Readers' Perceptions versus Computers' Interpretations of Text Meaning: the example of Lexical Cohesion

Jane Morris
Faculty of Information Studies, University of Toronto
Toronto, ON, Canada, M5S 3G6
Email: jane.morris@utoronto.ca

Abstract

The results of an empirical study of readers' perceptions of lexical cohesion in text are presented, showing approximately 40% subjectivity in their responses as measured by individual differences. Current computational applications of lexical cohesion such as summarization and information retrieval only process aspects of text meaning that are "in the text". This study highlights the potential importance of aspects of meaning that are "in the reader", and suggests the possibility of reader models to account for readers' subjective interpretations. No prior research on the subjectivity of lexical cohesion in text exists. This study indicates the importance of further research on the subjectivity of readers' perceptions of lexical cohesion, as well as further research on the subjectivity of other aspects of text meaning. Knowledge of subjectivity will become increasingly important as computers are used to provide increasingly complex interpretations of text. Current computational systems that use statistical and machine learning techniques to analyze large corpora ignore reader subjectivity and focus almost exclusively on aspects of text meaning that exist "in the text".

Introduction

In linguistics, lexical cohesion is used to explain one aspect of how a text's meaning is created, through "continuity of lexical meaning" (Halliday & Hasan, 1976). Lexical cohesion is the contribution to a text's meaning made by groups of related words found within it. Computationally, it has been used in determining the structure of text (Morris & Hirst, 1991), summarization (Barzilay & Elhadad, 1999; Silber & McCoy, 2002), spelling correction (Hirst & Budanitsky, 2004), and information retrieval (Voorhees, 1998). However, there is no research on how subjective readers' perceptions of lexical cohesion in text are. This research reports on results of an experimental study of readers' perceptions of lexical cohesion. The study was set up to answer the following research question:

> When readers (participants) mark the lexical cohesion that they perceive while reading a text, how much do their responses differ from one another, and how much are they similar?

The answer will give an indication of the subjectivity of readers' perceptions of lexical cohesion, as measured by individual differences in their responses.

This research is situated within the context of the larger question of how subjective text interpretation is in general, and was motivated in part by trends in computational linguistics over the years. Subjectivity (individual difference) is viewed here as an important part of research on what aspects of meaning are "in the text" versus what aspects are "in the reader" (and "in the writer"). Recent work that focuses on automatic corpus analysis in linguistics and computational linguistics is, by definition, only able to analyze or discover what is "in the text", although in some cases human-annotated corpora are used as starting points and results are judged by humans. A well-known earlier computational approach to text understanding (Grosz & Sidner, 1986) emphasized aspects of meaning found "in the text", in what was referred to as the linguistic structure, aspects found "in the writer", in what was called the intentional structure, and aspects found "in the reader", in the attentional structure which modeled readers' focus. Clearly their theory attempts to account for more than aspects of meaning that are "in the text".

In part because of difficulty understanding and analyzing writers' intentions and readers' subjective interpretations, as well as because of advances in statistical and machine learning techniques for corpus analysis, computational linguistics research has shifted away from including the reader (or writer) as being involved in text understanding/interpretation. However, as aptly stated by Olson (2004):

> This I now believe, is one of the fateful illusions of modernism, the idea that knowledge can be embodied in a text or a computer program or other artifact. Texts are more accurately seen as artifacts, notational devices for representation and thought. Knowledge remains the possession of the knower not of the artifact.

The view taken in this study is that knowledge exists in both the text and the reader, and that we need more research on the contributions of each to a text's meaning. The writer is also considered to be an important factor in the interpretation of a text's meaning, but is not considered further here.

Theoretical Background

The linguistic theory of lexical cohesion was presented by Halliday & Hasan in their 1976 classic *Cohesion in English*. Lexical cohesion is one of five types of cohesion detailed therein, and as stated earlier, its contribution to the meaning of text is provided by continuity of lexical meaning created by different groups of related words that run through a text.

Here is an example of lexical cohesion showing four word groups identified by a reader from the first paragraph of the first 1.5 pages of a *Reader's Digest* article referred to as Text 1 from the experimental study:

> I attended a *funeral service* recently. **Kind** words, *Communion*, *chapel* overflowing, speeches by lawyers, government workers, friends, all speaking of the *deceased's* **kindness**, his **brilliance** in mathematics, his **love** of SCRABBLE and CHESS, his great **humility** and **compassion**, his **sense of humour**.

Each different word group is shown with a different emphasis. The italicized word group is about funerals, the bold word group about positive human characteristics, the underlined group about types of people/jobs, and the capitalized group about board games. These inter-sentence word groups have no size or location restrictions and can therefore span any portion of the text.

Hasan extended the theory in 1984 to include reference cohesion, but the most significant addition was to combine inter-sentence word relations already a part of the original theory with intra-sentence word relations similar to those of Fillmore's (1968) case relations. The result consists of structured and more tightly knit units of lexical cohesion within text. Related ideas were found in Cruse's (1986) concept of patterns of lexical affinities, and in Barsalou's (1989) concept of ad hoc categories. Details of these ideas and their similarities are given in Morris, Beghtol & Hirst (2003). To date they have not been implemented in a computational system.

Experimental Study

Data Collection

This study is an empirical investigation of the subjectivity of readers' perceptions of lexical cohesion. Twenty-six participants as readers were each given one of three different texts. Each text consisted of the first page and a half of a general interest *Reader's Digest* article (approximately 390 words each). Nine participants read the first two texts and eight read the third. While reading their text, readers were instructed to mark groups of words that they perceived as being related by meaning. They were given a set of 30 coloured pencils with which to mark the word groups by underlining all words in a group with the same colour. These groups of words, referred to as *word groups*, were taken as representing their perceptions of lexical cohesion. After this task was completed, readers described (in one or two sentences) what each word group meant in the text.

Data Analysis

To summarize the degree of agreement between readers, pair-wise agreement statistics were calculated for all *common word groups* in each text and an average for each text and the overall study were computed. In this study, a common word group was defined as one that was contributed to by at least four readers. For a particular reader's word group to join a potential common word group at least half of the words in it must match with those in the potential common word group under formation and the reader's description of their word group must match with that of the common word group. As a potential common word group is formed, it contains all of the words of each word group allowed to join it, and its description becomes one that reflects the descriptions of word groups that join it. This process of common word group formation requires continual and thorough reassessment as word groups are added. In the end, all word groups marked by each reader will have been considered as candidates for each common word group formed.

For each common word group, readers' agreement on membership of the group was computed in the following manner: For all possible pairs of readers, the number of words on which they agreed as a percentage of the total number of words they marked was computed. The average of these pair-wise agreements was computed for each common word group.

Results

The following three tables show the average pair-wise agreements on word membership in the common word groups for each of the three texts in the study. The gloss of the common word group is also given, which is a description (created by the analyst) of the group based on individual descriptions of word groups that make it up.

Table 1: Text 1 - Average % Pair-wise Agreement between Readers on Word Membership in Common Word Groups (Cwgs)

Cwg	Gloss	Average % Pair-wise Agreement
1	positive human characteristics	62.0
2	alcoholics	82.4
3	funerals/death	53.4
4	life events	43.9
5	relationships	61.1
6	jobs/professions	77.6
7	places	72.4
8	places and programs to help people	34.1
9	games	63.3
10	speaking/communication	64.4
	Text 1 Average:	61.5

Table 1: Text 2 - Average % Pair-wise Agreement between Readers on Word Membership in Common Word Groups

Cwg	Gloss	Average % Pair-wise Agreement
1	emotions	60.9
2	research	55.9
3	material world/money and jobs	61.8
4	people/relationships	64.9
5	leisure activities	62.1
6	subjects/areas of study	47.1
7	geography/cultures/countries	65.7
8	measurement	61.3
9	life(time)	52.3
	Text 2 Average:	59.1

Table 3: Text 3 - Average % Pair-wise Agreement between Readers on Word Membership in Common Word Groups

Cwg	Gloss	Average % Pair-wise Agreement
1	places/locations	66.7
2	enforcing/rules	50.1
3	government/law/rules	59.3
4	people	69.8
5	garbage	63.7
6	items of garbage (grocery, paper, recycling, containers)	53.0
7	paper/text/print	57.4
8	war/confrontation	57.3
	Text 3 Average:	59.8

Tables 1, 2, and 3 above show that the text averages are similar. Table 4 below summarizes the average pair-wise agreements for the study.

Table 4: Study - Average % Pair-wise Agreement between Readers on Word Membership in Common Word Groups

	Average % Agreement
Text 1	61.5
Text 2	59.1
Text 3	59.8
Study	60.1

Table 4 indicates, that on average, approximately 60% of the words in the common word groups are agreed on, and that approximately 40% of the words reflect the individual differences (subjectivity) of the readers.

Discussion

One of the primary motivations for this study is that there is no prior research on readers' perceptions of lexical cohesion in text. Existing research does not take readers' perceptions and the potential subjectivity of these perceptions into account. A consequence is that there is no comparison data for the results presented above.

Lexical cohesion is assumed to contribute to the meaning of a text (Halliday & Hasan, 1976; Hasan, 1984; Martin, 1992). Study results indicate that subjectivity, as measured by individual differences in responses, exists in readers' perceptions of lexical cohesion. The major indicator of subjectivity used here, the average pair-wise agreement between readers on word membership in common word groups, shows a 40% amount of individual difference. This implies that readers' perceptions of lexical cohesion contribute to their subjective interpretation of the meaning of text.

The subjectivity found in this study contributes to research on what aspects of a text's meaning are "in the text" and "in the reader". The analysis in which researchers mark the lexical cohesion in text (Halliday & Hasan, 1976; Hasan, 1984; Martin, 1992) focuses on aspects of meaning that are presumed to objectively exist "in the text". Current computational applications based on lexical cohesion such as text summarization (Barzilay & Elhadad, 1999; Silber & McCoy, 2002) and spelling correction (Budanitsky & Hirst, 2001) also focus on meaning that is “in the text”.

This research builds on work, such as that of Olson (2004), that views meaning as something created by the reader (and writer) of text, as opposed to viewing meaning as something that somehow exists in (or can be determined from) text alone. If computers are relied on to interpret the meaning of texts, then the study results indicate the importance of research on how much and what aspects of text meaning are subjective. The view taken here is that text, reader, and writer all contribute to a text’s meaning. Current research on automated text understanding does not consider the potential subjectivity of human text interpretation, and proceeds as if there was one “correct” interpretation, and that this interpretation can obtain from the text itself (as well as any available codified information about the world).

Many natural language processing applications that adopt the “in the text” only view (information retrieval, machine translation, question answering systems etc.) produce adequate results in the context of their limited applications. These applications could be potentially improved by considering reader models that could help account for subjectivity observed between readers. Creating these models would require research on the nature of the subjectivity and whether it reflects attributes of the reader (such as predispositions to certain attitudes). While current applications are not meant to be full text interpretation systems, a goal is to have computational applications that are better able to interpret the meaning of text, perhaps somewhat uniquely for individual readers.

One area for future research is to explore the effect of using different kinds of texts and readers. The current study is limited to general-interest *Reader’s Digest* articles and adult Master’s students in the Faculty of Information Studies at the University of Toronto, and the effects of texts from different genres, styles, or levels of difficulty or readers from different ages, educational backgrounds, cultures, abilities or interests are not known. The results show that readers’ perceptions of lexical cohesion are subjective, which opens the door to research on what the subjectivity means or reflects. For example, does it reflect differences in readers’ attitudes about issues or concepts in the text? This research would require further investigation of the individual differences found between readers. Preliminary results are reported in Morris & Hirst (2005).

References

Barsalou, L. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), p. 211–227.

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

Barzilay, R., & Elhadad, M. (1999). Using lexical chains for text summarization. In I. Mani & M. Maybery (Eds.), *Advances in Text Summarization* (pp. 111–121). Cambridge, Mass.: The MIT Press.

Cruse, D. (1986). *Lexical Semantic Relations.* Cambridge, England: Cambridge University Press.

Fillmore, C. (1968). "The Case for Case". In E. Bach & R. Harms (Eds.), *Universals in Linguistic Theory*, p. 1-88. New York: Holt, Rinehart & Winston.

Grosz, B., & Sidner. C. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3). p. 175–204.

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

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.

Hirst, G, & Budanitsky, A. (2005). Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1), p. 87–111.

Martin, J. (1992). *English Text: System and Structure.* The Netherlands: John Benjamins Publishing Co.

Morris, J., Beghtol, C., & Hirst, G. (2003). Term Relationships and their Contribution to Text Semantics and Information Literacy through Lexical Cohesion. In *Proceedings of the 31st Annual Conference of the Canadian Association for Information Science,* Halifax, Nova Scotia, June 1-June 4.

Morris, J., & Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics.* 17(1), p. 21–48.

Morris, J. & Hirst, G. (2004). The Subjectivity of Lexical Cohesion in Text. In J. Shanahan et al. (eds.) *Computing Attitude and Affect in Text*. p. 41–48. Berlin: Springer.

Olson, D., R. (2004). Knowledge and its artifacts. In K. Chemla (Ed.). *History of science, history of text.* Dordrecht, NL: Kluwer.

Silber, H. Gregory and McCoy, Kathleen F. (2002). Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4), p. 487–496.

Voorhees, Ellen. (1998). Using WordNet for Text Retrieval. In C. Fellbaum, (Ed). *WordNet: An Electronic Lexical Database.* (pp. 285-303). Cambridge, Mass.: The MIT Press.