Res Lang Comput (2008) 6:241–245
DOI 10.1007/s11168-008-9057-3

Introduction to the Special Issue on Ambiguity and Semantic Judgments

Massimo Poesio · Ron Artstein

Published online: 25 October 2008

Most computational and psycholinguistic studies of language interpretation are based on the assumption that each use of a linguistic expression has a unique intended interpretation in context, and that addressees or readers can recover such a unique interpretation. The task for those aiming to develop a model of interpretation, then, is to identify which of the interpretations of a linguistic expression is intended in a given context by its speaker or preferred by its addressee, or to study the factors that affect the choice of interpretation. Most researchers in the area would probably accept that this assumption is just a convenient idealization, but they would also probably expect it to be harmless, in that it would hold of the overwhelming majority of cases of language use.

However, increasing evidence from both corpus-based work in computational linguistics and psychological experimentation suggests that this idealization may be more questionable that so far assumed. Computational work, in particular on word sense disambiguation and anaphora resolution, has revealed how hard it is to pinpoint

This work began at the University of Essex under the sponsorship of the Arrau project, EPSRC grant GR/S76434/01 (United Kingdom, both authors). The second author has continued this work at the Institute for Creative Technologies, and the following applies to this part of his effort: The project or effort described here has been sponsored by the US Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

M. Poesio
Università degli Studi di Trento, Trento, Italy
e-mail: massimo.poesio@unitn.it

M. Poesio
University of Essex, Colchester, UK

R. Artstein (✉)
Institute for Creative Technologies, University of Southern California, Marina del Rey, CA, USA
e-mail: artstein@ict.usc.edu

what, exactly, is the interpretation of many expressions in their context—a precondition for the successful creation of a gold standard annotation (this a problem long known to lexicographers—see, e.g., Kilgarriff 1997). For instance, studies of the extent of agreement among coders carried out by Véronis (1998) and others as part of work on the SENSEVAL-2 word sense disambiguation competition, revealed that for some words (*correct*, *historique*, *économie*, *comprendre*) there was no full agreement between coders on the interpretation of any of the uses of those words. In our own work on anaphoric interpretation (Poesio and Artstein 2005) we studied examples like (1), asking up to 20 subjects to identify the antecedent of each anaphoric expression.

(1)
1.1	M:	. . .
1.4		first thing I'd like you to do
1.5		is send engine E2 off with a boxcar to Corning to pick up oranges
1.6		as soon as possible
2.1	S:	okay
3.1	M:	and while it's there *it* should pick up the tanker

Some of the coders indicated *engine E2* as antecedent for the second *it* in utterance 3.1, whereas others indicated the immediately preceding pronoun, which they had previously marked as having *engine E2* as antecedent. The great majority of anaphoric expressions in the data investigated were marked as being implicitly ambiguous in this way, even after outlier removal. This work also revealed that the extent of disagreement on the interpretation of natural language expressions becomes apparent when large numbers of coders are asked to express judgments.

It may seem at first that these uses of anaphoric expressions are simply cases of infelicitous communication. But this would not do away with the problem of what to do with such uses when attempting to create a gold standard annotation for a corpus. Besides, it may be argued that the idealized view of communication, according to which natural language expressions are always meant to be uniquely disambiguated in context, is too simplistic. For starters, it has always been known that speakers and writers do not always attempt to avoid ambiguity: on the contrary, in certain genres at least, ambiguity can be deliberate—notoriously so in law and politics (Wagner and Cacciaguidi-Fahy 2006), but also in poetry (Su 1994) and in humor (Raskin 1985). But even in cases like (1), in which the speaker did not deliberately set out to be ambiguous, psycholinguistic evidence reviewed in Poesio et al. (2006) suggests that interpreters do not necessarily perceive the ambiguity as problematic—perhaps because they are satisfied with 'good enough' interpretations and therefore may not even recognize the ambiguity (Ferreira et al. 2002). Similar evidence for word sense disambiguation was presented by Frazier and Rayner (1990).

In fact, even the assumption that the speaker has in mind a unique intended interpretation has been challenged. Research on spoken dialogue communication by Herbert Clark and colleagues has shown that often the meaning attributed to a natural language expression is not determined a priori, but is the result of a complex negotiation between speaker and addressee (Clark 1996); it could be argued that this holds of the pronoun in (1). In such cases, attempting to build a system capable of recovering the originally intended meaning would not make any sense.

This special issue of *Research in Language and Computation* was born out of the belief that further study of ambiguous language use is essential both from a linguistic perspective and from a language technology perspective. From a linguistic perspective, we believe that there is a need to understand the extent to which actual communication approximates the idealized Gricean picture, and to identify problematic and unproblematic types of ambiguous language use. From a language technology perspective, the possibility that a linguistic expression may not have a single interpretation in context has implications both for corpus annotation and for system development. The developers of annotated corpora will need novel methods to assess agreement among coders, as well as methods to create gold standard annotations out of multiple reference judgments. For system development, supervised learning methods for tasks such as anaphora resolution will be needed that do not rely on the assumption that each item in the gold standard has a single classification (methods for working with multiple labels have already been developed for text classification and machine translation). The presence of 'dangerous' ambiguity in text may also lead to the development of a new task, ambiguity detection (see the article by Willis et al. in this issue).

The articles in this special issue further our understanding of ambiguous language use by addressing several of the questions raised above. A major emphasis is on investigating and quantifying the degree of ambiguity that naturally occurs in language use, and the effects and consequences of such ambiguity on interpretation (contributions by Boleda et al., Knees, Versley, and Willis et al.).

"An analysis of human judgements on semantic classification of Catalan adjectives" by Gemma Boleda, Sabine Schulte im Walde and Toni Badia reports a study in which large numbers of Web users were asked to classify Catalan adjectives according to a very basic classification scheme. The article conducts a detailed analysis of the subjects' judgments in order to understand the causes for disagreement among the subjects.

"The German temporal anaphor *danach*—ambiguity in interpretation and annotation" by Mareile Knees takes a look at a specific linguistic expression, German *danach* ("thereafter"). Using an annotation scheme which names the object that the pronoun refers to and also marks the text region which evokes that referent, the article develops a taxonomy of ambiguities and how they affect the overall interpretation of the pronoun.

"Mechanisms of semantic ambiguity resolution: insights from speech perception" by Daniel Mirman reviews psycholinguistic work on the resolution of word sense ambiguities. Drawing parallels with comparable approaches to speech perception studies, the article argues for an interactive model of bottom–up perceptual cues and top–down contextual information that select the preferred meaning.

"Disambiguating rhetorical structure" by Manfred Stede is focused on the methodological problems in corpus annotation raised by the presence of ambiguity in text. Work on ambiguous language use might raise the objection that it's not clear to what extent ambiguity is a real-world phenomenon, as opposed to being originated by theory-internal theoretical distinctions. This issue was already raised during the creation of the Penn Treebank (Marcus et al. 1993) but is particularly acute for the annotation of the higher levels of language interpretation. Discourse structure is perhaps

the most egregious example of a type of annotation in which many judgments are highly subjective, and the article points out a number of problems raised by annotating according to the notions of Rhetorical Structure Theory (Mann and Thompson 1988). The creators of language resources prefer to concentrate on annotating those aspects of such annotations that are the most theory-independent; the article proposes an analysis framework, Multi-Level Analysis, which aims at annotating those aspects of discourse structure which are theory-independent, therefore minimizing the extent of ambiguity.

"Vagueness and referential ambiguity in a large-scale annotated corpus" by Yannick Versley is a study of referential ambiguity in German text. The article argues that ambiguity appears to be acceptable in written text as well as in spoken dialogue, and in a wider range of cases than accounted for by the 'Justified Sloppiness Hypothesis' of Poesio et al. (2006). Consequently, the article proposes a *Generalized Sloppiness Hypothesis* covering such cases.

"Automatic Identification of Nocuous Ambiguity" by Alistair Willis, Francis Chantree and Anne De Roeck investigates the ambiguity of modifier attachment to coordinated phrases, with the aim of identifying cases which are likely to be interpreted differently by different people—these are called *nocuous ambiguity*. The study is based on a database of 138 sentences that are potentially ambiguous, interpreted by 17 subjects. This database is used not only to gather empirical evidence about such cases of ambiguity, but also to develop a system capable of recognizing cases of nocuous ambiguity, in order to edit or remove them (e.g., from technical manuals).

The articles by Knees and Versley are extended versions of papers presented at the workshop on ambiguity in anaphora, which was held at ESSLLI 2006 (the Eighteenth European Summer School in Logic, Language and Information) in Málaga, Spain; the remaining articles were solicited for this special issue. We are indebted to Ruth Kempson, who suggested that we broaden the scope of our workshop into this special issue and oversaw it during the early stages of the editing process, and to Shuly Wintner who took over as editor in chief and saw the issue through publication. We wish to thank the authors of all 13 submissions (of which we were only able to include 6), and the reviewers who provided invaluable input and comments for all the submissions: Jennifer Arnold, Chris Brew, Christian Chiarcos, Philipp Cimiano, Kees Van Deemter, Sonja Eisenbeiss, Christiane Fellbaum, Ruth Filik, Brendan Gillon, Daphna Heller, Ryu Iida, Eric Joanis, Lenhart K. Schubert, Andrew Kehler, Adam Kilgarriff, Tanya Kraljic, Sandra Kübler, Bernardo Magnini, Rada Mihalcea, Sergei Nirenburg, Kemal Oflazer, Becky Passonneau, Philip Resnik, Jennifer Rodd, Anne Pier Salverda, Christoph Scheepers, Suzanne Stevenson, Carlo Strapparava, Patrick Sturt, David Vinson, Bonnie Webber, and Florian Wolf.

Trento and Marina del Rey, September 2008.

References

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good enough representations in language comprehension. *Current Directions in Psychological Science, 11*, 11–15.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language, 29*, 181–200.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*, 91–113.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243–281.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.

Poesio, M., & Artstein, R. (2005). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*. Ann Arbor, Michigan, pp. 76–83, Association for Computational Linguistics.

Poesio, M., Sturt, P., Artstein, R., & Filik, R. (2006). Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse Processes, 42*(2), 157–175.

Raskin, V. (1985). *Semantic mechanisms of humor*. Dordrecht and Boston: D. Reidel.

Su, S. P. (1994). *Lexical ambiguity in poetry*. London: Longman.

Véronis, J. (1998). A study of polysemy judgments and inter-annotator agreement. In *Proceedings of SENS-EVAL-1*.

Wagner, A., & Cacciaguidi-Fahy, S. (Eds.). (2006). *Legal language and the search for clarity*. Peter Lang AG.