Named Entity Recognition without Gazetteers

Andrei Mikheev*, Marc Moens and Claire Grover
HCRC Language Technology Group,
University of Edinburgh,
2 Buccleuch Place, Edinburgh EH8 9LW, UK.
mikheev@harlequin.co.uk M.Moens@ed.ac.uk C.Grover@ed.ac.uk

Abstract

It is often claimed that Named Entity recognition systems need extensive gazetteers—lists of names of people, organisations, locations, and other named entities. Indeed, the compilation of such gazetteers is sometimes mentioned as a bottleneck in the design of Named Entity recognition systems.
We report on a Named Entity recognition system which combines rule-based grammars with statistical (maximum entropy) models. We report on the system's performance with gazetteers of different types and different sizes, using test material from the MUC–7 competition. We show that, for the text type and task of this competition, it is sufficient to use relatively small gazetteers of well-known names, rather than large gazetteers of low-frequency names. We conclude with observations about the domain independence of the competition and of our experiments.

1 Introduction

Named Entity recognition involves processing a text and identifying certain occurrences of words or expressions as belonging to particular categories of Named Entities (NE). NE recognition software serves as an important preprocessing tool for tasks such as information extraction, information retrieval and other text processing applications.

What counts as a Named Entity depends on the application that makes use of the annotations. One such application is document retrieval or automated document forwarding: documents annoted with NE information can be searched more accurately than raw text. For example, NE annotation allows you to search for all texts that mention the *company* "Philip Morris", ignoring documents about a possibly unrelated *person* by the same name. Or you can have all documents forwarded to you about a person called "Gates", without receiving documents about things called gates. In a document collection annotated with Named Entity information you can more easily find documents about Java the programming language without getting documents about Java the country or Java the coffee.

Most common among marked categories are names of people, organisations and locations as well as temporal and numeric expression. Here is an example of a text marked up with Named Entity information:

```
<ENAMEX TYPE='PERSON'>Flavel
Donne</ENAMEX> is an analyst with <ENAMEX
TYPE='ORGANIZATION'>General Trends
</ENAMEX>, which has been based in <ENAMEX
TYPE='LOCATION'>Little Spring</ENAMEX> since
<TIMEX TYPE='DATE'>July 1998</TIMEX>.
```

In an article on the Named Entity recognition competition (part of MUC-6) Sundheim (1995) remarks that "common organization names, first names of people and location names can be handled by recourse to list lookup, although there are drawbacks" (Sundheim 1995: 16). In fact, participants in that competition from the University of Durham (Morgan et al., 1995) and from SRA (Krupka, 1995) report that gazetteers did not make that much of a difference to their system. Nevertheless, in a recent article Cucchiarelli et al. (1998) report that one of the bottlenecks in designing NE recognition systems is the limited availability of large gazetteers, particularly gazetteers for different languages (Cucchiarelli et al. 1998: 291). People also use gazetteers of very different sizes. The basic gazetteers in the Isoquest system for MUC-7 contain 110,000 names, but Krupka and Hausman (1998) show that system performance does not degrade much when the

* Now also at Harlequin Ltd. (Edinburgh office)

gazetteers are reduced to 25,000 and 9,000 names; conversely, they also show that the addition of an extra 42 entries to the gazetteers improves performance dramatically.

This raises several questions: how important are gazetteers? is it important that they are big? if gazetteers are important but their size isn't, then what are the criteria for building gazetteers?

One might think that Named Entity recognition could be done by using lists of (e.g.) names of people, places and organisations, but that is not the case. To begin with, the lists would be huge: it is estimated that there are 1.5 million unique surnames just in the U.S. It is not feasible to list all possible surnames in the world in a Named Entity recognition system. There is a similar problem with company names. A list of all current companies worldwide would be huge, if at all available, and would immediately be out of date since new companies are formed all the time. In addition, company names can occur in variations: a list of company names might contain "The Royal Bank of Scotland plc", but that company might also be referred to as "The Royal Bank of Scotland", "The Royal" or "The Royal plc". These variations would all have to be listed as well.

Even if it was possible to list all possible organisations and locations and people, there would still be the problem of overlaps between the lists. Names such as Emerson or Washington could be names of people as well as places; Philip Morris could be a person or an organisation. In addition, such lists would also contain words like "Hope" and "Lost" (locations) and "Thinking Machines" and "Next" (companies), whereas these words could also occur in contexts where they don't refer to named entities.

Moreover, names of companies can be complex entities, consisting of several words. Especially where conjunctions are involved, this can create problems. In "China International Trust and Investment Corp decided to do something", it's not obvious whether there is a reference here to one company or two. In the sentence "Mason, Daily and Partners lost their court case" it is clear that "Mason, Daily and Partners" is the name of a company. In the sentence "Unfortunately, Daily and Partners lost their court case" the name of the company does not include the word "unfortunately", but it still includes the word "Daily", which is just as common a word as "unfortunately".

In this paper we report on a Named Entity recognition system which was amongst the highest scoring in the recent MUC-7 Message Understanding Conference/Competition (MUC). One of the features of our system is that even when it is run without any lists of names of organisations or people it still performs at a level comparable to that of many other MUC-systems. We report on experiments which show the difference in performance between the NE system with gazetteers of different sizes for three types of named entities: people, organisations and locations.

2 The MUC Competition

The MUC competition for which we built our system took place in March 1998. Prior to the competition, participants received a detailed coding manual which specified what should and should not be marked up, and how the markup should proceed. They also received a few hundred articles from the New York Times Service, marked up by the organisers according to the rules of the coding manual.

For the competition itself, participants received 100 articles. They then had 5 days to perform the chosen information extraction tasks (in our case: Named Entity recognition) without human intervention, and markup the text with the Named Entities found. The resulting marked up file then had to be returned to the organisers for scoring.

Scoring of the results is done automatically by the organisers. The scoring software compares a participant's answer file against a carefully prepared key file; the key file is considered to be the "correctly" annotated file. Amongst many other things, the scoring software calculates a system's recall and precision scores:

Recall: Number of correct tags in the answer file over total number of tags in the key file.

Precision: Number of correct tags in the answer file over total number of tags in the answer file.

Recall and precision are generally accepted ways of measuring system performance in this field. For example, suppose you have a text which is 1000 words long, and 20 of these words express a location. Now imagine a system which assigns the `LOCATION` tag to every single word in the text. This system will have tagged correctly all 20 locations, since it tagged everything as `LOCATION`; its recall score is 20/20, or 100%. But of the 1000 `LOCATION` tags it assigned, only those 20 were correct; its precision is therefore only 20/1000, or 2%.

category	learned lists		common lists		combined lists	
	recall	precision	recall	precision	recall	precision
organization	49	75	3	51	50	72
person	26	92	31	81	47	85
location	76	93	74	94	86	90

Figure 1: NE recognition with simple list lookup.

3 Finding Named Entities

3.1 A simple system

We decided first to test to what extent NE recognition can be carried out merely by recourse to list lookup. Such a system could be domain and language independent. It would need no grammars or even information about tokenization but simply mark up known strings in the text. Of course, the development and maintenance of the name lists would become more labour intensive.

(Palmer and Day, 1997) evaluated the performance of such a minimal NE recognition system equipped with name lists derived from MUC-6 training texts. The system was tested on newswire texts for six languages. It achieved a recall rate of about 70% for Chinese, Japanese and Portuguese and about 40% for English and French. The precision of the system was not calculated but can be assumed to be quite high because it would only be affected by cases where a capitalized word occurs in more than one list (e.g. "Columbia" could occur in the list of organisations as well as locations) or where a capitalised word occurs in a list but could also be something completely different (e.g. "Columbia" occurs in the list of locations but could also be the name of a space shuttle).

We trained a similar minimal system using the MUC-7 training data (200 articles) and ran it on the test data set (100 articles). The corpus we used in our experiments were the training and test corpora for the MUC-7 evaluation.

From the training data we collected 1228 person names, 809 names of organizations and 770 names of locations. The resulting name lists were the only resource used by the minimal NE recognition system. It nevertheless achieved relatively high precision (around 90%) and recall in the range 40–70%. The results are summarised in Figure 1 in the "learned lists" column.

Despite its simplicity, this type of system does presuppose the existence of training texts, and these are not always available. To cope with the absence of training material we designed and tested another variation of the minimal system. Instead of collecting lists from training texts we instead collected lists of commonly known entities—we collected a list of 5000 locations (countries and American states with their five biggest cities) from the CIA World Fact Book, a list of 33,000 organization names (companies, banks, associations, universities, etc.) from financial Web sites, and a list of 27,000 famous people from several websites. The results of this run can be seen in Figure 1 in the "common lists" column. In essence, this system's performance was comparable to that of the system using lists from the training set as far as location was concerned; it performed slightly worse on the person category and performed badly on organisations.

In a final experiment we combined the two gazetteers, the one induced from the training texts with the one acquired from public resources, and achieved some improvement in recall at the expense of precision. The results of this test run are given in the "combined lists" column in Figure 1.

We can conclude that the pure list lookup approach performs reasonably well for locations (precision of 90-94%; recall of 75-85%). For the person category and especially for the organization category this approach does not yield good performance: although the precision was not extremely bad (around 75-85%), recall was too low (lower than 50%)—i.e. every second person name or organization failed to be assigned.

For document retrieval purposes low recall is not necessarily a major problem since it is often sufficient to recognize just one occurrence of each distinctive entity per document, and many of the unassigned person and organization names were just repetitions of their full variants. But for many other applications, and for the MUC competition, higher recall and precision are necessary.

3.2 Combining rules and statistics

The system we fielded for MUC-7 makes extensive use of what McDonald (1996) calls *internal* (phrasal) and *external* (contextual) evidence in named entity recognition. The basic philosophy underlying our approach is as follows. A

Context Rule	Assign	Example
Xxxx+ is? a? JJ* PROF	PERS	Yuri Gromov, a former director
Xxxx+ is? a? JJ* REL	PERS	John White is beloved brother
Xxxx+ himself	PERS	White himself
Xxxx+, DD+,	PERS	White, 33,
shares in Xxxx+	ORG	shares in Trinity Motors
PROF of/at/with Xxxx+	ORG	director of Trinity Motors
Xxxx+ area	LOC	Beribidjan area

Figure 2: Examples of sure-fire transduction material for NE . Xxxx+ is a sequence of capitalized words; DD is a digit; PROF is a profession; REL is a relative; JJ* is a sequence of zero or more adjectives; LOC is a known location.

string of words like "Adam Kluver" has an internal (phrasal) structure which suggests that this is a person name; but we know that it can also be used as a shortcut for a name of organization ("Adam Kluver Ltd.") or location ("Adam Kluver Country Park"). Looking it up on a list will not necessarily help: the string may not be on a list, may be on more than one list, or may be on the wrong list. However, somewhere in the text, there is likely to be some contextual material which makes it clear what type of named entity it is. Our strategy is to only make a decision once we have identified this bit of contextual information.

We further assume that, once we have identified contextual material which makes it clear that "Adam Kluver" is (e.g.) the name of a company, then any other mention of "Adam Kluver" in that document is likely to refer to that company. If the author at some point in the same text also wants to refer to (e.g.) a *person* called "Adam Kluver", s/he will provide some extra context to make this clear, and this context will be picked up in the first step. The fact that at first it is only an assumption rather than a certainty that "Adam Kluver" is a company, is represented explicitly, and later processing components try to resolve the uncertainty.

If no suitable context is found anywhere in the text to decide what sort of Named Entity "Adam Kluver" is, the system can check other resources, e.g. a list of known company names and apply compositional phrasal grammars for different categories. Such grammars for instance can state that if a sequence of capitalized words ends with the word "Ltd." it is a name of organization or if a known first name is followed by an unknown capitalized word this is a person name.

In our MUC system, we implemented this approach as a staged combination of a rule-based system with probabilistic partial matching. We describe each stage in turn.

3.3 Step 1. Sure-fire Rules

In the first step, the system applies sure-fire grammar rules. These rules combine internal and external evidence, and only fire when a possible candidate expression is surrounded by a suggestive context. Sure-fire rules rely on known corporate designators (Ltd., Inc., etc.), person titles (Mr., Dr., Sen.), and definite contexts such as those in Figure 2. The sure-fire rules apply after POS tagging and simple semantic tagging, so at this stage words like "former" have already been identified as JJ (adjective), words like "analyst" have been identified as PROF (professions), and words like "brother" as REL (relatives).

At this stage our MUC system treats information from the lists as *likely* rather than definite and always checks if the context is either suggestive or non-contradictive. For example, a likely company name with a conjunction (e.g. "China International Trust and Investment Corp") is left untagged at this stage if the company is not listed in a list of known companies. Similarly, the system postpones the markup of unknown organizations whose name starts with a sentence initial common word, as in "Suspended Ceiling Contractors Ltd denied the charge".

Names of possible locations found in our gazetteer of place names are marked as LOCATION only if they appear with a context that is suggestive of location. "Washington", for example, can just as easily be a surname or the name of an organization. Only in a suggestive context, like "in Washington", will it be marked up as location.

3.4 Step 2. Partial Match 1

After the sure-fire symbolic transduction the system performs a probabilistic partial match of the identified entities. First, the system collects all named entities already identified in the document.

It then generates all possible partial orders of the composing words preserving their order, and marks them if found elsewhere in the text. For instance, if "Adam Kluver Ltd" had already been recognised as an organisation by the sure-fire rule, in this second step any occurrences of "Kluver Ltd", "Adam Ltd" and "Adam Kluver" are also tagged as *possible* organizations. This assignment, however, is not definite since some of these words (such as "Adam") could refer to a different entity.

This information goes to a pre-trained maximum entropy model (see Mikheev (1998) for more details on this aproach). This model takes into account contextual information for named entities, such as their position in the sentence, whether they exist in lowercase in general, whether they were used in lowercase elsewhere in the same document, etc. These features are passed to the model as attributes of the partially matched words. If the model provides a positive answer for a partial match, the system makes a definite assignment.

3.5 Step 3. Rule Relaxation

Once this has been done, the system again applies the grammar rules. But this time the rules have much more relaxed contextual constraints and extensively use the information from already existing markup and from the lexicon compiled during processing, e.g. containing partial orders of already identified named entities.

At this stage the system will mark word sequences which look like person names. For this it uses a grammar of names: if the first capitalized word occurs in a list of first names and the following word(s) are unknown capitalized words, then this string can be tagged as a PERSON. Note that it is only at this late stage that a list of names is used. At this point we are no longer concerned that a person name can refer to a company. If the name grammar had applied earlier in the process, it might erroneously have tagged "Adam Kluver" as a PERSON instead of an ORGANIZATION. But at this point in the chain of NE processing, that is not a problem anymore: "Adam Kluver" will by now already have been identified as an ORGANIZATION by the sure-fire rules or during partial matching. If it hasn't, then it is likely to be the name of a person.

At this stage the system will also attempt to resolve conjunction problems in names of organisations. For example, in "China International Trust and Investment Corp", the system checks if possible parts of the conjunctions were used in the text on their own and thus are names of different organizations; if not, the system has no reason to assume that more than one company is being talked about.

In a similar vein, the system resolves the attachment of sentence initial capitalized modifiers, the problem alluded to above with the "Suspended Ceiling Contractors Ltd" example: if the modifier was seen with the organization name elsewhere in the text, then the system has good evidence that the modifier is part of the company name; if the modifier does not occur anywhere else in the text with the company name, it is assumed not to be part of it.

This strategy is also used for expressions like "Murdoch's News Corp". The genitival "Murdoch's" could be part of the name of the organisation, or could be a possessive. Further inspection of the text reveals that Rupert Murdoch is referred to in contexts which support a person interpretation; and "News Corp" occurs on its own, without the genitive. On the basis of evidence like this, the system decides that the name of the organisation is "News Corp", and that "Murdoch" should be tagged separately as a person.

At this stage known organizations and locations from the lists available to the system are marked in the text, again without checking the context in which they occur.

3.6 Step 4. Partial Match 2

At this point, the system has exhausted its resources (rules about internal and external evidence for named entities, as well as its gazetteers). The system then performs another partial match to annotate names like "White" when "James White" had already been recognised as a person, and to annotate company names like "Hughes" when "Hughes Communications Ltd." had already been identified as an organisation.

As in Partial Match 1, this process of partial matching is again followed by a probabilistic assignment supported by the maximum entropy model. For example, conjunction resolution makes use of the fact that in this type of text it is more common to have conjunctions of like entities. In "he works for Xxx and Yyy", if there is evidence that Xxx and Yyy are two entities rather than one, then it is more likely that Xxx and Yyy are two entities of the same type, i.e. both organisations or are both people, rather than a mix of the two. This means that, even if only one of the entities in the conjunction has been recognised as definitely of a certain type, the conjunction rule will help decide on the type of the other entity. One of the texts in the competition contained the string "U7ited States and Russia". Because of the typo in "U7ited States", it wasn't found in a gazetteer. But there was internal evidence that it could be

Stage	ORGANIZATION		PERSON		LOCATION	
Sure-fire Rules	R: 42	P: 98	R: 40	P: 99	R: 36	P: 96
Partial Match 1	R: 75	P: 98	R: 80	P: 99	R: 69	P: 93
Relaxed Rules	R: 83	P: 96	R: 90	P: 98	R: 86	P: 93
Partial Match 2	R: 85	P: 96	R: 93	P: 97	R: 88	P: 93
Title Assignment	R: 91	P: 95	R: 95	P: 97	R: 95	P: 93

Figure 3: Scores obtained by the system through different stages of the analysis. R - recall P - precision.

a location (the fact that it contained the word "States"); and there was external evidence that it could be a location (the fact that it occurred in a conjunction with "Russia", a known location). These two facts in combination meant that the system correctly identified "U7ited States" as a location.

3.7 Step 5. Title Assignment

Because titles of news wires are in capital letters, they provide little guidance for the recognition of names. In the final stage of NE processing, entities in the title are marked up, by matching or partially matching the entities found in the text, and checking against a maximum entropy model trained on document titles. For example, in "GENERAL TRENDS ANALYST PREDICTS LITTLE SPRING EXPLOSION" "GENERAL TRENDS" will be tagged as an organization because it partially matches "General Trends Inc" elsewhere in the text, and "LITTLE SPRING" will be tagged as a location because elsewhere in the text there is supporting evidence for this hypothesis. In the headline "MURDOCH SATELLITE EXPLODES ON TAKE-OFF", "Murdoch" is correctly identified as a person because of mentions of Rupert Murdoch elsewhere in the text. Applying a name grammar on this kind of headline without checking external evidence might result in erroneously tagging "MURDOCH SATELLITE" as a person (because "Murdoch" is also a first name, and "Satellite" in this headline starts with a capital letter).

4 MUC results

In the MUC competition, our system's combined precision and recall score was 93.39%. This was the highest score, better in a statistically significant way than the score of the next best system. Scores varied from 93.39% to 69.67%. Further details on this can be found in (Mikheev et al., 1998).

The table in Figure 3 shows the progress of the performance of the system we fielded for the MUC competition through the five stages.

As one would expect, the sure-fire rules give very high precision (around 96-98%), but very low recall—in other words, they don't find many named entities, but the ones they find are correct. Subsequent phases of processing add gradually more and more named entities (recall increases from around 40% to around 90%), but on occasion introduce errors (resulting in a slight drop in precision). Our final score for `ORGANISATION`, `PERSON` and `LOCATION` is given in the bottom line of Figure 3.

5 The role of gazetteers

Our system fielded for the MUC competition made extensive use of gazetteers, containing around 4,900 names of countries and other place names, some 30,000 names of companies and other organisations, and around 10,000 first names of people. As explained in the previous section, these lists were used in a judicious way, taking into account other internal and external evidence before making a decision about a named entity. Only in step 3 is information from the gazetteers used without context-checking.

It is not immediately obvious from Figure 3 what exactly the impact is of these gazetteers. To try and answer this question, we ran our system over 70 articles of the MUC competition in different modes; the remaining 30 articles were used to compile a limited gazetteer as described below and after that played no role in the experiments.

Full gazetteers. We first ran the system again with the full gazetteers, i.e. the gazetteers used in the official MUC system. There are minor differences in Recall and Precision compared to the official MUC results, due to the fact that we were using a slightly different (smaller) corpus.

No gazetteers. We then ran the system without any gazetteers. In this mode, the system can still use internal evidence (e.g. indicators such as "Mr" for people or "Ltd" for organisations) as well as external evidence (contexts such as "XXX, the chairman of YYY" as evidence that XXX is a person and YYY an organisation).

The hypothesis was that names of organisations

	Full gazetteer		Ltd gazetteer		Some locations		No gazetteers	
	recall	prec'n	recall	prec'n	recall	prec'n	recall	prec'n
organisation	90	93	87	90	87	89	86	85
person	96	98	92	97	90	97	90	95
location	95	94	91	92	85	90	46	59

Figure 4: Our MUC system with extensive gazetteers, with limited gazetteers, with short list of locations, and without gazetteers, tested on 70 articles from the MUC-7 competition.

and names of people should still be handled relatively well by the system, since they have much internal and external evidence, whereas names of locations have fewer reliable contextual clues. For example, expressions such as "XXX is based in YYY" is not sure-fire evidence that YYY is a location – it could also be an organisation. And since many locations are so well-known, they receive very little extra context ("in China", "in Paris", vs "in the small town of Ekeren").

Some locations. We then ran the system with some locational information: about 200 names of countries and continents from `www.yahoo.com/Regional/` and, because MUC rules say explicitly that names of planets should be marked up as locations, the names of the 8 planets of our solar system. The hypothesis was that even with those reasonably common location names, Named Entity recognition would already dramatically improve. This hypothesis was confirmed, as can be seen in Figure 4.

Inspection of the errors confirms that the system makes most mistakes when there is no internal or external evidence to decide what sort of Named Entity is involved. For example, in a reference to "a Hamburg hospital", "Hamburg" no longer gets marked up as a location, because the word occurs nowhere else in the text, and that context is not sufficient to assume it indicates a location (cf. a Community Hospital, a Catholic Hospital, an NHS Hospital, a Trust-Controlled Hospital, etc). Similarly, in a reference to "the Bonn government", "Bonn" is no longer marked up as a location, because of lack of supportive context (cf. the Clinton government, the Labour government, etc). And in financial newspaper articles NYSE will be used without any indication that this is an organisation (the New York Stock Exchange).

Limited gazetteers. The results so far suggest that the most useful gazetteers are those that contain very common names, names which the authors can expect their audience already to know about, rather than far-fetched examples of little known places or organisations.

This suggests that it should be possible to tune a system to the kinds of Named Entities that occur in its particular genre of text. To test this hypothesis, we wanted to know how the system would perform if it started with no gazetteers, started processing texts, then built up gazetteers as it goes along, and then uses these gazetteers on a new set of texts in the same domain. We simulated these conditions by taking 30 of the 100 official MUC articles and extracting all the names of people, organisations and locations and using these as the only gazetteers, thereby ensuring that we had extracted Named Entities from articles in the same domain as the test domain.

Since we wanted to test how easy it was to build gazetteers automatically, we wanted to minimise the amount of processing done on Named Entities already found. We decided to only used first names of people, and marked them all as "likely" first names: the fact that "Bill" actually occurs as a first name does not guarantee it will definitely be a first name next time you see it. Company names found in the 30 articles were put in the company gazetteer, irrespective of whether they were full company names (e.g. "MCI Communications Corp" as well as "MCI" and "MCI Communications"). Names of locations found in the 30 texts were simply added to the list of 200 location names already used in the previous experiments.

The hope was that, despite the little effort involved in building these limited gazetteers, there would be an improved performance of the Named Entity recognition system.

Figure 4 summarises the Precision and Recall results for each of these modes and confirms the hypotheses.

6 Discussion

The hypotheses were correct: without gazetteers the system still scores in the high eighties for names of organisations and people. Locations come out badly. But even with a very small number of country names performance for those named entities also goes up into the mid-

eighties. And simple techniques for extending the gazetteers on the basis of a sample of just 30 articles already makes the system competitive again.

These experiments suggest that the collection of gazetteers need not be a bottleneck: through a judicious use of internal and external evidence relatively small gazetteers are sufficient to give good Precision and Recall. In addition, when collecting these gazetteers one can concentrate on the *obvious* examples of locations and organisations, since these are exactly the ones that will be introduced in texts without much helpful context.

However, our experiments only show the usefulness of gazetteers on a particular type of text, viz. journalistic English with mixed case. The rules as well as the maximum entropy models make use of internal and external evidence in that type of text when trying to identify named entities, and it is obvious that this system cannot be applied without modification to a different type of text, e.g. scientific articles. Without further formal evaluations with externally supplied evaluation corpora it is difficult to judge how general this text type is. It is encouraging to note that Krupka and Hausman (1998) point out that the MUC-7 articles which we used in our experiments have less external evidence than do Wall Street Journal articles, which suggests that on Wall Street Journal articles our system might perform even better than on MUC-7 articles.

Acknowledgements

The work reported in this paper was supported in part by grant GR/L21952 (Text Tokenisation Tool) from the Engineering and Physical Sciences Research Council, UK. We would like to thank Steve Finch and Irina Nazarova as well as Colin Matheson and other members of the Language Technology Group for help in building various tools and other resources that were used in the development of the MUC system.

References

Alessandro Cucchiarelli, Danilo Luzi, and Paola Velardi. 1998. Automatic semantic tagging of unknown proper names. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and Proceedings of the 17th International Conference on Computational Linguistics*, pages 286–292, Montréal, Canada, August 10–14.

George R. Krupka and Kevin Hausman. 1998. Isoquest, Inc: Description of the NetOwl(TM) extractor system as used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference held in Fairfax, Virginia, 29 April–1 May, 1998.* `http://www.muc.saic.com/proceedings/muc_7_toc.html`.

George R. Krupka. 1995. Description of the SRA system as used for MUC-6. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference held in Columbia, Maryland, November 6–8, 1995*, pages 221–235, Los Altos, Ca. Morgan Kaufmann.

David D. McDonald. 1996. Internal and external evidence in the identification and semantic categorization of proper names. In Bran Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, chapter 2, pages 21–39. The MIT Press, Cambridge, MA.

Andrei Mikheev, Claire Grover, and Marc Moens. 1998. Description of the LTG system used for MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference held in Fairfax, Virginia, 29 April–1 May, 1998.* `http://www.muc.saic.com/proceedings/muc_7_toc.html`.

Andrei Mikheev. 1998. Feature lattices for maximum entropy modelling. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and Proceedings of the 17th International Conference on Computational Linguistics*, pages 848–854, Montreal, Quebec, August 10–14.

Richard Morgan, Roberto Garigliano, Paul Callaghan, Sanjay Poria, Mark Smith, Agnieszka Urbanowicz, Russel Collingham, Marco Costantino, and Chris Cooper. 1995. Description of the LOLITA system as used in MUC-6. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference held in Columbia, Maryland, November 6–8, 1995*, pages 71–86, Los Altos, Ca. Morgan Kaufmann.

D. Palmer and D. Day. 1997. A statistical profile of the Named Entity task. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 190–193, Washington D.C.

Beth Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference held in Columbia, Maryland, November 6–8, 1995*, pages 13–32, Los Altos, Ca. Morgan Kaufmann.