and verb–object. Cruse (1986) briefly discusses a related concept of "patterns of lexical affinities", where similar intra-sentence patterns called "syntagmatic affinities" can create more-general inter-sentence patterns (relations) called "paradigmatic affinities". Cohesive harmony and the concept of patterns of lexical affinities make the important contribution of linking lexical (and grammatical, in the case of reference cohesion) inter-sentence cohesion with grammatical intra-sentence cohesion. But no analysis of these concepts has been done using readers of text. It is therefore not known how subjective the process is.

Lexical semantic relations are the building blocks of lexical cohesion, cohesive harmony, and the concept of patterns of lexical affinity. The original view of them by Halliday and Hasan (1976) was very broad and general; the only criterion was that there had to be a recognizable relation between two words. Many of these relations were found in *Roget's Thesaurus* by Morris and Hirst (1991) in an application of the theory. The more-recent view of Hasan (1984; Halliday and Hasan, 1989) is to only use classical relations, since the rest are "too intersubjective", and both Hasan and Cruse (1986) indicate that they focus on classical relations because of prior historical focus. In psychology, the focus has been mostly on classical relations; however, there have been recent calls to broaden the focus and include non-classical relations as well (McRae and Boisvert, 1998; Hodgson, 1991). Some researchers have always included some non-classical relations, such as Evens et al. (1983), Chaffin and Herrmann (1984), and researchers in Library and Information Science. However, as stated earlier, the research on lexical semantic relations has been done out of the context of text, and then assumed to be relevant within it, and in lexical cohesion research, the analysis of lexical semantic relations was done by experienced linguists with particular points of view.

Experimental study

We are interested in analyzing readers' perceptions and interpretations of the lexical cohesion in text for individual differences. To this end, a pilot study was conducted with five participants as readers of the first 1.5 pages of a general-interest article from the *Reader's Digest* on the topic of movie actors and movie characters as possibly inappropriate role models for children.

Subjects were instructed to first read the article and mark the word groups that they perceived, using a different color of pencil for each different group. Once this task was completed, they transferred each separate word group to a new data sheet, and then, for each word group, indicated which pairs of words they perceived as related and what the relation was. Finally, they described the meaning of each word group in the text.

This data was analyzed to determine the degree of individual differences in the responses. For each of these groups, we computed the subjects' agreement on membership of the group in following manner: We took all possible pairs of subjects, and for each pair computed the number of words on which they agreed as a percentage of the total number of words they used. Averaged over all possible pairs of subjects, the agreement was 63%. Next, we looked at agreement on the word pairs that were identified as directly related (within the groups that were identified by a majority of subjects). We restricted this analysis to *core* words, which we defined to be those marked by a majority of subjects. We counted all distinct instances of word pairs that were marked by at least 50% of the subjects, and divided this by the total number of distinct word pairs marked. We found that 13% of the word pairs were marked by at least 50% of the subjects. For the set of word pairs used by at least two subjects, we then computed agreement on what the relation between the pair was deemed to be. We found that the subjects agreed in 70% of the cases.

Table 1: Word group similarity among readers: Average agreement between pairs of readers.

Gloss of word group	Average pairwise agreement (%)
Movies	71
Communications[a]	69
Smoking	73
Groups and causes	63
Bad behaviors	41

[a]Only 3 subjects used this group.

Table 1 summarizes the results for the major word groups found in the text by the readers. Individual differences showed up as different non-core words within a group, or as a different focus for the same group. As an example of the latter case, one reader added idiosyncratic attitude-bearing choices to the 'bad behaviors' word group, reflecting a "law-and-order" focus on bad behaviors. This is shown in Table 2, where the readers largely agree on the core words of the group, but one reader adds a group of seven "law-and-order" words that no other reader includes. (The number of readers who used each word is shown in the left column of the table.)

Table 1 shows a "trend" of 60–70% agreement (average of 63%) on word groups (though the sample of five readers and one text is small). The outlier group of 'bad behaviors' was much lower at 41% and seems to reflect the fact that judgment of bad behavior is an inherently value-laden human endeavor. For example, two out of five readers included *witchcraft*, two out of five did not include smoking-related

Table 2: 'Bad behaviors' word group: an example of subjectivity reflecting reader attitude.

Core words (chosen by ≥ 3 readers)	
5	shooting
4	sex
4	drinking
4	dangerous
3	drag racing
3	irresponsible [behaviors]

"Law / order / authority" outliers (all chosen only by 1 reader)	
1	police
1	caught
1	British Intelligence Service
1	gun control lobby
1	Department of Role Model Development
1	M.A.D.D. [Mothers Against Drunk Driving]
1	spies

Table 3: Lexical semantic relation similarity among readers: Average agreement on related word pairs and on the nature of the relation in agreed-on word pairs.

Gloss of word group	Word pairs agreed on (%)	Relation agreement (%)
Movies	10	75
Communications[a]	12	20
Smoking	13	85
Groups and causes	18	69
Bad behaviors	12	100

[a]Only 3 subjects used this group.

words, and, as noted, one reader included a law-and-order focus while the other four did not.

Agreement on which word pairs within a group are related is much lower at around 13% (Table 3). This could be a reflection of the following two factors:

- This is a much more indirect task than identifying word groups. It is also cumbersome (as reported by some subjects) in that the potential number of pairs of related words is large. They were asked to be exhaustive (*i.e.,* give all word pairs that they perceived as related), but complained and were not. In contrast to forming word groups, this process was not intuitive for the readers.
- The word groups might be comprehended as gestalts or wholes, and words entering the category or group are, in some way, all related. That is, the relations are not perceived as binary, but holistically. In fact many of the relation descriptions were context specific. For example, one subject said that the relation in the word pair *sex–smoking* is that "both are undesirable activities for kids in the article".

In cases where subjects identified identical word pairs as related, they also showed a marked tendency (at an average of 70% agreement) to agree on what the relation was. In fact, they showed a notable ability and ease at being able to explain how words are related in context. This contrasts sharply with the commonly known fact, noted by Cruse (1986), that people find words hard to define out of context. This high level of reader agreement on what the relations were is a reflection of the importance of considering aspects of text understanding such as lexical semantic relations as being situated within their surrounding context. In other words, while explaining or perceiving linguistic meaning out of context is hard, doing so within text seems here not to be, and is therefore likely a rich and meaningful area for further study.

Discussion

The subjects in this small study identified a common "core" of groups of related words in the text, as well as exhibiting subjectivity or individual differences. It might be objected that our subjects simply showed "a low kappa" (or "a *bad* kappa"), and all this shows is that we asked them "the wrong question". We disagree. Rather, we believe that these preliminary results indicate that lexical cohesion is useful both as a theory and as a practical tool for determining both the commonly agreed on and the subjective aspects of text understanding. In fact, the kappa statistic doesn't apply here, as the words in a word group are not independent, and so agreement by chance cannot be computed.

Our work here does not investigate cases where the author of a text either implicitly or explicitly marks the text as being a subjective point of view taken by a particular person. Rather, we focus on the overall subjectivity in readers' perceptions of a text's meaning (i.e., aspects that are inherent in the word groups and lexical semantic relations). We consider this subjectivity to be a crucial aspect of text understanding in that it builds on research that views meaning as something created by the reader or processor of text, as opposed to meaning as something that somehow exists in text alone, separate from the reader/processor (Olson, 1994). For automation purposes it will be useful to have a clear understanding of what aspects of text meaning do exist "in the text", and what aspects can be expected to contribute to individual differences in comprehension.

Our next step will be the larger study for which this was a pilot; we will use three different texts and at least ten readers per text. We will look for overall patterns in the types of words and relations that form part of the core group and those that do not. We intend to focus on aspects of word pairs and relations such as whether they are classical or non-classical and text-general or text-specific. We also intend to analyze the relations to determine whether a common set of relation types is being used by readers. These non-classical relation types could be used to augment future or existing lexical resources.

An obvious area for future research is the effect of different types of texts and readers. We are interested in how text-specific the word groups and relations are, since non–text-specific information can be added to existing resources, but text-specific knowledge will require further complex interaction with the rest of the text. We also intend to inves-

tigate the potential linkages between the word groups in the texts for evidence of cohesive harmony or any other relations to other theories of pre-determined mechanisms of text understanding.

Acknowledgment This research was supported by the Natural Sciences and Engineering Research Council of Canada. We are grateful to Clare Beghtol for helpful discussions.

References

Barsalou, L. (1989). Intra-concept similarity and its implications for inter-concept similarity. In S. Vosniadou and A. Ortony (eds.), *Similarity and analogical reasoning* (pp. 76–121). Cambridge: Cambridge University Press.

Chaffin, R., and Herrmann, D. (1984). The similarity and diversity of semantic relations. *Memory and Cognition*, 12(2), 134–141.

Cruse, D. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.

Evens, M., Markowitz, J., Smith, R., and Werner, O. (eds.). (1983). *Lexical semantic relations: A comparative survey*. Edmonton, Alberta: Linguistic Research Inc.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, Mass.: The MIT Press.

Fillmore, C. (1968). The Case for Case. In E. Bach and R. Harms (eds.), *Universals in linguistic theory* (pp. 1–88). New York: Holt, Rinehart and Winston.

Grosz, Barbara J. and Sidner, Candace L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175–204.

Halliday, M.A.K., and Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M.A.K., and Hasan, R. (1989). *Language, context, and text: aspects of language in a social-semiotic perspective*. (2nd ed.). Oxford: Oxford University Press.

Hasan, R. (1984). Coherence and Cohesive Harmony. In J. Flood (ed.), *Understanding reading comprehension: Cognition, language and the structure of prose* (pp. 181–219). Newark, Delaware: International Reading Association.

Hodgson, J. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes*, 6(3), 169–205.

Lakoff, G. (1987). *Women, fire and dangerous things.* Chicago: University of Chicago Press.

Mann, William C. and Thomson, Sandra A. (1988). Rhetorical structure theory. *Text*, 8, 243–281.

McRae, K., and Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24(3), 558–572.

Milstead, J.L. (2001). Standards for relationships between subject indexing terms. In C.A. Bean and R. Green (eds.). *Relationships in the organization of knowledge* (pp. 53–66). Kluwer Academic Publishers.

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

Neelameghan, A. (2001). Lateral relationships in multicultural, multilingual databases in the spiritual and religious domains: The OM Information Service. In C. Bean and R. Green (eds.), *Relationships in the organization of knowledge* (pp. 185–198). Norwell, Mass.: Kluwer Academic Publishers.

Olson, David (1994). *The World on Paper*. Cambridge University Press.

Roget, Peter Mark. *Roget's International Thesaurus*. Many editions and publishers.