Introduction to the Special Issue on Computational Anaphora Resolution

Ruslan Mitkov*
University of Wolverhampton

Branimir Boguraev†
IBM T. J. Watson Research Center

Shalom Lappin‡
King's College, London

Anaphora accounts for cohesion in texts and is a phenomenon under active study in formal and computational linguistics alike. The correct interpretation of anaphora is vital for natural language processing (NLP). For example, anaphora resolution is a key task in natural language interfaces, machine translation, text summarization, information extraction, question answering, and a number of other NLP applications.

After considerable initial research, followed by years of relative silence in the early 1980s, anaphora resolution has attracted the attention of many researchers in the last 10 years and a great deal of successful work on the topic has been carried out. Discourse-oriented theories and formalisms such as Discourse Representation Theory and Centering Theory inspired new research on the computational treatment of anaphora. The drive toward corpus-based robust NLP solutions further stimulated interest in alternative and/or data-enriched approaches. Last, but not least, application-driven research in areas such as automatic abstracting and information extraction independently highlighted the importance of anaphora and coreference resolution, boosting research in this area.

Much of the earlier work in anaphora resolution heavily exploited domain and linguistic knowledge (Sidner 1979; Carter 1987; Rich and LuperFoy 1988; Carbonell and Brown 1988), which was difficult both to represent and to process, and which required considerable human input. However, the pressing need for the development of robust and inexpensive solutions to meet the demands of practical NLP systems encouraged many researchers to move away from extensive domain and linguistic knowledge and to embark instead upon knowledge-poor anaphora resolution strategies. A number of proposals in the 1990s deliberately limited the extent to which they relied on domain and/or linguistic knowledge and reported promising results in knowledge-poor operational environments (Dagan and Itai 1990, 1991; Lappin and Leass 1994; Nasukawa 1994; Kennedy and Boguraev 1996; Williams, Harvey, and Preston 1996; Baldwin 1997; Mitkov 1996, 1998b).

The drive toward knowledge-poor and robust approaches was further motivated by the emergence of cheaper and more reliable corpus-based NLP tools such as part-of-speech taggers and shallow parsers, alongside the increasing availability of corpora and other NLP resources (e.g., ontologies). In fact, the availability of corpora, both raw and annotated with coreferential links, provided a strong impetus to anaphora resolu-

---

* School of Humanities, Language and Social Sciences, Stafford Street, Wolverhampton WV1 1SB, UK.
    E-mail: r.mitkov@wlv.ac.uk
† 30 Saw Mill River Road, Hawthorne, NY 10532, USA. E-mail: bkb@watson.ibm.com
‡ Department of Computer Science, King's College, The Strand, London WC2R 2LS, UK.
    E-mail: lappin@dcs.kcl.ac.uk

tion with regard to both training and evaluation. Corpora (especially when annotated) are an invaluable source not only for empirical research but also for automated learning (e.g., machine learning) methods aiming to develop new rules and approaches; they also provide an important resource for evaluation of the implemented approaches. From simple co-occurrence rules (Dagan and Itai 1990) through training decision trees to identify anaphor-antecedent pairs (Aone and Bennett 1995) to genetic algorithms to optimize the resolution factors (Orăsan, Evans, and Mitkov 2000), the successful performance of more and more modern approaches was made possible by the availability of suitable corpora.

While the shift toward knowledge-poor strategies and the use of corpora represented the main trends of anaphora resolution in the 1990s, there are other significant highlights in recent anaphora resolution research. The inclusion of the coreference task in the Sixth and Seventh Message Understanding Conferences (MUC-6 and MUC-7) gave a considerable impetus to the development of coreference resolution algorithms and systems, such as those described in Baldwin et al. (1995), Gaizauskas and Humphreys (1996), and Kameyama (1997). The last decade of the 20th century saw a number of anaphora resolution projects for languages other than English such as French, German, Japanese, Spanish, Portuguese, and Turkish. Against the background of a growing interest in multilingual NLP, multilingual anaphora/coreference resolution has gained considerable momentum in recent years (Aone and McKee 1993; Azzam, Humphreys, and Gaizauskas 1998; Harabagiu and Maiorano 2000; Mitkov and Barbu 2000; Mitkov 1999; Mitkov and Stys 1997; Mitkov, Belguith, and Stys 1998). Other milestones of recent research include the deployment of probabilistic and machine learning techniques (Aone and Bennett 1995; Kehler 1997; Ge, Hale, and Charniak 1998; Cardie and Wagstaff 1999; the continuing interest in centering, used either in original or in revised form (Abraços and Lopes 1994; Strube and Hahn 1996; Hahn and Strube 1997; Tetreault 1999); and proposals related to the evaluation methodology in anaphora resolution (Mitkov 1998a, 2001b). For a more detailed survey of the state of the art in anaphora resolution, see Mitkov (forthcoming).

The papers published in this issue reflect the major trends in anaphora resolution in recent years. Some of them describe approaches that do not exploit full syntactic knowledge (as in the case of Palomar et al.'s and Stuckardt's work) or that employ machine learning techniques (Soon, Ng, and Lim); others present centering-based pronoun resolution (Tetreault) or discuss theoretical centering issues (Kibble). Almost all of the papers feature extensive evaluation (including comparative evaluation as in the case of Tetreault's and Palomar et al.'s work) or discuss general evaluation issues (Byron as well as Stuckardt).

Palomar et al.'s paper describes an approach that works from the output of a partial parser and handles third person personal, demonstrative, reflexive, and zero pronouns, featuring among other things syntactic conditions on Spanish NP-pronoun noncoreference and an enhanced set of resolution preferences. The authors also implement several known methods and compare their performance with that of their own algorithm. An indirect conclusion from this work is that an algorithm requires semantic knowledge in order to hope for a success rate higher than 75%.

Soon, Ng, and Lim describe a C5-based learning approach to coreference resolution of noun phrases in unrestricted text. The approach learns from a small, annotated corpus and tackles pronouns, proper names, and definite descriptions. The coreference resolution module is part of a larger coreference resolution system that also includes sentence segmentation, tokenization, morphological analysis, part-of-speech tagging, noun phrase identification, named entity recognition, and semantic class determination (via WordNet). The evaluation is carried out on the MUC-6 and MUC-7 test

corpora. The paper reports on experiments aimed at quantifying the contribution of each resolution factor and features error analysis.

Stuckardt's work presents an anaphor resolution algorithm for systems where only partial syntactic information is available. Stuckardt applies Government and Binding Theory principles A, B, and C to the task of coreference resolution on partially parsed texts. He also argues that evaluation of anaphora resolution systems should take into account several factors beyond simple accuracy of resolution. In particular, both developer-oriented (e.g., related to the selection of optimal resolution factors) and application-oriented (e.g., related to the requirement of the application, as in the case of information extraction, where a proper name antecedent is needed) evaluation metrics should be considered.

Tetreault's contribution features comparative evaluation involving the author's own centering-based pronoun resolution algorithm called the Left-Right Centering algorithm (LRC) as well as three other pronoun resolution methods: Hobbs's naive algorithm (Hobbs 1978), BFP (Brennan, Friedman, and Pollard 1987), and Strube's S-list approach (Strube 1998). The LRC is an alternative to the original BFP algorithm in that it processes utterances incrementally. It works by first searching for an antecedent in the current sentence; if none can be found, it continues the search on the Cf-list of the previous and the other preceding utterances in a left-to-right fashion.

In her squib, Byron maintains that additional kinds of information should be included in an evaluation in order to make the performance of algorithms on pronoun resolution more transparent. In particular, she suggests that the pronoun coverage be explicitly reported and proposes that the evaluation details be presented in a concise and compact tabular format called standard disclosure. Byron also proposes a measure, the resolution rate, which is computed as the number of pronouns resolved correctly divided by the number of (only) referential pronouns.

Finally, in his squib Kibble discusses a reformulation of the centering transitions (Continue, Retain, and Shift), which specify the center movement across sentences. Instead of defining a total preference ordering, Kibble argues that a partial ordering emerges from the interaction among cohesion (maintaining the same center), salience (realizing the center as subject), and cheapness (realizing the anticipated center of a following utterance as subject).

The last years have seen considerable advances in the field of anaphora resolution, but a number of outstanding issues either remain unsolved or need more attention and, as a consequence, represent major challenges to the further development of the field (Mitkov 2001a). A fundamental question that needs further investigation is how far the performance of anaphora resolution algorithms can go and what the limitations of knowledge-poor methods are. In particular, more research should be carried out on the factors influencing the performance of these algorithms. One of the impediments to the evaluation or fuller utilization of machine learning techniques is the lack of widely available corpora annotated for anaphoric or coreferential links. More work toward the proposal of consistent and comprehensive evaluation is necessary; so too is work in multilingual contexts. Some of these challenges have been addressed in the papers published in this issue, but ongoing research will continue to address them in the near future.

References

Abraços, Jose and José Lopes. 1994. Extending DRT with a focusing mechanism for pronominal anaphora and ellipsis resolution. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING'94), pages 1128–1132, Kyoto, Japan.

Aone, Chinatsu and Scott Bennett. 1995. Evaluating automated and manual

acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (ACL'95), pages 122–129, Las Cruces, NM.

Aone, Chinatsu and Douglas McKee. 1993. A language-independent anaphora resolution system for understanding multilingual texts. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (ACL'93), pages 156–163, Columbus, OH.

Azzam, Saliha, Kevin Humphreys, and Robert Gaizauskas. 1998. Coreference resolution in a multilingual information extraction. In *Proceedings of a Workshop on Linguistic Coreference*, Granada, Spain.

Baldwin, Breck. 1997. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, Madrid, Spain.

Baldwin, Breck, Jeff Reynar, Mike Collins, Jason Eisner, Adwait Ratnaparki, Joseph Rosenzweig, Anoop Sarkar, and Srivinas Bangalore. 1995. Description of the University of Pennsylvania system used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference* (MUC-6), pages 177–191, Columbia, MD.

Brennan, Susan, Marilyn Friedman, and Carl Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics* (ACL'87), pages 155–162, Stanford, CA.

Carbonell, Jaime and Ralf Brown. 1988. Anaphora resolution: A multi-strategy approach. In *Proceedings of the 12th International Conference on Computational Linguistics* (COLING'88), volume 1, pages 96–101, Budapest, Hungary.

Cardie, Claire and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89, College Park, MD.

Carter, David M. 1987. *Interpreting Anaphors in Natural Language Texts*. Ellis Horwood, Chichester, UK.

Dagan, Ido and Alon Itai. 1990. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics* (COLING'90), volume 3, pages 1–3, Helsinki, Finland.

Dagan, Ido and Alon Itai. 1991. A statistical filter for resolving pronoun references. In Yishai A. Feldman and Alfred Bruckstein, editors, *Artificial Intelligence and Computer Vision*. Elsevier Science Publishers B.V. (North-Holland), Amsterdam, pages 125–135.

Gaizauskas, Robert and Kevin Humphreys. 1996. Quantitative evaluation of coreference algorithms in an information extraction system. Presented at Discourse Anaphora and Anaphor Resolution Colloquium (DAARC-1), Lancaster, UK. Reprinted in Simon Botley and Tony McEnery, editors, *Corpus-Based and Computational Approaches to Discourse Anaphora*. John Benjamins, Amsterdam, 2000, pages 143–167.

Ge, Niyu, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170, Montreal, Canada.

Hahn, Udo and Michael Strube. 1997. Centering-in-the-large: Computing referential discourse segments. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (ACL'97/EACL'97), pages 104–111, Madrid, Spain.

Harabagiu, Sanda and Steven Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of Conference on Applied Natural Language Processing/North American Chapter of the Association for Computational Linguistics* (ANLP-NAACL2000), pages 142–149, Seattle, WA.

Hobbs, Jerry. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Kameyama, Megumi. 1997. Recognizing referential links: An information extraction perspective. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53, Madrid, Spain.

Kehler, Andrew. 1997. Probabilistic coreference in information extraction. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing* (EMNLP-2), pages 163–173, Providence, RI.

Kennedy, Christopher and Branimir Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics* (COLING'96), pages 113–118, Copenhagen, Denmark.

Lappin, Shalom and Herbert Leass. 1994. An algorithm for pronominal anaphora

resolution. *Computational Linguistics*, 20(4):535–561.

Mitkov, Ruslan. 1996. Pronoun resolution: The practical alternative. Presented at the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC-1), Lancaster, UK. Reprinted in Simon Botley and Tony McEnery, editors, *Corpus-Based and Computational Approaches to Discourse Anaphora*. John Benjamins, Amsterdam, 2000, 189–212.

Mitkov, Ruslan. 1998a. Evaluating anaphora resolution approaches. In *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium* (DAARC-2), Lancaster, UK.

Mitkov, Ruslan. 1998b. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics* (COLING'98/ACL'98), pages 869–875, Montreal, Canada.

Mitkov, Ruslan. 1999. Multilingual anaphora resolution. *Machine Translation*, 14(3–4):281–299.

Mitkov, Ruslan. 2001a. Outstanding issues in anaphora resolution. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*. Springer, Berlin, pages 110–125.

Mitkov, Ruslan. 2001b. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. *Applied Artificial Intelligence: An International Journal*, 15:253–276.

Mitkov, Ruslan. Forthcoming. *Anaphora Resolution*. Longman, Harlow, UK.

Mitkov, Ruslan, Lamia Belguith, and Malgorzata Stys. 1998. Multilingual robust anaphora resolution. In *Proceedings of the Third International Conference on Empirical Methods in Natural Language Processing* (EMNLP-3), pages 7–16, Granada, Spain.

Mitkov, Ruslan and Malgorzata Stys. 1997. Robust reference resolution with limited knowledge: High precision genre-specific approach for English and Polish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP'97), pages 74–81, Tzigov Chark, Bulgaria.

Mitkov, Ruslan and Catalina Barbu. 2000. Improving pronoun resolution in two languages by means of bilingual corpora. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference* (DAARC 2000), pages 133–137, Lancaster, UK.

Nasukawa, Tetsuya. 1994. Robust method of pronoun resolution using full-text information. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING'94), pages 1157–1163, Kyoto, Japan.

Orăsan, Constantin, Richard Evans, and Ruslan Mitkov. 2000. Enhancing preference-based anaphora resolution with genetic algorithms. In *Proceedings of NLP-2000*, pages 185–195, Patras, Greece.

Rich, Elaine and Susann LuperFoy. 1988. An architecture for anaphora resolution. In *Proceedings of the Second Conference on Applied Natural Language Processing* (ANLP-2), pages 18–24, Austin, TX.

Sidner, Candace. 1979. Toward a computational theory of definite anaphora comprehension in English. Technical Report AI-TR-537, MIT, Cambridge, MA.

Strube, Michael. 1998. Never look back: An alternative to centering. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics* (COLING'98/ACL'98), pages 1251–1257, Montreal, Canada.

Strube, Michael and Udo Hahn. 1996. Functional centering. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (ACL'96), pages 270–277, Santa Cruz, CA.

Tetreault, Joel. 1999. Analysis of syntax-based pronoun resolution methods. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL'99), pages 602–605, College Park, MD.

Williams, Sandra, Mark Harvey, and Keith Preston. 1996. Rule-based reference resolution for unrestricted text using part-of-speech tagging and noun phrase parsing. In *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium* (DAARC-1), pages 441–456, Lancaster, UK.