Attributions

Sabine Bergler, Monia Doandes, Christine Gerard, and René Witte
Department of Computer Science
Concordia University
1455 de Maisonneuve Blvd. West
Montreal, QC, Canada, H3G 1M8
email contact: bergler@cs.concordia.ca

Abstract

We present here the outline of an ongoing research effort to recognize, represent, and interpret attributive constructions such as reported speech in newspaper articles. The role of reported speech is attribution: the statement does not assert some information as 'true' but attributes it to some source. The description of the source and the choice of the reporting verb can express the reporter's level of confidence in the attributed material.

Introduction

Many broad coverage systems employ shallow, often stochastic means to ensure robustness and forego any form of linguistic analysis. Some successful systems use hybrid technology (cf. (Clarke *et al.* 2000), (Harabagiu *et al.* 2001), (Surdeanu *et al.* 2003), (Hovy *et al.* 2001)). Our research is similarly positioned for broad coverage, using shallow tools, yet embedding linguistically motivated modules for performance enhancement. We present here the outline of an ongoing research effort to recognize, represent, and interpret attributive constructions such as reported speech[1] in newspaper articles.[2]

Newspaper articles have a standardized way to relate opinions, point-of-view, and sentiments: reported speech. In some articles, 90% of the sentences are direct or indirect reported speech. In most current language processing systems it is, however, ignored.

The role of reported speech is attribution: the statement does not assert 'true' what amounts to the information content of the sentence, but a situation in which this content was proffered by some source. This device can be used both, to bolster a claim made in the text already, or to distance the author from the attributed material, implicitly lowering its credibility (Anick & Bergler 1992). Thus the credibility or reliability of the attributed information is always in question for reported speech and other attributions. If the attribution is used to bolster a claim already made by citing a particularly strong source for endorsement, ignoring the fact that an explicit attribution was made will do no great harm. This is in fact a frequent case in the type of newspaper articles typically used for large scale system development and testing (as in MUC, TREC, DUC, ...) and this is why ignoring attribution has been tolerable. But when a text is argumentative (opposing two or more points of view on a topic) or speculative (when the final outcome of an event is not yet decided and the text uses different sources as predictors), text meaning depends on proper attribution recognition (Bergler 1995).[3]

To test the practicability of attribution recognition and its impact, we present three stand alone[4] systems previously developed in our lab. The systems roughly cover *recognition* of attribution and its *representation*. We also outline what is required for the *interpretation* of attribution in order to perform standard belief revision and maintenance.

Note that this work has important differences with traditional work on *belief* reports. Utility texts such as newspaper articles expressly avoid using belief reports. They represent after all an evaluation by the reporter which the reader might not share. What occurs in newspaper articles are instead *evidence* reports[5] which do not give rise to beliefs without an additional interpretation step. (Gerard 2000) presents one possible way to transform these evidence reports into first potential, and eventually held beliefs. But the focus here is on the extraction of a proper representation of the evidence reports from real texts that will enable the intricate reasoning about beliefs discussed in the literature (see for instance (Rapaport, Shapiro, & Wiebe 1997; Rapaport 1986)).

Recognizing reported speech and selected other construc-

---

[1] *Reported speech* here refers to both direct or quoted speech, and indirect reported speech. We are working towards including all attributions, including "according to the New York Times" and "the report shows ...". The systems presented in this report, however, are limited to reported speech of the different patterns possible for matrix clause/subordinate clause, where the matrix clause contains at least a source NP and a reporting verb.

[2] The recognition and representation of attribution in other genres is analogous, but the interpretation strategies might differ.

---

[3] An argumentative or speculative structure is not limited to newspaper articles. Scientific articles, too, use reported speech for this purpose. And multi-participant political analysis segments on newscasts form the same phenomenon.

[4] We anticipate integrating the prototype systems for large scale automatic testing.

[5] See (Bergler 1992) for a detailed analysis.

tions is based on a reported speech recognition and representation system (Doandes 2003). Interpretation will be based on an exploratory prototype of belief acquisition called *Percolator* (Gerard 2000). The bridge between them is the fuzzy-set based coreference resolution system Fuzzy-ERS (Witte & Bergler 2003). We describe the modules in the following sections on an example text presented in Figure 1.

Politics & Policy: Democrats Plan Tactic to Block Tax-Cut Vote

Threat of Senate Filibuster Could Indefinitely Delay Capital-Gains Package

(S_1) Democratic leaders have bottled up President Bush's capital-gains tax cut in the Senate and may be able to prevent a vote on the issue indefinitely.

(S_2) Yesterday, Sen. Packwood acknowledged, "We don't have the votes for cloture today."

(S_3) The Republicans contend that they can garner a majority in the 100-member Senate for a capital-gains tax cut. (S_4) They accuse the Democrats of unfairly using Senate rules to erect a 60-vote hurdle.

(S_5) Democrats asserted that the proposal, which also would create a new type of individual retirement account, was fraught with budget gimmickry that would lose billions of dollars in the long run.

Figure 1: Adapted from Jeffrey H. Birnbaum, The Wall Street Journal, 10/27/89

Recognizing Reported Speech

Doandes (Doandes 2003) developed a context-free grammar to chunk contiguous verb groups and extracts grammatical information from the chunks (tense, aspect, modality, polarity). Using a list of possible[6] reported speech indicating verbs, she then applies the known patterns for reported speech (Quirk *et al.* 1985) to separate the reporting clause from the reported material.

The development was done on 65,739 sentences from the Wall Street Journal, testing on 2404 sentences taken mainly from the Wall Street Journal, with a few articles from the DUC 2003 corpus. 513 occurrences of reported speech were found, which makes for a precision of 98.65% and a recall of 63%.

Recall problems are linked to tagging errors (a basic version of the Brill tagger was used), an incomplete list of reported speech verbs, the chunking process (the NP chunker occasinaly splits heavy NPs into several smaller chunks, thus obfuscating the reported speech pattern), and the fact that the chunker used a chart parser which returned the first parse tree according to some internal criterion, not necessarily the best.

Figure 2 shows the condensed version of the system's output in the form of *reported speech frames* which are counted as correct only if they fully match the hand annotated Gold standard. Additional information in the matrix clause (so called circumstantial information, such as time, place, manner adverbials or prepositional phrases) are sometimes not

---
[6]No word sense disambiguation is attempted.

S_2
textsource Sen. Packwood
textverb acknowledged
reported speech 'We don't have the votes for cloture today.'
circumstantial Yesterday,
semantic dimensions EXPLICIT:yes
                POLARITY:positive
                SPEECHACT:inform
                STRENGTH:high

S_3
textsource The Republicans
textverb contend
reported speech they can garner a majority in the 100-member Senate for a capital-gains tax cut.
semantic dimensions EXPLICIT:yes
                POLARITY:positive
                SPEECHACT:inform

S_4
textsource They
textverb accuse
reported speech the Democrats of unfairly using Senate rules to erect a 60-vote hurdle.
semantic dimensions EXPLICIT:yes
                POLARITY:positive
                SPEECHACT:inform

S_5
textsource Democrats
textverb asserted
reported speech the proposal, which also would create a new type of individual retirement account, was fraught with budget gimmickry that would lose billions of dollars in the long run.
semantic dimensions EXPLICIT:yes
                POLARITY:positive
                SPEECHACT:inform
                STRENGTH:high

Figure 2: Internal structure of the reported speech sentences S_2–S_5.

parsed correctly. In this case, the frame is not considered correct, even though the main aspects of the reported speech are correctly analyzed. The implementation has a precision of 88% on identifying reported speech and determining its internal structure, its recall is 56%

Representation

The sentence-by-sentence representation of Figure 2 is of course but a beginning. Even in our simple text we see that the same source occurs twice (*republicans, They*), suggesting strongly that both sentences form a logical unit. The basic frames for S_3 and S_4 should be merged into one larger *profile* for that source (see (Bergler 1995) for more detail on profiles and their use in text representation.).

The use of profiles is simple: profiles provide a partition of the text according to the source of the information transmitted. Any evaluation of the reliability/credibility will thus affect the entire set. Moreover, the set of sources in a text in

turn can influence the evaluation of their credibility: compare the status of *Police Officer XYZ* compared to *the thief* (status: high reliability) or *disciplinary commission* (status: neutral). We see that while we cannot reliably determine the meaning of the text, organizational ontologies might provide a first step towards these evaluations. Thus the basic frames representing a single utterance should be grouped by coreference resolution on the source into profiles.

(Witte & Bergler 2003) describes coreference resolution of isolated NPs using fuzzy set theory to explicitly model the inherent uncertainty in our heuristic-based, non-probabilistic approach to coreference resolution. In short, a coreference chain is represented as a fuzzy set mapping all noun phrases in a text to the $[0, 1]$-interval, which represents the degree of membership of each NP in the coreference chain. For further processing a threshold value determines which NP is considered "in" and which not. These coreference chains have been employed as the main criterion for selecting NPs in a text summarization system (Bergler *et al.* 2003). While the coreference chains are not always exact, they work best for named entities, which are the most frequent source descriptions. This system groups S_3 and S_4 as having the same source.

Interpretation

While the proper representation of the argumentative or speculative structure of a text is important by itself and can directly be used for information retrieval and extraction, for instance, it begs the question of how to interpret the attributions made. To this end, (Gerard 2000) developed a system *Percolator*, extending the basic notions of (Wilks & Ballim 1991). While Wilks and Ballim were concerned with the proper attribution of surmised beliefs to dialog partners, Gerard used their representation scheme of nested beliefs and introduced the notion of belief acquisition through percolation.

Our text would be translated into the nested structures of Figure 3[7] based on the profile structure.

In this representation, each reported speech complement is indexed by its *Source-list*. The Source-list holds the description of the source (as given in the subject noun phrase of the matrix clause) and four evaluation features, which represent in turn: the reporter's confidence in the source(s)[8], the reporter's confidence in the reported information[9], the reader's confidence in the source[10], and the reader's confidence in the reported information[11]. Important here is the

---

[7]In fact Figures 3 and 4 have been edited for display. *Percolator* uses a hand-crafted predicate argument structure representation of the reported material, having been developed first. But little rests on this point, since it is a feasibility study of the accumulation of the Source-list information and its potential for interpretation, which are not affected by the representation of the complement clause.

[8]as expressed by the lexicalization of the matrix clause's subject NP

[9]as expressed by the lexicalization of the reporting verb

[10]potentially from prior knowledge or beliefs

[11]again, potentially from previous knowledge or beliefs

System believes
  Reader believes
    Reporter Jeffrey H. Birnbaum believes
      Sen. Packwood said
        ["We don't have the votes for cloture today."
         Source-list(Sen. Packwood, h, n, n, n)]
      Republicans said
        [Republicans can garner a majority in the 100-member
         Senate for a capital-gains tax cut.
         Source-list(Republicans, n, n, n, n)]
        [the Democrats are unfairly using Senate rules to
         erect a 60-vote hurdle.
         Source-list(Republicans, n, h, n, n)]
      Democrats said
        [the proposal, which also would create a new type of
         individual retirement account, was fraught with
         budget gimmickry that would lose billions of dollars
         in the long run.
         Source-list(Democrats, n, n, n, n)]

Figure 3: *Percolator* representation of the Profiles in Figure 2

translation of the *strength* feature of the reporting verb into a credibility rating of the reported information: *acknowledge* as a reporting verb here carries an implication that the information of the complement clause is negative for the subject. If a source is reported to *acknowledge* information this has to be rated as information of high reliability[12] because sources will not falsely make detrimental statements.

The Source-list pairs the reporter's apparent evaluation of the source and the reported information from lexical semantics with the reader's evaluation of the source and the reported information. This reflects a reader's ability to immediately discount the reporter's apparent evaluation based on previous knowledge or on previous beliefs (about the reporter, the source, or the information). But a reader with no relevant previous beliefs has to rely solely on the intrinsic evaluation of the reporter.

Percolator transforms this representation in several steps by pulling information from nested contexts into their surrounding contexts (a process called "percolation" by (Wilks & Ballim 1991)). In every step the source list grows to reflect the nesting that has been lost, thus making explicit the path of attribution. Gerard showed with several examples that an extension of Wilks' and Ballim's percolation mechanism allows to properly combine subjective evaluations of each level of attribution. But the representation of (Wilks & Ballim 1991) was based on beliefs, so Gerard introduced the notion of a *potential belief*, defined as information that might or might not turn into a *held belief* given further evidence. This reflects a reader's ability to accommodate contradictory information: having read an article that presents two contradictory theories, one can argue both ways if one has no own opinion on the matter until there is evidence that "settles" the issue. This representational device allows to delay the decision as to whether information is believed until a certain threshold of comfort is reached. The text here would

---

[12]Possible values for the prototype were **h**igh, **n**eutral, and **l**ow.

System believes
  Reader believes

  —

  Reader potentially believes
    ["We don't have the votes for cloture today."
     Source-list(Sen. Packwood, Reporter Jeffrey H. Birnbaum, h, n, n, n)]
    [Republicans can garner a majority in the 100-member
     Senate for a capital-gains tax cut.
     Source-list(Republicans, Reporter Jeffrey H. Birnbaum, n, n, n, n)]
    [the Democrats are unfairly using Senate rules to
     erect a 60-vote hurdle.
     Source-list(Republicans, Reporter Jeffrey H. Birnbaum, n, h, n, n)]
    [the proposal, which also would create a new type of
     individual retirement account, was fraught with
     budget gimmickry that would lose billions of dollars
     in the long run.
     Source-list(Democrats, Reporter Jeffrey H. Birnbaum, n, n, n, n)]

Figure 4: Final *Percolator* structure of Text 1 in the case where the reader has no prior knowledge or beliefs.

be transformed into a structure as given in Figure 4.

Conclusion

Attribution is a phenomenon of great interest and a principled treatment is important beyond the realm of newspaper articles. We contend that the way natural language has evolved to reflect a culture's understanding of attribution[13] can serve as a guide to a principled representation that can form the basis for a belief revision or maintenance system. We have distilled three major steps for this process: recognition, representation, and interpretation. Here we have only been able to hint at how each step might be performed. We will integrate the individual modules into an attribution resolution module, which will provide a test bed for two different kinds of experiments: without taking the interpretations computed by *Percolator* into account we will measure the usefulness of representing attributions and the argumentative or speculative structure of a text explicitly for Question Answering, Summarization, and Information Extraction. With the *Percolator* output we will be able to determine experimentally exactly what the semantic contribution of different lexical items is on text comprehension and belief acquisition.

Acknowledgments

References

Anick, P., and Bergler, S. 1992. Lexical structures for linguistic inference. In Pustejovsky, J., and Bergler, S., eds., *Lexical Semantics and Knowledge Representation*. Berlin: Springer Verlag. 121–135.

Bergler, S.; Witte, R.; Khalife, M.; Li, Z.; and Rudzicz, F. 2003. Using knowledge-poor coreference resolution for text summarization. In *On-line Proceedings of the Workshop on Text Summarization, Document Understanding Conference (DUC 2003)*. http://www-nlpir.nist.gov/projects/duc/pubs.html.

Bergler, S. 1992. *Evidential Analysis of Reported Speech*. Ph.D. Dissertation, Brandeis University.

Bergler, S. 1995. From lexical semantics to text analysis. In Saint-Dizier, P., and Viegas, E., eds., *Computational Lexical Semantics*. Cambridge, UK: Cambridge University Press.

Clarke, C.; Cormack, G.; Kisman, D.; and Lynam, T. 2000. Question answering by passage selection. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9), Gaithersburg, Maryland*.

Doandes, M. 2003. Profiling for belief acquisition from reported speech. Master's thesis, Concordia University, Montreal, Canada.

Gerard, C. 2000. Modelling readers of news articles using nested beliefs. Master's thesis, Concordia University, Montreal, Canada.

Harabagiu, S.; Moldovan, D.; Pasca, M.; Mihalcea, R.; Surdeanu, M.; Bunescu, R.; Girju, R.; Rus, V.; and Morarescu, P. 2001. The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001), Toulouse France*, 274–281.

Hovy, E.; Gerber, L.; Hermjakob, U.; Lin, C.-Y.; and Ravichandran, D. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the DARPA Human Language Technology Conference (HLT), San Diego, CA*.

Quirk, R.; Greenbaum, S.; Leech, G.; and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rapaport, W. J.; Shapiro, S. C.; and Wiebe, J. 1997. Quasi-indexicals and knowledge reports. *Cognitive Science* 21(1):63–107.

Rapaport, W. J. 1986. Logical foundations for belief representation. *Cognitive Science* 10:371–422.

Surdeanu, M.; Harabagiu, S.; Williams, J.; and Aarseth, P. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), Sapporo, Japan*, 8–15.

Wilks, Y., and Ballim, A. 1991. *Artificial Believers*. Norwood, NJ: Erlbaum.

Witte, R., and Bergler, S. 2003. Fuzzy coreference resolution for summarization. In *Proceedings of the International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS 2003)*, 43–50.

---

[13]See, for instance, Native American languages which grammaticalize attribution.