Evaluation of an Algorithm for the Recognition and Classification of Proper Names

Takahiro Wakao **Robert Gaizauskas** **Yorick Wilks**
Department of Computer Science,
University of Sheffield
{T.Wakao,R.Gaizauskas,Y.Wilks}@dcs.shef.ac.uk

Abstract

We describe an information extraction system in which four classes of naming expressions – organisation, person, location and time names – are recognised and classified with nearly 92% combined precision and recall. The system applies a mixture of techniques to perform this task and these are described in detail. We have quantitatively evaluated the system against a blind test set of *Wall Street Journal* business articles and report results not only for the system as a whole, but for each component technique and for each class of name. These results show that in order to have high recall, the system needs to make use not only of information internal to the naming expression but also information from outside the name. They also show that the contribution of each system component varies from one class of name expression to another.

1 Introduction

The appropriate treatment of proper names is essential in a natural language understanding system which processes unedited newswire text, since up to 10 % of this type of text may consist of proper names (Coates-Stephens, 1992). Nor is it only the sheer volume of names that makes them important; for some applications, such as information extraction (IE), robust handling of proper names is a prerequisite for successfully performing other tasks such as template filling where correctly identifying the entities which play semantic roles in relational frames is crucial. Recent research in the fifth and sixth Message Understanding Conferences (MUC5, 1993) (MUC6, 1995) has shown that the recognition and classification of proper names in business newswire text can now be done on a large scale and with high accuracy: the success rates of the best systems now approach 96%.

We have developed an IE system – *LaSIE* (Large Scale Information Extraction) (Gaizauskas *et al*, 1995) – which extracts important facts from business newswire texts. As a key part of the extraction task, the system recognises and classifies certain types of naming expressions, namely those specified in the MUC-6 named entity (NE) task definition (MUC6, 1995). These include organisation, person, and location names, time expressions, percentage expressions, and monetary amount expressions. As defined for MUC-6, the first three of these are proper names, the fourth contains some expressions that would be classified as proper names by linguists and some that would not, while the last two would generally not be thought of as proper names. In this paper we concentrate only the behaviour of the *LaSIE* system with regards to recognising and classifying expressions in the first four classes, i.e. those which consist entirely or in part of proper names (though nothing hangs on omitting the others). The version of the system reported here achieves almost 92% combined precision and recall scores on this task against blind test data.

Of course the four name classes mentioned are not the only classes of proper names. Brand names, book and movie names, and ship names are just a few further classes one might chose to identify. One might also want to introduce subclasses within the selected classes. We have not done so here for two reasons. First, and foremost, in order to generate quantitative evaluation results we have used the MUC-6 data and scoring resources and these restrict us to the above proper name classes. Secondly, these four name classes account for the bulk of proper name occurrences in business newswire text. Our approach could straightforwardly be extended to account for additional classes of proper names, and the points we wish to make about the approach can be adequately presented using only this restricted set.

Our approach to proper name recognition is heterogeneous. We take advantage of graphological, syntactic, semantic, world knowledge, and discourse level information to perform the task. In the paper we present details of the approach, describing those data and processing components of the overall IE system which contribute to proper name recognition and classification. Since name

recognition and classification is achieved through the activity of four successive components in the system, we quantitatively evaluate the successive contribution of each component in our overall approach. We perform this analysis not only for all classes of names, but for each class separately. The resulting analysis

1. supports McDonald's observation (McDonald, 1993) that external evidence as well as internal evidence is essential for achieving high precision and recall in the recognition and classification task; i.e. not just the name string itself must be examined, but other information in the text must be used as well;
2. shows that all components in our heterogeneous approach contribute significantly;
3. shows that not all classes of proper names benefit equally from the contributions of the different components in our system: in particular, organisation names benefit most from the use of external evidence.

In the second section an overview of the LaSIE system is presented. The third section explains in detail how proper names are recognised and classified in the system. The results of evaluating the system on a blind test set of 30 articles are presented and discussed in section 4. Section 5 concludes the paper.

2 *LaSIE* system overview

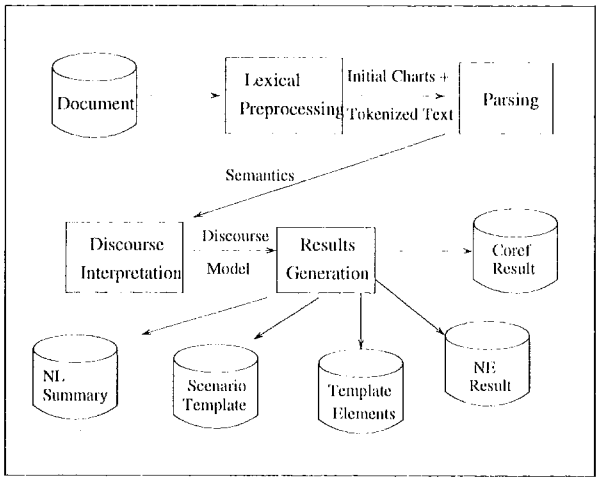

Figure 1: LaSIE System Architecture

LaSIE has been designed as a general purpose IE research system, initially geared towards, but not solely restricted to, carrying out the tasks specified by the sixth Message Understanding Conference: named entity recognition, coreference resolution, template element filling, and scenario template filling tasks (see (MUC6, 1995) for further details of the task descriptions). In addition, the system can generate a brief natural language summary of the scenario it has detected in the text.

All of these tasks are carried out by building a single rich model of the text – the discourse model – from which the various results are read off.

The high level structure of LaSIE is illustrated in **Figure 1**. The system is a pipelined architecture which processes a text sentence-at-a-time and consists of three principal processing stages: lexical preprocessing, parsing plus semantic interpretation, and discourse interpretation. The overall contributions of these stages may be briefly described as follows:

- **lexical preprocessing** reads and tokenises the raw input text, tags the tokens with parts-of-speech, performs morphological analysis, performs phrasal matching against lists of proper names, and builds lexical and phrasal chart edges in a feature-based formalism for hand-over to the parser;
- **parsing** does two pass parsing, pass one with a special proper name grammar, pass two with a general grammar and, after selecting a 'best parse', passes on a semantic representation of the current sentence which includes name class information;
- **discourse interpretation** adds the information in its input semantic representation to a hierarchically structured semantic net which encodes the system's world model, adds additional information presupposed by the input to the world model, performs coreference resolution between new instances added and others already in the world model, and adds information consequent upon the addition of the input to the world model.

For further details of the system see (Gaizauskas *et al*, 1995).

3 How proper names are recognised and classified

As indicated in section 1, our approach is a heterogeneous one in which the system makes use of graphological, syntactic, semantic, world knowledge, and discourse level information for the recognition and classification of proper names. The system utilises both the information which comes from the name itself (internal evidence in McDonald's sense (McDonald, 1993)) as well as the information which comes from outside the name, from its context in the text (external evidence).

In what follows we describe how proper names are recognised and classified in *LaSIE* by considering the contribution of each system component.

3.1 Lexical preprocessing

The input text is first tokenised and then each token is tagged with a part-of-speech tag from the Penn Treebank tagset (Marcus *et al*, 1993) using a slightly customised [1] version of Brill's tag-

[1] The tagger has been customised by adding some entries to its lexicon and by adding several special tags

ger (Brill, 1994). The tagset contains two tags for proper nouns – `NNP` for singular proper nouns and `NNPS` for plurals. The tagger tags a word as a proper noun as follows: if the word is found in the tagger's lexicon and listed as a proper noun then tag it as such; otherwise, if the word is not found in the lexicon and is uppercase initial then tag it as a proper noun. Thus, capitalised unknown tokens are tagged as proper nouns by default.

Before parsing an attempt is made to identify proper name phrases – sequences of proper names – and to classify them. This is done by matching the input against pre-stored lists of proper names. These lists are compiled via a *flex* program into a finite state recogniser. Each sentence is fed to the recogniser and all single and multi-word matches are tagged with special tags which indicate the name class.

Lists of names used include:

- organisation : about 2600 company and governmental institution names based on an organisation name list which was semi-automatically collected from the MUC-5 answer keys and training corpus (*Wall Street Journal* articles);
- location : about 2200 major country, province/state, and city names derived from a gazetteer list of about 150,000 place names;
- person : about 500 given names taken from a list of given names in the Oxford Advanced Learner's Dictionary (Hornby, 1980);
- company designator : 94 designators (e.g. 'Co.','PLC'), based on the company designator list provided in the MUC6 reference resources.
- human titles : about 160 titles, (e.g. 'President','Mr.'), manually collected;

As well as name phrase matching, another technique is applied at this point. Inside multi-word proper names, certain words may function as *trigger words*. A trigger word indicates that the tokens surrounding it are probably a proper name and may reliably permit the class or even subclass [2] of the proper name to be determined. For example, 'Wing and Prayer Airlines' is almost certainly a company, given the presence of the word 'Airlines'. Trigger words are detected by matching against lists of such words and are then specially tagged. Subsequently these tags are used by the proper name parser to build complex proper name constituents.

The lists of trigger words are:

- Airline company: 3 trigger words for finding airline company names, e.g. 'Airlines';
- Governmental institutions: 7 trigger words for governmental institutions, e.g 'Ministry';
- Location: 8 trigger words for location names, e.g. 'Gulf';
- Organisation: 135 trigger words for organisation names, e.g 'Association'.

These lists of trigger words were produced by hand, though the organisation trigger word lists were generated semi-automatically by looking at organisation names in the MUC-6 training texts and applying certain heuristics. So, for example, words were collected which come immediately before 'of' in those organisation names which contain 'of', e.g. 'Association' in 'Association of Air Flight Attendants'; the last words of organisation names which do not contain 'of' were examined to find trigger words like 'International'.

3.2 Grammar rules for proper names

The LaSIE parser is a simple bottom-up chart parser implemented in Prolog. The grammars it processes are unification-style feature-based context free grammars. During parsing, semantic representations of constituents are constructed using Prolog term unification. When parsing ceases, i.e. when the parser can generate no further edges, a 'best parse selection' algorithm is run on the final chart to chose a single analysis. The semantics are then extracted from this analysis and passed on to the discourse interpreter.

Parsing takes place in two passes, each using a separate grammar. In the first pass a special grammar is used to identify proper names. These constituents are then treated as unanalysable units during the second pass which employs a more general 'sentence' grammar.

Proper Name Grammar The grammar rules for proper names constitute a subset of the system's noun phrase (NP) rules. All the rules were produced by hand. There are 177 such rules in total of which 94 are for organisation, 54 for person, 11 for location, and 18 for time expressions.

Here are some examples of the proper name grammar rules:

```
NP --> ORGAN_NP
ORGAN_NP --> LIST_LOC_NP NAMES_NP CDG_NP
ORGAN_NP --> LIST_ORGAN_NP NAMES_NP CDG_NP
ORGAN_NP --> NAMES_NP '&' NAMES_NP
NAMES_NP --> NNP NAMES_NP
NAMES_NP --> NNP PUNC(_) NNP
NAMES_NP --> NNP
```

The non-terminals `LIST_LOC_NP`, `LIST_ORGAN_NP` and `CDG_NP` are tags assigned to one or more input tokens in the name phrase tagging stage of lexical preprocessing. The non-terminal `NNP` is the tag for proper name assigned to a single token by the Brill tagger.

The rule `ORGAN_NP --> NAMES_NP '&' NAMES_NP` means that if an as yet unclassified or ambiguous proper name (`NAMES_NP`) is followed by '&' and another ambiguous proper name, then it is an organisation name. So, for example, 'Marks & Spen-

for word classes such as days of the week and months.

[2] `company` and `governmental institution` are subclasses of the class `organisation`. `airline` is a subclass of `company`.

cer' and 'American Telephone & Telegraph' will be classified as organisation names by this rule.

Nearly half of the proper name rules are for organisation names because they may contain further proper names (e.g. person or location names) as well as normal nouns, and their combinations. There are also a good number of rules for person names since care must be taken with given names, family names, titles (e.g. 'Mr.','President'), and special lexical items such as 'de' (as in 'J. Ignacio Lopez de Arriortua') and 'Jr.','II', etc.

There are fewer rules for location names, as they are identified mainly in the previous preprocessing stage by look-up in the mini-gazetteer.

Sentence Grammar Rules The grammar used for parsing at the sentence level contains approximately 110 rules and was derived automatically from the Penn TreeBank-II (PTB-II) (Marcus *et al*, 1993), (Marcus *et al*, 1995). When parsing for a sentence is complete the resultant chart is analysed to identify the 'best parse'. From the best parse the associated semantics are extracted to be passed on to the discourse interpreter.

Rules for compositionally constructing semantic representations were assigned by hand to the grammar rules. For simple verbs and nouns the morphological root is used as a predicate name in the semantics, and tense and number features are translated directly into the semantic representation where appropriate. For named entities a token of the most specific type possible (e.g. `company` or perhaps only `object`) is created and a `name` attribute is associated with the entity, the attribute's value being the surface string form of the name. So, for example, assuming 'Ford Motor Co.' has already been classified as a company name, its semantic representation will be something like `company(e23) & name(e23,'Ford Motor Co.')`.

3.3 Discourse interpretation

The discourse interpreter module performs two activities that contribute to proper name classification (no further recognition of proper names goes on at this point, only a refining of their classification). The first activity is coreference resolution – an unclassified name may be coreferred with a previously classified one by virtue of which the class of the unclassified name becomes known. The second activity, which is arguably not properly 'discourse interpretation' but nevertheless takes place in this module, is to perform inferences about the semantic types of arguments in certain relations; for example, in compound nominals such as 'Erikson stocks' our semantic interpreter will tell us that there is a qualifier relation between 'Erikson' and 'stocks' and since the system stores the fact that named entities qualifying things of type `stock` are of type `company` it can classify the proper name 'Erikson' as a company.

Note that both of these techniques make use of external evidence, i.e. rely on information supplied by the context beyond the words in the instance of the proper name being classified.

3.3.1 Proper name coreference

Coreference resolution for proper names is carried out in order to recognise alternative forms, especially of organisation names. For example, 'Ford Motor Co.' might be used in a text when the company is first mentioned, but subsequent references are likely to be to 'Ford'. Similarly, 'Creative Artists Agency' might be abbreviated to 'CAA' later on in the same text. Such shortened forms must be resolved as names of the same organisation.

In order to determine whether given two proper names match, various heuristics are used. For example, given two names, Name1 and Name2:

- if Name2 is consists of an initial subsequence of the words in Name1 then Name2 matches Name1 – e.g. 'American Airlines Co.' and 'American Airlines';
- if Name1 is a person name and Name2 is either the first, the family, or both names of Name1, then Name2 matches Name1 – e.g. 'John J. Major Jr.' and 'John Major'.

There are 31 such heuristic rules for matching organisation names, 11 heuristics for person names, and 3 rules for location names.

When an unclassified proper noun is matched with a previously classified proper name in the text, it is marked as a proper name of the class of the known proper name. Thus, when we know 'Ford Motor Co.' is an organisation name but have not classified 'Ford' in the same text, coreference resolution determines 'Ford' to be an organisation name.

3.3.2 Semantic Type Inference

In the following contexts, semantic type information about the types of arguments in certain relations is used to drive inferences permitting the classification of proper names. The system uses these techniques in a fairly limited and experimental way at present, and there is much room for their extension.

- noun-noun qualification: when an unclassified proper name qualifies an organisation-related thing then the name is classified as an organisation; e.g. in 'Erickson stocks' since 'stock' is semantically typed as an organisation-related thing, 'Erickson' gets classified as an organisation name.
- possessives: when an unclassified proper name stands in a possessive relation to an organisation post, then the name is classified as an organisation; e.g. 'vice president of ABC', 'ABC's vice president'.
- apposition : when an unclassified proper name is apposed with a known location name,

the former name is also classified as a location; e.g. given 'Fort Lauderdale, Fla.' if we know 'Fla.' is a location name, 'Fort Lauderdale' is also classified as a location name.

- verbal arguments: when an unclassified proper name names an entity playing a role in a verbal frame where the semantic type of the argument position is known, then the name is classified accordingly; e.g. in 'Smith retired from his position as ...' we can infer that 'Smith' is a person name since the semantic type of the logical subject of 'retire' (in this sense) is `person`.

4 Results and Evaluation

After these processing stages, the results generator produces a version of the original text in which all the proper names which have been detected are marked up with pre-defined SGML tags, specifying their classes. These marked up texts are then automatically scored against manually marked up texts.

A series of evaluations has been done on the system using a blind test set consisting of 30 *Wall Street Journal* texts. In these texts there are 449 organisation names, 373 person names, and 110 location names and 111 time expressions in total. The overall precision and recall scores for the four classes of proper names are shown in **Table 1**.

Proper Name Class	Recall	Precision
Organisation	91 %	91 %
Person	90 %	95 %
Location	88 %	89 %
Time	94 %	97 %
Overall	91 %	93 %

Table 1: Overall Precision and Recall Scores

4.1 System module contribution

We have analysed the results in terms of how much each module of the system contributes to the proper name task.

Table 2 illustrates the contribution of each system module to the task for all classes of proper names. In addition to recall and precision scores, we have added Van Rijsbergen's F-measure which combines these scores into a single measure (Rijsbergen, 1979). The F-measure (also called P&R) allows the differential weighting of precision and recall. With precision and recall weighted equally it is computed by the formula:

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

There are four different settings of the system.

- **setting 1** : Only the lexical preprocessing techniques are used – part-of-speech tagging and name phrase matching.
- **setting 2** : Two-stage parsing is added to **1**.
- **setting 3** : Coreference resolution for proper names is added to **2**.
- **setting 4** : Full discourse interpretation is added to **3**. This is the full-fledged system

Setting	Recall	Precision	P&R
1	49	89	63.01
2	79	94	85.82
3	89	94	91.13
4	91	93	91.75

Table 2: Module Contribution Scores

Table 2 shows that we can attain reasonable results using tagging, exact phrase matching, trigger word detection, and parsing (setting 2). Note that this amounts to making use of only internal evidence. However, to achieve higher recall, we need coreference resolution for proper names (setting 3) and other context information (setting 4).

4.2 Different classes of proper names

We have also examined how the contribution of each component varies from one class of proper name to another.

For organisation names, using the same settings as above, scores are shown in **Table 3**.

Setting	Recall	Precision	P&R
1	46	87	59.91
2	65	92	76.15
3	87	93	89.84
4	91	91	91.13

Table 3: Module Contributions for Org Names

For person names, location names and time expressions the results are shown in **Tables 4-6**.

Setting	Recall	Precision	P&R
1	47	88	61.64
2	89	95	92.34
3	90	95	92.14
4	90	95	92.14

Table 4: Module Contributions for Person Names

Figure 2 shows graphically how the system components contribute for each of the four different classes of proper names as well as for all classes combined.

5 Conclusion

We have described an IE system in which four classes of naming expressions (organisation, person, and location names and time expressions) are recognised and classified. The system was tested on 30 unseen *Wall Street Journal* texts and the results were analysed in terms of major system components and different classes of referring expression.

Setting	Recall	Precision	P&R
1	81	94	86.84
2	88	90	88.99
3	88	89	88.58
4	88	89	88.58

Table 5: Module Contributions for Location Names

Setting	Recall	Precision	P&R
1	32	100	48.97
2	94	97	95.41
3	94	97	95.41
4	94	97	95.41

Table 6: Module Contributions for Time Expressions

Tables 3-6 and **Figure 2** enable us to make the following observations:

1. Techniques relying on internal evidence only – exact word and phrase matching, grapholological conventions, and parsing – are not sufficient to recognise and classify organisation names. It is clear that in order to have high recall for organisation names, we need to be able to make good use of external evidence as well, i.e. proper name coreference resolution and information from the surrounding context.
2. On the other hand, for person and location names and time expressions, techniques relying solely on internal evidence do permit us to attain high recall whilst maintaining high precision. Thus, the contribution of different system components varies from one class of proper name to another.
3. However, given that in a reasonable sample of business newswire text, 43 % of the proper names are organisation names, it is evident that for a system to achieve high overall precision and recall in the name recognition and classification task on this text type, it must utilise not only internal evidence but also external evidence.

More generally, two conclusions can be drawn. First, the results presented above suggest that when a system for proper name recognition and/or classification is evaluated, much benefit can be gained by analysing it not only in terms of overall recall and precision figures, but also in terms of system components and classes of names. Second, a heterogeneous approach to the recognition and classification of proper names in newswire text such as described here is appropriate since it provides mechanisms that can utilise the variety of internal and external evidence which is available and which needs to be taken into account.

6 Acknowledgements

This research has been made possible by the grants from the U.K. Department of Trade and Industry (Grant Ref. YAE/8/5/1002) and the Engineering and Physical Science Research Council (Grant # GR/K25267).

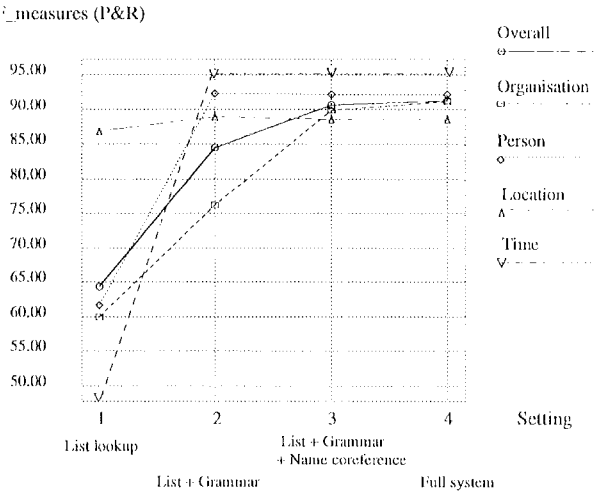

Figure 2: Contribution of System Components

References

Brill, E. (1994). "Some Advances in Transformation Based Part of Speech Tagging." In *Proc. AAAI.*

Coates-Stephens, S. (1992). *The Analysis and Acquisition of Proper Names for Robust Text Understanding.* PhD thesis, Department of Computer Science, City University, London.

Gaizauskas, R.; Wakao, T.; Humphreys, K.; Cunningham, H. and Wilks, Y. (1995). "University of Sheffield: Description of LaSIE system as used for MUC-6." In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufman.

Hornby A.S. (Ed.). (1980). *Oxford Advanced Learner's Dictionary of Current English.* London: Oxford University Press.

Marcus, M.; Santorini, B. and Marcinkiewicz, M.A. (1993). "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics*, 19 (2): 313 330.

Marcus, M.; Kim, G; Marcinkiewicz, M.A.; MacIntyre, R.; Bies, A.; Ferguson, M.; Katz, K. and Schasberger, B. (1995). "The Penn Treebank: Annotating Predicate Argument Structure." Distributed on The Penn Treebank Release 2 CD-ROM by the Linguistic Data Consortium.

McDonald, D.D. (1993). "Internal and External Evidence in the Identification and Semantic Categorisation of Proper Names." In *Proceedings of SIGLEX workshop on "Acquisition of Lexical Knowledge from Text"*, pp. 32 43.

MUC-5. (1993). *Proceedings of the Fifth Message Understanding Conference (MUC-5).* Morgan Kaufman.

MUC-6. (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6).* Morgan Kaufman.

Van Rijsbergen, C.J. (1979). *Information Retrieval.* London: Butterworths.