Non-Classical Lexical Semantic Relations

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

NLP methods and applications need to take account not only of “classical” lexical relations, as found in WordNet, but the less-structural, more context-dependent “non-classical” relations that readers intuit in text. In a reader-based study of lexical relations in text, most were found to be of the latter type. The relationships themselves are analyzed, and consequences for NLP are discussed.

1 Introduction

Many NLP applications, such as text summarization and discourse segmentation, require, or can be helped by, the identification of lexical semantic relations in text. However, the resources that are presently available, such as WordNet (Fellbaum, 1998) provide only “classical” relations: taxonomy or hyponymy (*robin* / *bird*), hypernymy (*tool* / *hammer*), troponymy (*drink* / *guzzle*), meronymy (*hand* / *finger*), antonymy (*go* / *come*), and synonymy (*car* / *automobile*). These relations, which have been widely studied and applied, are characterized by a sharing of the same individual defining properties between the words and a requirement that the words be of the same syntactic class.[1]

Intuitively, however, we see many other kinds of lexical relations in text. As an example, consider the following two sentences taken from a *Reader’s Digest* article:

> I attended a *funeral service* recently. **Kind** words, *Communion*, *chapel* overflowing, *speeches* by <u>lawyers</u>, <u>government workers</u>, 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 **humor**.

There are four groups of related words in this text: the italicized group is about funerals, the bolded group is positive human characteristics, the underlined group is job types, and the capitalized group is games. Some of the lexical relations here are of the classical kind that we mentioned earlier (e.g., *chess* and *Scrabble* have a common subsumer); but others are examples of relations that we will refer to as “non-classical”, such as *funeral* / *chapel* and *humility* / *kindness*. The goal of this research is to investigate these non-classical relations, and to determine what the different types are and how they are used, with a view to eventual automatic detection of the relationships in text.

Most prior research on types of lexical semantic relations has been context-free: the relations are considered out of any textual context and are then assumed to be relevant within textual contexts. And in lexical cohesion research, the analysis of lexical relations has been done by professional linguists with particular points of view (Hasan, 1984; Martin, 1992). A better understanding of the types of lexical semantic relations that are actually identified in context by readers of text will potentially lead to improvements in the types of relations used in NLP applications.

2 Theoretical Background

2.1 The lexical semantic relations used in lexical cohesion

When people read a text, the relations between the words contribute to their understanding of it. Related word pairs may join together to form larger groups of related words that can extend freely over sentence

[1] Causality as a lexical relation (*teach* / *learn*), of which there are just a few examples in WordNet, falls in a grey area here.

boundaries. These larger word groups contribute to the meaning of text through "the *cohesive* effect achieved by the continuity of lexical meaning" (Halliday and Hasan, 1976, p. 320, emphasis added). Lexical semantic relations are the building blocks of *lexical cohesion*, and so a clear understanding of their nature and behavior is crucial. Lexical cohesion analysis has been used in such NLP applications as determining the structure of text (Morris and Hirst, 1991) and automatic text summarization (Barzilay and Elhadad, 1999).

In recent lexical cohesion research in linguistics (Hasan, 1984; Halliday and Hasan, 1989; Martin, 1992) non-classical relations are largely ignored, and the same is true in implementations of lexical cohesion in computational linguistics (Barzilay and Elhadad, 1999; Silber and McCoy, 2002), as the lexical resource used is WordNet. It is notable, however, that the original view of lexical semantic relations in the lexical cohesion work of Halliday and Hasan (1976) was very broad and general; the only criterion was that there had to be a recognizable relation between two words. Most research on lexical semantic relations in linguistics (Cruse, 1986) and psychology has also ignored non-classical relations (with the exception of Chaffin and Herrmann, 1984); however there have been recent calls to broaden the focus and include non-classical relations as well (McRae and Boisvert, 1998; Hodgson, 1991).

A notable exception to this trend is in library and information science (LIS), and is likely a pragmatic reflection of the fact that it is a field with a large user base that demanded this type of access to reference materials. In LIS thesauri, most of the word pairs that are classed as Related Terms (RTs) are related non-classically, but unfortunately are listed as an undifferentiated group. Standards for their use have been developed (ISO, 1986); but since 1985, the Library of Congress has been encouraging a minimization of their use (El-Hoshy, 2001). Since RTs are all grouped together in an unclassified manner, the result has been inconsistencies and subjective judgments about what word pairs are included; but this is an issue of implementation rather than whether RTs can, in principle, be useful.

Roget's Thesaurus, which was used to form the lexical chains in Morris and Hirst (1991), also gives non-classically related word groups. Although this thesaurus is hierarchically classified, it makes frequent use within its basic categories of unclassified pointers to other widely dispersed basic categories. In this respect the structure of LIS thesauri and *Roget's Thesaurus* are similar. They are both hierarchically organized — *Roget's* by Roget's own principles of domain and topic division and LIS thesauri by a broad-term / narrow-term structure — but they also both have a non-hierarchical, non-classified "structure" (or at least mechanism) for representing non-classical relations. But while both, unlike WordNet, give access to non-classically related word pairs, they don't give any indication of what the actual relation between the words is. Other recent computational work such as that of Ji, Ploux, and Wehrli (2003) suffers from the same problem, in that groups of related words are created (in this case through automatic processing of text corpora), but the actual relations that hold between the members of the groups are not determined.

2.2 Non-classical lexical semantic relations

Lakoff (1987) gives the name "classical" to categories whose members are related by shared properties. We will extend Lakoff's terminology and refer to relations that depend on the sharing of properties of classical categories as *classical* relations. Hence we will use the term *non-classical* for relations that do *not* depend on the shared properties required of classical relations. Lakoff emphasizes the importance of non-classical *categories*, providing support for the importance of non-classical *relations*. The classical category structure has been a limiting factor in the study of lexical relations: since relations create categories (and vice versa), if the categories that are considered are severely restricted in nature, so too will be the relations; and, as mentioned, related words must be of the same part of speech. This is thus a restriction found in both Hasan's (1984) relations in lexical cohesion work and Cruse's (1986, p. 16) concept of patterns of lexical affinity, where a mechanism is given for relating inter-sentence and, in fact, inter-text words that are both in the same grammatical class. The lexical chains of Morris and Hirst (1991) had no such restriction, and frequently nouns, verbs, adjectives, adverbs, and verbs were joined together in one chain.

Lakoff (1987) mentions Barsalou's (1989) concept of creating *ad hoc categories*, his term for categories that are "made up on the fly for some immediate purpose", which would presumably require some type of processing interaction with a specific text instead of the assumption that all categories pre-exist (Lakoff, 1987, p. 45). Two examples of these categories are "things to take on a camping trip" and "what to do for entertainment on a weekend" (*ibid*, p. 45). Barsalou's ad hoc categories seem to be of (at least) two types: (1) different activities or actions pertaining to the same or similar objects; (2) different objects pertaining to the same or similar activities or actions. This process has similarities to the mechanisms of Hasan (1984), Martin (1992), and Cruse (1986) that use *both* intra-sentence case-like relations and inter-sentence classical relations. Categories created this way are not classical, as they seem to be ways of joining "different" objects, actions, or activities, and so the relations between their members are not classical either. The mix of classical categories and relations with non-classical categories and relations appears to be a rich source of lexico-grammatical cohesion.

The following are the major (not necessarily mutually exclusive) types of non-classical relations found in the literature:

- Relations between members of Lakoff's non-classical categories: *ball*, *field* and *umpire*, that are part of the structured activity of *cricket* (or *baseball*).
- Case relations:
  - General: *dog* / *bark* (Chaffin and Herrmann, 1984).
  - Sentence-specific (Fillmore, 1968): *stroke* / *it* in the sentence: They *stroked it*.
- LIS RTs (Milstead, 2001).

The relations between members of non-classical categories are unnamable except with reference to the category name (one can't describe the relations between *ball* / *field* or *ball* / *umpire* without using the word *cricket*). For word pairs consisting of a member and the category name, the relation has often been covered, either as a general case relation (*ball* / *cricket* as instrument / activity) or as an RT (*field* / *cricket* as the activity / location relation of Neelameghan (2001), or the locative general case relation).

Case relations come in two varieties: general and specific (to a sentence). The general inter-sentence and inter-text case relations (Chaffin and Herrmann, 1984) are given also by several of the LIS researchers who have provided lists of RT types (Neelameghan, 2001; Milstead, 2001). Cruse deals almost exclusively with classical relations, but does mention two general case-like relations that he calls "zero-derived paronymy" (1986, p. 132). The instrumental case (*dig* / *spade* or *sweep* / *broom*) and the objective case (*drive* / *vehicle* or *ride* / *bicycle*) are given as examples. He observes that in the instrumental case, the definition of the noun will most likely contain the verb, and in the objective case, the definition of the verb will most likely contain the noun. To Cruse, these are not "real" relations but merely "quasi" relations, as the word classes involved differ.

The case relations as defined by Fillmore (1968) are intra-sentence grammatical relations that always apply to the specific text and sentence they are situated in. Sometimes these relations can be both text-specific and general at the same time (*dog* / *barked* in *The dog barked*). Hasan (1984) and Martin (1992) also use these intra-sentence case relations to further link together word groups that have been created through classical relations, as does Cruse (1986) with his concept of patterns of lexical affinity mentioned above.

LIS can lay claim to the most extensive amount of research on non-classical relations. It is interesting to note that during the development of the *Art and Architecture Thesaurus* (AAT), RTs were not included in the initial design, but rather added in afterwards due to user demand (Moholt, 1996). Of the LIS researchers, Neelameghan (2001) has produced the most extensive list of non-classical relations, which has changed little since Neelameghan and Ravichandra (1976). Apart from relations between members of non-classical categories (see above), his list includes most of the text-general relations (recognizable out of the context of a text) mentioned by other researchers. Obviously any text-specific relations such as sentence-specific case cannot be included, since word pairs are considered out of text. Note again, however, that both Hasan (1984) and Martin (1992) use relations similar to text-specific case relations to strengthen cohesive ties created by the classical relations. This combination of text-specific and text-general relations could prove to be useful computationally. A couple of exceptions to the above mentioned relation types have been noted. Evens et al. (1983) have a *provenience* relation (*water* / *well*), and Cruse (1986) has a *proportional series* relation made up of what he calls recurring *endonymy* (*university* / *lecturer* / *student*, *prison* / *warden* / *convict*, *hospital* / *doctor* / *patient*), that is a relation that "involves the incorporation of the meaning of one lexical item in the meaning of another", such as *education* in *university* / *lecturer* / *student* (1986, p. 123–125).

In the research on domain-neutral lexical semantic relations, hundreds (Cassidy, 2000) or thousands (Lenat, 1995) of relations are defined, or perhaps even more in the case of Barrière and Popowich (2000). The question of whether there is a smallish set of field- (domain-) neutral non-classical relations that will provide (good) coverage for all (or most) fields is one of the questions we are investigating. Encouragingly, LIS has tackled an extensive number of specific domains with just such a smallish set of field-neutral non-classical relations. However, due to the reportedly subjective implementation of these relations, this may not in fact be true in practice. WordNet's approach uses domain-neutral relations for a general domain, but mostly for classical relations. Databases use domain-specific relations for specific domains.

3 Experiment

3.1 Introduction

We are interested in determining and analyzing the types of lexical semantic relations that can be identified in text. To this end, a study was conducted with nine participants who read the first 1.5 pages of a general-interest article from the *Reader's Digest* on the topic of the funeral of a homeless alcoholic who had nonetheless achieved many positive aspects and qualities in his life. The study reported here is part of a larger study of three texts from the *Reader's Digest* that investigates not only the relation types used but also the nature of the larger word groups, the interactions among the word groups, how much of and what type of text meaning this information represents, and the degree of subjectivity in the readers' perceptions of both the relation types and word groups as measured by individual differences (see Morris and Hirst, 2004).

3.2 Method

Subjects were given a large set of colored pencils and a supply of data sheets for recording their observations. They were instructed to first read the article and mark the words that they perceived to be related, using a different color of pencil to underline the words of each different group of related words. (In effect, they built lexical chains; two words could be in the same group even if not directly related to one another if both were related to another word in the group.) They were told that they could re-read the text and add new underlining at any time during this part of the study. Once this task was completed, the subjects were instructed to transfer each separate word group to a new data sheet, and for each group to indicate which pairs of words within the group they perceived to be related,and what the relation was. Finally, they were asked to describe what each word group was "about", and to indicate whether and how any of the word groups themselves were related to another.

3.3 Results

We will briefly present some statistics that summarize the degree of agreement between the subjects, and then turn to a qualitative analysis.

In general, the subjects were in broad agreement about many of the groups of related words — for example, that there was a "funerals" group and a "positive human qualities" group — but, as one would expect, they differed on the exact membership of the groups. Eleven groups were identified by at least four of the nine subjects. 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 at least four 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 25% of the word pairs were marked by at least 50% of the subjects.

For this set of word pairs that were identified by more than one subject, we then computed agreement on what the relationship between the pair was deemed to be. We found that the subjects agreed in 86% of the cases.

We now turn to the nature of lexical relations that the subjects reported perceiving in the text in each of the eleven word groups that were used by at least four of the readers. As we would expect, the individual wording of the descriptions of relation types varied greatly by reader: the subjects often used different ways to describe what were clearly intended to the same relations. Thus, we had to analyze and interpret their descriptions. We were careful in this analysis to try to determine the subjects' intent and generalize the conceptual meaning of the individual wordings that were given, but not impose any view of what the relations "should be".

We found that for this one text, there seems to be an emerging "smallish" set of 13 commonly used relations, listed below. Not included in the list are the outlier relations — the relation types used only by one reader.

1. Positive qualities (*brilliant* / *kind*).
2. Negative qualities (*homeless* / *alcoholic*).
3. Qualities in opposition (*drunk* / *drying out*).
4. Large categories such as positive human characteristics (*humility* / *humour*), typical major life events (*funeral* / *born* / *married*), and jobs / types of people (*lawyer* / *volunteer*).
5. Words that are each related to a third concept; for example caring (*kind* / *gentlemanly*), remember (*speeches / deceased*), and education (*people / professors*).
6. Descriptive noun / adjective pairs (*born* / *young*, *professors* / *brilliant*).
7. Commonly co-occurring words often described as words that are associated, or related: (*alcoholic* / *beer*). In many cases the readers used subgroups of this category:
   a. Location (*homeless* / *shelter*, *funeral* / *chapel*, *kitchen* / *home*)
   b. Problem / solution / cause / one word leads to the other (*homeless* / *drunk*, *date* / *love*, *date* / *relationship*, *alcoholic* / *rehab program*).
   c. Case relations (*volunteer* / *service*, *people* / *living*, *speeches* / *friends*).
   d. Aspects of an institution: married (*son* / *married*), funeral (*speeches* / *communion*), and education (*college* / *jobs*).
8. Stereotypical relations (*homeless* / *drunk*, *people* / *home*).
9. One word related to a large group of words, seemingly with a lot of import: (*homeless* /

⟨the group of positive human characteristics such as *brilliant* / *kind* / *humility*⟩).

10. Definitional: (*alcoholic / drunk*) .
11. Quasi-hyponymy relations (*friend* / *relationship*).
12. Synonymy (*relaxed* / *at ease*).
13. Antonymy (*died* / *born*).

The data show that while individual differences occur (Morris and Hirst, 2004), the readers in the study identified a common core of groups of related words in the text. Agreement on which exact word pairs within a group are related is much lower at 25%, and possible reasons for this are, briefly, that this is a much more indirect task for the readers than initially identifying word groups and that the word groups might be comprehended more as gestalts or wholes. In cases where subjects identified word pairs as related, they also showed a marked tendency, at an average of 86%, to agree on what the relation was. This high level of reader agreement on what the relations were is a reflection of the importance of considering lexical semantic relations as being situated in their surrounding context. In other words, while explaining or perceiving linguistic meaning out of context is hard, as noted by Cruse (1986), doing so within text seems here not to be, and is therefore likely a meaningful area for further study.

One clear pattern was evident in the analysis: the overwhelming use of non-classical relations. There were a few uses of hyponymy, synonymy, and antonymy (relations 11, 12, and 13 above), but these classical relations were used only for a minority of the word pairs identified by the readers from within the word groups in the text.

4 Discussion

The subjects in this study identified a common "core" of groups of related words in the text, as well as exhibiting subjectivity or individual differences. Within these word groups, the subjects identified a "smallish" group of common relation types. Most of these relation types are non-classical. This result supports the integration of these relations into lexical resources or methods used by NLP applications that need to identify and use lexical semantic relations and lexical cohesion in text. There are two related computational issues. The easier one is to be able to automatically identify words in a text that are related. Much harder is to be able to provide the semantically rich information on what the relation actually is.

Clearly this work is preliminary in the sense that, to date, one text has been analyzed. Our next step is to complete the analysis of the data from the other two texts in this study, which has been collected but not yet analyzed. An obvious area for future research is the effect of different types of both texts and readers. Our readers were all masters-level students from the Faculty of Information Studies, and the three texts are all general-interest articles from *Reader's Digest.*

It would be very useful to do a thorough analysis of the correspondence between the readers' relation types reported above, and the relation types discussed earlier from the literature. A preliminary look indicates overlap, for example of inter-sentence case relations, ad hoc non-classical categories, and words related through a third concept. We would like to investigate the potential of using both classical and non-classical relation types along with the intra-sentence case relations for the automatic generation of relations and relation learning. This work would incorporate and build on the related ideas discussed above of Cruse (1986), Hasan (1984), and Barsalou (1989), along with the actual relation types and word group interactions found by readers.

We are also 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 intend to investigate any potential linkages between the word groups in the texts and other theories that provide pre-determined structures of text, such as Rhetorical Structure Theory (Marcu, 1997). It will also be useful for computational purposes to have a clearer understanding of what aspects of text understanding exist "in it" and what can be expected to contribute to subjectivity of interpretation or individual differences in comprehension.

Acknowledgments

This research was supported by a grant and scholarship from the Natural Sciences and Engineering Research Council of Canada. We are grateful to Clare Beghtol for ongoing comments.

References

Barrière, Caroline and Popowich, Fred (2000). Expanding the type hierarchy with nonlexical concepts. In Howard Hamilton (Ed.), *Canadian AI 2000* (pp. 53–68). Berlin: Springer-Verlag.

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, England: Cambridge University Press.

Barzilay, Regina and Elhadad, Michael (1999). Using lexical chains for text summarization. In Inderjeet Mani and Mark Maybury (Eds.), *Advances in text summarization* (pp. 111–121). Cambridge, Mass.: The MIT Press.

Cassidy, P. (2000). An investigation of the semantic relations in the Roget's Thesaurus: Preliminary results. In A. Gelbukh (Ed.). *CICLing-2000: Conference on Intelligent Text Processing and Computational Linguistics, February 13–19, Mexico City, Mexico*, 181–204.

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

Cruse, D. (1986). *Lexical semantic relations.* Cambridge, England: Cambridge University Press.

El-Hoshy, S. (2001). Relationships in Library of Congress Subject Headings. In C. Bean, and R. Green (Eds.), *Relationships in the organization of knowledge* (pp. 135–152). Norwell, Mass: Kluwer Academic Publishers.

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

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

Fillmore, Charles (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.

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

Halliday, M.A.K., and Hasan, Ruqaiya (1989). *Language, Context and Text: Aspects of language in a social-semiotic perspective*. Geelong, Victoria: Deakin University Press. (republished by Oxford University Press, 1989).

Hasan, Ruqaiya (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 prelexical priming. *Language and Cognitive Processes*, 6(3), 169–205.

ISO. (1986). Guidelines for the establishment and development of monolingual thesauri. [Geneva:]. ISO. (ISO2788-1986(E)).

Ji, Hyungsuk, Ploux, Sabine, and Wehrli, Eric (2003). Lexical knowledge representation with contexonyms. *Proceedings, Machine Translation Summit IX,* New Orleans, September 2003.

Lakoff, George (1987). *Women, Fire and Dangerous Things.* Chicago: University of Chicago Press.

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11), 33–38.

Marcu, Daniel (1997). From Discourse Structures to Text Summaries. In Inderjeet Mani, and Mark Maybury (Eds.), *Intelligent Scalable Text Summarization, Proceedings of a Workshop Sponsored by the ACL*, 82–88. Somerset NJ: Association for Computational Linguistics.

Martin, James (1992). *English text: System and structure.* Amsterdam: John Benjamins Publishing Co.

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.

Molholt, P. (1996). *A Model for Standardization in the Definition and Form of Associative, Interconcept Links.* (Doctoral dissertation, Rensselaer Polytechnic Institute).

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

Morris, Jane and Hirst, Graeme (2004). The subjectivity of lexical cohesion in text. *AAAI Spring Symposium on Exploring Affect and Attitude in Text*, Stanford.

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.

Neelameghan, A., and Ravichandra, R. (1976). Nonhierarchical associative relationships: Their types and computer generation of RT links. *Library Science*, (13), 24–42.

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

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