Machine Learning of Entity Recognizers for Modular Retargetable Natural Language Processing

Claire Grover · Ewan Klein · Chris Manning · Katja Markert · Malvina Nissim

1 OBJECTIVES

Our overall aim is to provide high-quality, general purpose entity recognition components. Specific objectives include:

- To generalise the class of entities that can be recognized within a wide variety of domains.
- To make innovative use of machine-learning techniques to enable the use of unnannotated corpora as training material.
- To achieve portability of components through modular XML-based architecture grounded in the LTG's XML toolset.
- To improve the results of entity recognition by carrying out Metonymy analysis and Word Sense Disambiguation.
- To demonstrate the utility of generalised entity recognition in the Question-Answering application area.

2 SUMMARY

This project will extend and generalise the current state-of-the-art in entity recognition so that it can become a reliable enabling technology for a wide range of applications and higher-level NLP tasks. Our goal is to develop the means to recognize and classify a much wider range of entities than are traditionally treated, and to develop techniques which can be applied in a wide range of text types. Rapid adaptation to new domains is a priority and this requires modularity in the processing pipeline as well as the innovative use of machine-learning techniques to enable the use of unnannotated corpora as training material wherever possible.

3 BENEFICIARIES

Named Entity Recognition (NER) technology is an essential ingredient in Information Extraction (IE) and there are a number of commercial IE systems currently available. Clearly, our improvements on current NER technology would be of immediate benefit to builders of such systems. However, more broadly it has the potential to support a variety of applications which require domain-specific semantic information while not requiring full Natural Language Understanding. Such an application might, for example, assist participants in an instant messaging session by triggering a database retrieval on detecting relevant named entities in the discourse [cf. Adams, 2002].

A somewhat different application is ontology acquisition. With the advent of the semantic web and the related technology of web services, there is a growing belief that ontologies are crucial for business-to-business transactions. NER can play a significant role in the semi-automatic acquisition of ontological concepts and relations from domain-specific corpora [Kietz et al., 2000].

Question Answering (QA) is the application area we foresee as a main beneficiary since this requires an open class of entities to be recognized. High quality entity recognition will also provide a foundation for more linguistically-oriented NLP activities, which will yield further improvements in performance in IE and QA.

4 TRACK RECORD

The groups at Edinburgh and Stanford have extensive experience with information extraction and related technologies, including Named Entity Recognition, Word Sense Disambiguation, and XML processing. They have not previously worked together, but see the Edinburgh-Stanford link as an effective way to combine their respective strengths, and in particular coordination between two sites will drive the goal of producing modular, reusable components which can be flexibly integrated into larger systems.

Claire Grover. Grover's main LTG activities in the past few years have been in the area of XML tools, Named Entity Recognition, Information Extraction and robust deep and shallow parsing. In the TTT project she cooperated on the development of the LT TTT toolset [Grover et al., 2000b] which was the technology behind the LTG's highly successful NER system entered in the MUC7 competition [Mikheev et al., 1998]. In the CROSSMARC project, she has been working on multi-lingual information extraction from web pages [Grover et al., 2002]. This has involved the recognition of non-standard sets of entities and the challenge of extracting information from deeply-structured HTML documents, for which the LTG's XML tools have proved invaluable. In a project funded by Edify, she has been adapting NER technology to identify names and addresses in customer emails to assist in the automatic processing of email enquiries. In DISP, which is a more linguistically oriented NLP project, Grover has used the LT TTT tools to perform a foundation of tokenisation and entity recognition in processing Medline abstracts [Grover and Lascarides, 2001]. The resultant corpus was then parsed using a variety of shallow and deep robust techniques with the aim of automatically acquiring lexical semantic information [Grover et al., submitted].

Ewan Klein. Klein has had over 20 years experience working on research issues in natural language, and has been involved in numerous academic-industrial collaborative projects on the development of language technology. Research topics include computational approaches to phonology, syntax and semantics, multimodal interfaces, controlled languages for hardware design [Holt et al., 1999, Holt and Klein, 1999, Grover et al., 2000a, Holt et al., 2000], and dialogue with intelligent systems. He has had management responsibilities on numerous projects concerned with natural language processing and semantics, with funding from the EU ESPRIT programme and from the UK EPSRC research council.

Chris Manning. Manning has been involved for a decade in probabilistic and machine learning approaches to natural language processing, and is a coauthor of the leading textbook on statistical NLP [Manning and Schütze, 1999]. He has longterm interests in lexical acquisition [Manning, 1993], and has worked on commercial information extraction systems [Manning, 1998]. His recent work includes researching probabilistic and other broad coverage parsing [Klein and Manning, 2001], unsupervised grammar induction [Klein and Manning, 2002b], conditional log-linear models, including for part-of-speech tagging [Toutanova and Manning, 2000, Klein and Manning, 2002a], parse selection models based on the Redwoods HPSG treebank [Oepen et al., 2002, Toutanova and Manning, 2002], and word sense disambiguation [Klein et al., 2002]. He also has extensive experience working with XML and related technologies through his work on computational lexicography for indigenous languages [Jansz et al., 2000, Manning et al., 2001]. His recent work has been funded by the US government (NSF and ARDA) and various corporations (IBM, NTT, and NHK).

Katja Markert. Markert has worked on metonymy resolution and Word Sense Disambiguation [Markert and Hahn, 1997, 2002, Markert and Nissim, 2002b,a], discourse phenomena [Markert et al., 1996, Hahn et al., 1996b,a] and the semantic interpretation of grammatical relations [Romacker et al., 1999]. She has experience with symbolic, knowledge-intensive methods as well as with semantic annotation and supervised machine learning. Her recent work has been funded by the German as well as the British Government.

Malvina Nissim. Nissim has broad experience of working on lexical semantics [Nissim, to appear] and discourse phenomena such as anaphora [Nissim and Sansò, 1999, Nissim et al., 1999], and their interaction under several factors [Nissim, 2001, 2002]. This research—based on a strong linguistics background—has equipped her with experience of working with large corpora and with semantic annotation. Recently, she has been involved in large-scale metonymy resolution, and has developed skills in working with XML tools and in supervised machine learning [Markert and Nissim, 2002b,a]. Her recent work has been funded by the Italian (MURST/MIUR) as well as the British Government (ESRC).

5 PROPOSED ACTIVITIES

5.1 Background

Named Entity Recognition (NER) is best known as the first step in the Information Extraction (IE) task, which was particular codified in the US DARPA-sponsored MUC competitions. The overall aim is to extract 'facts' from text in the form of multi-place relations and NER provides the entities which fill the argument slots. In MUC, the entire process was driven by pre-defined templates, which effectively provided an ontology of the entities that should be recognized, making the class of types of entities a closed one. For instance, in MUC7 [Chinchor, 1998], the task was to process newspaper articles and fill templates concerning airplane disasters. The entities to be recognized were a small fixed set of basic entities (Person, Organization, Location, Money, Date and Time), and teams variously built custom recognizers for these components, ranging from hand-tuned recognizers using extensive lists of names and rules, to machine learning components based on extensive supervised training data.

There is an older tradition in Natural Language Processing (NLP), still much used in applications like spoken language systems, where people build *semantic grammars*: phrase-structure grammars where the categories are no longer grammatical categories like noun phrase and verb phrase but semantic categories like Flight, Flight-Destination, City [Allen, 1995, 332–334]. These grammars work very well in limited domains, but suffer from brittleness, duplication of grammatical structure in different semantic categories, and lack of portability. Nevertheless, it has been insufficiently appreciated that the low level categories of semantic grammars are equivalent to Named Entities, and that this work dramatically shows how in particular applications there is an unbounded variety of application-specific named entities of interest, often even representing different classifications of the same denotational range.

Indeed, this issue becomes even clearer when one considers the range of current and possible applied uses of NER technology. There are many activities which are variants of the NER task of identifying substrings in text and labeling them as tokens of various semantic kinds:

- In a current LTG IE project, CROSSMARC [Grover et al., 2002], the domain is laptops from computer vendor websites and the entities to be recognized include e.g. Manufacturer, Model, Processor, Price.

- In the rapidly growing related field of Question Answering (QA), NER is standardly used as a component in the processing of both questions and the text which supplies the answers [Vorhees and Tice, 2001]. Here, since the topic area is potentially unlimited, a more open class of entities, much broader than the common MUC entities, is potentially very useful, assuming that they can be found reliably.

- In the LTG's DISP project, which performed higher level NLP on Medline abstracts, we included in the pre-processing an Entity Recognition component to identify entities such as drug names, formulae, dosage indications etc. [Grover and Lascarides, 2001, Grover et al., submitted].

- The LTG project MAScARA [Markert and Nissim, 2002b] aims to use learning techniques for the automatic annotation of metonymic usage and a prerequisite for the task is accurate Entity Recognition that identifies e.g. *Berlin* as a location in order that its capital-for-government metonymic use can be computed in examples such as *whether Berlin would meet demands*.

Activities such as these are broader than MUC NER, but they are essentially the same, and certainly the tools and techniques needed are the same. We use the term Entity Recognition to cover the broader class. The same technology can support both language engineering tasks such as IE, Question Answering, and Text Mining as well as tasks which are part of research in computational linguistics.

5.2 Aims

The overall aim of this project is to extend and generalise the current state-of-the-art in entity recognition so that it can become a reliable enabling technology for a wide range of applications and higher-level NLP tasks. Our goal is to develop the means to recognize and classify a much wider range of entities than are traditionally treated, and to develop techniques which can be applied in a wide range of text types. Rapid

adaptation to new domains is a priority and this requires modularity in the processing pipeline as well as the use of unsupervised learning techniques wherever possible.

Bespoke information extraction (IE) systems deal effectively with texts in fixed domains, but they are time-consuming and costly to build. Our goals require the use of machine-learning methods for learning entities on the basis of unannotated corpora and we will incorporate these methods into the existing pipeline architecture provided by the LT TTT and LT XML toolsets. The techniques we develop will be immediately useable in IE tasks where the entities required are highly constrained but we aim for our techniques to be equally useful in more linguistically-oriented, less constrained tasks. For this reason we build on current work in the DISP and MAScARA projects plus Word Sense Disambiguation work at Stanford to compute semantic annotations on top of Entity Recognition output.

The project builds on expertise at the two sites: Edinburgh LTG's expertise and highly successful tools in the field of Named Entity Recognition (NER) and the Stanford NLP group's expertise in probabilistic and machine learning techniques for robust NLP.

5.3 Methodology

5.3.1 Machine Learning for Entity Recognition

Although standard NER systems, such as the LTG's contribution to MUC7 [Mikheev et al., 1998], can deal effectively with texts in fixed domains where the application clearly defines the entities to be recognized, porting to new domains and building entity recognition systems for less constrained applications is a considerable challenge. We believe that it is necessary to incorporate machine learning techniques for automatically acquiring patterns from data.

When the entity set is fixed, as with the MUC entities, recent work has shown that considerable mileage can be gained from supervised learning techniques [Freitag and McCallum, 2000]. For example, Collins [2002] describes a method which uses a maximum entropy word tagger to label words according to whether or not they belong to a Named Entity. The tagger was trained on 1 million words of web page data, annotated for Named Entities by AT&T. Interesting though this approach is, the reliance on large amounts of annotated data for supervised training is problematic since there is commonly very little annotated data available for applied tasks, or even for MUC entities. For our purposes, unannotated corpora are preferable since we are interested in wider Entity Recognition techniques applicable to a variety of text types where the entities could be e.g., product names, drug names, mathematical formulae, legal terms, etc. We thus propose the use of cotraining and other bootstrapping methods to learn instances of named entities.

Some existing work has used unannotated corpora. One method is to start with a small number of *seed* rules which are then generalized so as to cover a much wider data set [cf. Collins and Singer, 1999, Yangarber and Grishman, 2000, Riloff and Jones, 1999]. Such methods exploit the idea of *cotraining* where two complementary sources of information (here the content of a slot, and the environments in which it occurs) are used to constrain each other. This method has enjoyed some success, but it is somewhat unstable: the bounds of the concept desired by the human is unclear, and such systems commonly have to be stopped at just the right moment in their learning process. We plan to approach solving this problem by contrasting desired concepts with *anti-concepts* that will learn non-examples.

An interesting middle ground is beginning from more limited forms of supervision such as incomplete lists of examples. Smarr and Manning [2002] show how one can very effectively work from lists (simple lexicons) of words in a class, which are much more commonly available than annotated texts, to build models of what terms of a certain entity tend to look like. We believe that these two approaches can be productively combined in the work proposed here: an initial model of what entities look like can bootstrap the cotraining process of finding indicative contexts.

Both groups have existing sets of data suitable for NER research collected in recent projects, and access to MUC data from the Linguistic Data Consortium. A first task would be to agree on particular domains for experimentation and to identify any additional data resources that should be aquired. We do not anticipate devoting any researcher time to corpus annotation except for the annotation of testing material wherever evaluation requires it. Another task in the workplan will be agreeing on evaluation methodologies for each of the subtasks, where the methodologies will differ depending on the subtask and corpora resources.

5.3.2 Modular Portable Components

NER is an *enabling technology* which provides foundations upon which further levels of processing can be built. An important failure of much NLP work is the production of one-off custom architectures and complete systems, which at best see use by a few people or sites beside the creators. A central success of part-of-speech tagging systems is that they have provided a simple reusable component. So far, NER has achieved that only for the kind of simple, fixed tagsets explored in the MUC competition. To achieve this goal more broadly, we firstly need the unsupervised and semi-supervised methods referred to above, and secondly we need an architecture which will support modular component-based processing and reuse. For the latter goal, we will make use of XML representations, the emerging *lingua franca* of structured data representation. We will incorporate these methods into the LTG's existing toolsets, which use XML representations throughout. The pipeline architecture provided in this framework is particularly suitable for modularising tasks and for exploring the interactions between sub-tasks. As with the LTG's MUC7 system, hybrid pipelines can be built which interleave rule-based modules with probabilistic and machine-learning components, thereby facilitating experimentation with different combinations of techniques.

One of the strengths of the pipeline approach is the way in which knowledge of the text, represented as XML annotations, is incrementally computed by the use of successive modules. This gives a clear framework in which underspecification can be used, so that an early module can provide partial information which can be further instantiated by a later module. The architecture is highly flexible so that, for example, feedback loops can be incorporated to allow a second pass through a module in order to benefit from computations that may have occurred subsequent to the first pass.

5.3.3 Entity Recognition Subtasks

The traditional NER task can be regarded as being composed of two distinct parts: identification of the sub-strings in texts which are named entities (*segmentation*) and assigning the named entities to their semantic classes (*classification*). Many NER systems, including the LTG's own MUC7 system, perform these two tasks simultaneously, but approaching them separately leads to greater conceptual clarity and to better performance on each task (see for example Wacholder et al. [1997] and Mikheev [1999] for segmentation and Smarr and Manning [2002] for classification). In this project, therefore, we advocate a strong conceptual separation of the two parts. In this way our goals of modularity and portability will also be addressed and the conceptual separation of the two parts can be mirrored in the implementation using our pipelined XML approach which encourages the decomposition of a task into clear sub-modules. Furthermore, the exploration of appropriate machine learning techniques for the NER task will benefit from the separation of the two parts since the techniques that are appropriate for the segmentation may well be different from the best techniques for classification.

Our aim of widening the task to include a large variety of entities means that segmentation must identify not only standard named entities which are typically capitalized, but also terms which have no capitalisation to function as a cue. In this case, the segmentation task is harder and must rely on learning techniques which examine both word internal evidence and external evidence such as contextual clues and the phrase boundaries provided by chunkers etc. In Medline abstracts the names of drugs and diseases (e.g. *esophageal dysphagia*, *tetracycline*) are typically all lower case but instances of mixed capitalization/lower case are also very frequent (e.g. *nonP-fimbriated Escherichia coli*, *Lyme disease*). In less technical text, for example newspaper articles, there are instances of entities which appear in lower case but which can act in the same way as traditional named entities in filling slots in templates. For example, in *he entertained the whole table with his witty remarks*, *the table* is not capitalized but is an entity that might need to be recognized. In this case, Word Sense Disambiguation can be applied to identify that the correct interpretation of *table* in this context is the concrete object interpretation (as opposed to the diagram interpretation), and therefore *the table* can be marked as an entity.

5.3.4 Semantic Annotation

As a step immediately subsequent to Entity Recognition, we will improve and extend the work on metonymy resolution currently being investigated in the LTG's MAScARA project. The process of identifying a metonymic use of a named entity (e.g. *We drank Bordeaux*, where the region name is used to refer to its product), can be thought of as a processing stage which refines or corrects the classification assigned by the

entity recognizer. In the example of *the table* above, metonymy resolution will compute a object-for-people metonymic interpretation. Markert and Nissim [2002a] describe the disambiguation of metonymic uses of location entities using supervised classification algorithms. Previous studies have shown metonymy to be pervasive: Markert and Hahn [2002] found a metonymy in 15% of all utterances in 27 German magazine texts. Distinguishing names of artefacts (like "BMW") from names of organisations (like "BMW") as part of entity recognition can help make more fine-grained searches in document collections — we found that of 100 random occurrences of "BMW" in the British National Corpus, about 50% referred to the firm and 50% to cars or motorcycles. Since metonymic readings are very frequent, reliable resolution of these ambiguities would make a significant contribution to general purpose application areas such as Question Answering. Our aim is to extend the range of metonymic types that we currently handle and to apply some of the semi-supervised learning techniques used for classification to the metonymy resolution task.

Named entity recognition connects closely to Word Sense Disambiguation. There has been more focus in NER on recognizing multiword units, but this seems like an underexplored issue in word sense disambiguation (given an expression like *golden handshake*, one should be identifying it as a collocational unit and determining its meaning, rather than trying to disambiguate which sense of *handshake* is being invoked).

Both research groups have expertise in Word Sense Disambiguation, the discovery of metonymies, and the use of WordNet and other semantic resources. We will investigate using our tools and techniques to automatically perform Word Sense Disambiguation and other computations on semantic information. The virtue of this route is that semantic annotation is both an instance of a higher-level linguistic process which depends on reliable Entity Recognition and at the same time a processing stage which can be performed using some of the same machine learning methods as Entity Recognition.

We suggest that one application area that might ultimately benefit most from our combined work on Entity Recognition and semantic annotation would be Question Answering. NER is an accepted subcomponent of many of the systems competing in the TREC-9 question answering competition and many also use information from Wordnet in analysing both questions and candidate answers. Stallard [1993] reports a 27% performance improvement when metonymy resolution was included in a QA system. Question Answering is becoming an active research area in ICCS so we expect there to be collaboration opportunities in the near future.

5.3.5 Overview of Tasks

The following list divides the proposed work into tasks. We have not included start and end times for these tasks since this constitutes only a very schematic workplan which we propose to elaborate more fully at the kick-off meeting for the project. There is some temporal order inherent in the work since some tasks need to build on the output of others.

1. **Corpora selection and preparation**: obtain corpora from diverse domains containing annotations appropriate for the degree of supervision anticipated for machine learning.
2. **Segmentation**: evaluate state-of-the-art in segmentation and implement flexible segmentation methods utilising XML tools and combining rule-based and machine-learning techniques.
3. **Classification**: evaluate existing research in Stanford and elsewhere and extend and improve methods, with the particular aim of using unannotated corpora.
4. **Word Sense Disambiguation**: build on existing experience in both sites to implement this module.
5. **Metonymy Resolution**: extend Edinburgh's existing metonymy resolution to a wide-range of metonymy types and experiment with machine-learning techniques to improve performance.
6. **Integration with XML tools**: integrate individual components into the XML pipeline architecture to create modular and reusable programs.
7. **Evaluation**: for each task design an appropriate evaluation methodology and acquire or create testing materials.

5.4 Relevance to Beneficiaries

The LTG's current NER technology has already been commercialised in improvements to search engine technology by Infogistics. Moreover, the LTG is using it in a collaborative project with Edify and in the EU funded IE project, CROSSMARC, which aims to produce a close-to-market system for multi-lingual IE in the e-retail domain.

IE is being taken up as a niche technology by companies such as Whizbang! Labs or Knowmadic, but has a much wider commercial potential, in our view. According to a recent Datamonitor (`www.datamonitor.com` report, natural language understanding applications are predicted to have a cumulative annual growth rate of around 70% during the period 2000–2006. Yet as a technology, full NL understanding is not mature enough for general industrial use within this time window. IE is the obvious alternative, since it represents a significant step towards domain-specific understanding while remaining very robust.

The work we propose in this project will advance our technology significantly by addressing the need for swift adaptation and portability. In directing our efforts at the QA task we are being quite ambitious, and the exploitation possibilities would be longer term. However, the extensions to our existing technology would produce immediate improvements to the more constrained task of IE.

The development of modular, portable techniques for entity recognition will allow for the rapid adaptation of our systems to new and specialised domains. The results of this project will feed into any existing and future research projects which use large corpora and specialised domains. For example, a current EPSRC-funded summarisation project (SUM) uses a corpus of appeal court rulings from the legal domain. Entity recognition of legal terms as well as standard names is a necessary component of the processing of the corpus. The DISP project has already made some progress with entity recognition in the medical domain where the technical terms are highly complex. In another Edinburgh-Stanford link proposal (Webber, Altman and Grover), a variety of XML-based NLP techniques will be used to examine whether this can improve the performance of literature-based medical informatics applications. Results of the current project are likely to be directly relevant to the implementation of the Webber, Altman and Grover project.

The LTG's technology has been used by PhD students within ICCS and more recently by a group of MSc students working collaboratively to build a Question Answering System. While it is usable in its current form, we would welcome an opportunity to make it more accessible than it currently is, for researchers and students at both Edinburgh and Stanford.

5.5 Dissemination and Exploitation

As with most academic research, the primary vehicle for dissemination will be through the academic literature. We will seek to publicize our results through conference presentations as well as in peer-reviewed journal articles. We will also be sensitive to the fact that much of the recent explosion of activity within the Semantic Web area is dependent on XML-based (or XML-inspired) standards, and we would like to explore the potential for contributing to standardization in relevant aspects of Information Extraction technology. There are opportunities for NER and IE to be used to provide semi-automatic annotation of the sort of ontological information desired for Semantic Web work, so this would be a promising avenue for exploitation.

The most innovative aspect of the proposed research is machine learning using unannotated corpora and, assuming that our research is successful, the techniques we develop will be directly relevant to commercial interests. Ewan Klein is currently seconded to head the research laboratory of Edify in Edinburgh, and applications that Edify is developing in the area of customer contact management would benefit from our results.

REFERENCES

D J Adams. *Programming Jabber*. O'Reilly, Sebastopol, 2002.

James Allen. *Natural Language Understanding*. Benjamin Cummings, Redwood City, CA, 1995.

Nancy A. Chinchor, editor. *Proceedings of Seventh Message Understanding Conference (MUC–7)*, Fairfax, Virginia, 1998. URL `http://www.muc.saic.com/proceedings/muc_7_toc.html`.

M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

Michael Collins. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In *ACL 40*, 2002. URL `http://www.research.att.com/~mcollins/`.

Dayne Freitag and Andrew McCallum. Information extraction with HMM structures learned by stochastic optimization. In *Proceedings of AAAI*, pages 584–589, 2000. URL `citeseer.nj.nec.com/freitag00information.html`.

Claire Grover, Alexander Holt, Ewan Klein, and Marc Moens. Designing a controlled language for interactive model checking. In *CLAW 2000: Proceedings of the Third International Workshop on Controlled Language Applications, 29-30 April 2000*, pages 90–104, Seattle, April 2000a.

Claire Grover, Ewan Klein, Maria Lapata, and Alex Lascarides. NLP tools for processing and interpreting biomedical language. In *Submitted to the 2nd Workshop on NLP and XML (NLPXML-2002)*, submitted.

Claire Grover and Alex Lascarides. XML-based data preparation for robust deep parsing. In *Proceedings of the Joint EACL-ACL Meeting (ACL-EACL 2001)*, 2001.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. LT TTT - a flexible tokenisation tool. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, 2000b.

Claire Grover, Scott McDonald, Donnla Nic Gearailt, Vangelis Karkaletsis, Dimitra Farmakiotou, Georgios Samaritakis, Georgios Petasis, Maria Teresa Pazienza, Michele Vindigni, Frantz Vichot, and Francis Wolinski. Multilingual XML-based named entity recognition for e-retail domains. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 2002.

Udo Hahn, Katja Markert, and Michael Strube. A conceptual reasoning approach to textual ellipsis. In *Proceedings of the 12th European Conference on Artificial Intelligence*, pages 572–576, 1996a.

Udo Hahn, Michael Strube, and Katja Markert. Bridging textual ellipses. In *Proceedings of the 16th International Conference on Computational Linguistics*, 1996b.

Alexander Holt and Ewan Klein. A semantically-derived subset of English for hardware verification. In *37th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference: 20–26 June 1999, University of Maryland, College Park, Maryland, USA*, pages 451–456. Association for Computational Linguistics, 1999. ISBN 1-55860-609-2.

Alexander Holt, Ewan Klein, and Claire Grover. Natural language for hardware verification: Semantic interpretation and model checking. In *Proceedings of ICoS-1: Inference in Computational Semantics, August 15, 1999*, pages 133–137, Amsterdam, August 1999. Institute for Logic, Language and Computation, University of Amsterdam.

Alexander Holt, Ewan Klein, and Claire Grover. Natural language specifications for hardware verification. *Journal of Language and Computation*, 1(2):275–282, 2000.

Kevin Jansz, Wee Jim Sng, Nitin Indurkhya, and Christopher Manning. Using XSL and XQL for efficient customised access to dictionary information. In *Proceedings of AusWeb 2000, the Sixth Australian World Wide Web Conference*, pages 167–181, 2000.

Jörg-Uwe Kietz, Raphael Volz, and Alexander Maedche. Extracting a domain-specific ontology from a corporate intranet. In Claire Cardie, Walter Daelemans, Claire Nédellec, and Erik Tjong Kim Sang, editors, *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop, Lisbon, 2000*, pages 167–175. Association for Computational Linguistics, Somerset, New Jersey, 2000. URL `citeseer.nj.nec.com/319456.html`.

Dan Klein and Christopher D. Manning. Parsing with treebank grammars: Empirical bounds, theoretical models, and the structure of the Penn treebank. In *ACL 39/EACL 10*, pages 330–337, 2001.

Dan Klein and Christopher D. Manning. Conditional structure versus conditional estimation in NLP models. In *EMNLP 2002*, 2002a.

Dan Klein and Christopher D. Manning. A generative constituent-context model for improved grammar induction. In *ACL 40*, 2002b.

Dan Klein, Kristina Toutanova, H. Tolga Ilhan, Sepandar D. Kamvar, and Christopher D. Manning. Combining heterogeneous classifiers for word-sense disambiguation. In *Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, 2002. At ACL 40.

Christopher Manning. Rethinking text segmentation models: An information extraction case study. Technical Report SULTRY-98-07-01, University of Sydney, 1998.

Christopher D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *ACL 31*, pages 235–242, 1993.

Christopher D. Manning, Kevin Jansz, and Nitin Indurkhya. Kirrkirr: Software for browsing and visual exploration of a structured Warlpiri dictionary. *Literary and Linguistic Computing*, 16:123–139, 2001.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Boston, MA, 1999.

Katja Markert and Udo Hahn. On the interaction of metonymies and anaphora. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, volume 2, pages 1010–1015, 1997.

Katja Markert and Udo Hahn. Understanding metonymies in discourse. *Artificial Intelligence*, 135(1/2): 145–198, 2002.

Katja Markert and Malvina Nissim. Metonymy resolution as a classification task. In *To appear in Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 2002a.

Katja Markert and Malvina Nissim. Towards a corpus annotated for metonymies: the case of location names. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 2002b.

Katja Markert, Michael Strube, and Udo Hahn. Inferential realization constraints on functional anaphora in the centering model. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pages 609–614, 1996.

A. Mikheev. A knowledge-free method for capitalized word disambiguation. In *In Proceedings of the 37th Annual Meeting of the ACL*, pages 159–166, 1999.

Andrei Mikheev, Claire Grover, and Marc Moens. Description of the LTG system used for MUC-7. In Chinchor [1998]. URL `http://www.muc.saic.com/proceedings/muc_7_toc.html`.

Malvina Nissim. Underlying relations in genitives and bridging. In Eniko Nemeth, editor, *Pragmatics in 2000*. International Pragmatics Association, 2001.

Malvina Nissim. *Bridging Definites and Possessives: Distribution of Determiners in Anaphoric Noun Phrases*. PhD thesis, Dipartimento di Linguistica, Università di Pavia, 2002.

Malvina Nissim. Lexical constraints on the use of the English neuter possessive determiner. In Malvina Nissim and Andrea Sansò Cristiano Broccias, editors, *Semantics Inside and Outside the Clause*. to appear.

Malvina Nissim and Andrea Sansò. Discarding expectations when building bridges: The need for a revision procedure. In *Proceedings of VEXTAL: Venezia per il Trattamento Automatico delle Lingue*, Venezia, Italy, November 22-24 1999.

Malvina Nissim, Andrea Sansò, and Claudia Soria. Towards a compositional frame semantics. In *Proceedings of ECCS99, European Conference on Cognitive Science*, Siena, Italy, October 27-30 1999.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christoper D. Manning, Dan Flickinger, and Thorsten Brants. The LinGO Redwoods treebank: Motivation and preliminary applications. In *COLING 19*, 2002.

Ellen Riloff and Rosie Jones. Learning Dictionaries for Information Extraction by Multi-level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 1044–1049. The AAAI Press/MIT Press, 1999. URL `citeseer.nj.nec.com/riloff99learning.html`.

Martin Romacker, Katja Markert, and Udo Hahn. Lean semantic interpretation. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1999.

Joseph Smarr and Christopher Manning. Automatic classification of previously unseen proper noun phrases. submitted, 2002.

David Stallard. Two kinds of metonymy. In *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*, pages 87–94, Columbus, Ohio, 1993.

Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *EMNLP/VLC 2000*, pages 63–70, 2000.

Kristina Toutanova and Christopher D. Manning. Feature selection for a rich HPSG grammar using decision trees. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, 2002.

Ellen M. Vorhees and Dawn M. Tice, editors. *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, 2001.

N. Wacholder, Y. Ravin, and M. Choi. Disambiguation of proper names in text. In *Proceedings of Fifth Conference on Applied Natural Language Processing*, pages 202–208, 1997.

Roman Yangarber and Ralph Grishman. Machine learning of extraction patterns from unannotated corpora. In *Proceedings of 14th European Conference on Artificial Intelligence (ECAI 2000), Workshop on Machine Learning for Information Extraction*, Berlin, 2000.