A Framework for Named Entity Recognition in the Open Domain

Richard Evans
Research Group in Computational Linguistics
School of Humanities, Languages, and Social Sciences
University of Wolverhampton
Stafford Street
Wolverhampton WV1 1SB, UK
{R.J.Evans@wlv.ac.uk}

Abstract

In this paper, a system for Named Entity Recognition in the Open domain (NERO) is described. It is concerned with recognition of various types of entity, types that will be appropriate for Information Extraction in any scenario context. The recognition task is performed by identifying normally capitalised phrases in a document and then submitting queries to a search engine to find potential hypernyms of the capitalised sequences. These hypernyms are then clustered to derive a typology of named entities for the document. The hypernyms of the normally capitalised phrases are used to classify them with respect to this typology. The method is tested on a small corpus and its classifications are evaluated. Finally, conclusions are drawn and future work considered.

1 Introduction

Information Extraction (IE) is defined as the automatic identification of selected types of entities, relations, or events in free text (Grishman 03). Some of the most significant multi-site evaluations of IE have been carried out in the Message Understanding Competitions (eg. MUC-7 (Chinchor 98b)).

In the context of IE, the events of interest described in documents are encoded using templates (Chinchor 98a). An IE system attempts to assign the participants of an event to functional *slots* in the template. The slots accept elements of particular types. For instance, a template corresponding to management succession events includes slots for the person who is leaving the post; the organisation in which the event is taking place; etc. Each slot has a particular semantics and participants that are appropriate for each slot are subject to that semantics.

The templates used in MUC-7 have slots for ENTITY and LOCATION elements. ENTITY elements are divided into classes for PERSON, ORGANIZATION, and ARTIFACT. The goal of named entity recognition (NER) is to identify these elements in texts automatically.

MUC-7 represents just one IE scenario and many more types of entity must be recognised for effective IE in different domains. In the domains of medicine, e-retail, or entertainment, systems will need to identify pharmaceutical names, product names, or pop groups. The set of required name types varies from case to case. The NER approaches developed for MUC-7 are able to recognise a small set of named entity (NE) types with considerable accuracy. However, in most cases, the classification models used rely on specific domain knowledge. They cannot easily be extended to recognise the pertinent entities occurring in documents from other domains.

In the current state of the art, IE systems must either be re-tuned or else implemented from scratch for new scenarios. In this paper, the goal is to automatically identify the entities that are likely to be of interest in any scenario context, with no knowledge *a priori*. A system called NERO[1] is described that embodies a framework for NER in the open domain.

The paper is structured as follows. Section 2 describes the methods by which NERO addresses its goals - deriving a document specific ontology for NEs (Section 2.1); identifying sequences of words that are normally capitalised in the document (Section 2.2); and then classifying normally capitalised words with respect to the derived typology (Section 2.3). In Section 3, the small corpus used to test the system is described and in Section 4 the resulting evaluation is reported. In Section 5, related work is reviewed and in Section 6 conclusions are drawn and directions for future research considered.

2 The Method for Named Entity Recognition in the Open Domain

The process of open-domain NER is tackled in three stages. Firstly, a document-specific typology for NEs is derived automatically (Section

[1] Named Entity Recognition in the Open domain

2.1). Secondly, NEs are identified (Section 2.2). Thirdly, NEs are classified in line with the derived typology (Section 2.3).

2.1 Typology Derivation

The typology is obtained by collecting the hypernyms of capitalised phrases (Section 2.1.1), clustering the hypernyms (Section 2.1.2), and labelling those clusters (Section 2.1.3).

2.1.1 Collection of Hypernyms

The method for identification of hyponyms described in (Hearst 92) was applied in order to identify potential hypernyms of sequences of capitalised words appearing in the document. Here, sequences include single words. In the first sentence of the abstract of this paper, the sequences of capitalised words are $\{$*In*, *Named Entity Recognition*, *Open*, and *NERO*$\}$. Numerous patterns were used to produce queries that were submitted to the google[2] search engine. The summaries returned by *google* were then used to derive the hypernyms. Following (Hearst 92), when X is a capitalised sequence, the query `such as X`, was submitted to *google*. The FDG Parser (Tapanainen & Järvinen 97) was used to find the part of speech and lemma of words in the returned summaries. The lemma of the immediately preceding noun was chosen as the hypernym of X. Three other patterns (Figure 1) were included in an effort to improve coverage. In these patterns, Y is a phrase whose head is a potential hypernym of X.

- Y such as X
- Y like X
- X or other Y
- X and other Y

Figure 1: Hypernymy Patterns

One problem that results from exploiting *google* for information retrieval is that case and punctuation are ignored in the queries. This is a particular problem in the current context because many effective patterns are best expressed using a combination of words and punctuation marks. To illustrate, the pattern *X and other Y* shown in Figure 1 was originally expressed as $NP\{, NP\}^*\{,\}$ *and other* NP in (Hearst 92). Despite this, the coverage provided by *google* means that it is currently preferred over alternative search engines.

When running the system, queries were submitted for each capitalised sequence and all of its substrings. The first 1000 results from *google* were processed in each case. A sample of output from the system is shown in Figure 2. The result of the method on processing substrings of `Internet Explorer` cannot be displayed due to space restrictions. The ten most frequent potential hypernyms are shown with their frequency of occurrence as returned by *google*.

Use of the Internet, rather then the documents in which NEs appear, as the source of potential hypernyms is justified because the documents used to test the NERO system in the present study are rather small, and the four patterns shown in Figure 1 are used very rarely within them. In fact, for these documents only 1.19% of the NEs appear in those patterns. This contrasts with 96.46% when the Internet is taken as the source of potential hypernyms. Intuitively, it is expected that a document specific approach would yield high precision, and be less affected by the problems of word sense ambiguity that the Internet based method is vulnerable to (see Section 4.2 for more details). Unfortunately, poor recall is the cost of applying patterns under a document specific framework.

Coverage could be improved under a document specific approach by identifying additional patterns. In its favour, this would allow for the exploitation of such features as punctuation in the patterns. The drawbacks are that additional manual effort would be required to formulate them, and there is a risk that such patterns would only apply well in the document for which they were identified. For these reasons, this approach was not implemented in the current study.

2.1.2 Clustering Hypernyms

Having obtained sets of potential hypernyms for all sequences of capitalised words in the input text, the system clusters the global set of hypernyms in an attempt to find the general types of NE that appear in that document. NEs will subsequently be classified on the basis of the resultant typology.

A hard, bottom-up, hierarchical clustering algorithm was used (Manning & Schütze 99). It is presented in Figure 3.

The similarity function *sim* is a group average

[2] www.google.com

```
CAPITALISED SEQUENCE: Internet Explorer

  SUBSTRING: Internet Explorer
      Y such as X: (file 173, browser 120, web 69, program 54, window 40,
                    application 35, internet 25, suv 16, software 16,
                    browser 421)
    X and other Y: (browser 175, application 77, window 54, Microsoft 46,
                    feature 38, software 38, web 37, internet 24, program 23,
                    key 8)
         Y like X: (browser 311, program 35, application 27, product 15,
                    software 14, window 10, page 7, somthing 6, thing 6,
                    Netscape 5)
     X or other Y: (browser 250, web 96, program 43, window 42, internet 38,
                    software 16, application 14, compatible-level 12, Microsoft 11,
                    product 10)
```

Figure 2: Hypernyms for Internet Explorer

```
• Given:
    – a set $H := \{\chi_1, ..., \chi_n\}$ of hypernyms
    – a group-average similarity function $sim$,
      described in Section 2.1.2.
    – $\Upsilon = 1$

• for $i := 1$ to $n$ do
  $h_i := \{\chi_i\}$ end
  $C := \{h_1, ...h_n\}$
  $j := n + 1$

• while $\Upsilon > \tau$
    1. $\Upsilon := max_{(h_u,h_v) \in C \times C} sim(h_u, h_v)$
    2. $(h_{n_1}, h_{n_2}) := argmax_{(h_u,h_v) \in C \times C} sim(h_u, h_v)$
    3. $h_j := h_{n_1} \cup h_{n_2}$
    4. $C := C \setminus \{h_{n_1}, h_{n_2}\} \cup \{h_j\}$
    5. $j := j + 1$
```

Figure 3: The alorithm used to cluster hypernyms

method that assesses similarity between hypernyms on the basis of their taxonomic similarity in WordNet (Fellbaum 98). Taxonomic similarity was computed using a measure known as *Learning Accuracy* which was implemented to assist evaluation in (Hahn & Schnattinger 98). The clustering process is halted when the similarity measure for the two most similar clusters proposed for merging drops below a threshold, τ. Empirical observation indicated that a threshold of 0.5 was suitable. The stopping condition is set in an attempt to prevent hypernyms indicative of distinct types from being merged.

This type of clustering algorithm was suitable for the task because no information on the desired properties of the set of derived types is available *a priori.*

```
0.5 SURFACE:
  port, interface, panel, window,
  workstation, machine, server,
  computer, ski, dial, storage, memory,
  store, surface, device
```

Figure 4: The cluster labelled SURFACE

2.1.3 Labelling Clusters

The WordNet package (Fellbaum 98) was used in order to assign easy-to-read labels to the clusters resulting from the algorithm described in Section 2.1.2. For all senses of all words in a cluster, increasingly general hypernyms were listed. These lists were compared within the cluster and the most specific hypernym common to all words[3] was used to label the cluster as a whole. Each label was assigned a number, hereafter referred to as *height* that is the mean depth of the words in the cluster measured from the common hypernym. This measure is used to weight the classification of named entities, as described in Section 2.3. When no such common hypernym exists, the cluster is simply labelled MISCn where n is a unique reference number for that cluster.

When testing NERO, it was found that the cluster labels were often unhelpful, as illustrated in Figure 4. The clusters themselves were referred to in the evaluation process to determine whether or not a NE had been classified appropriately.

2.2 Identification of Named Entities (Capitalised Word Normalisation)

Capitalisation is one signal that can distinguish NEs from other phrases in texts. Unfortunately,

[3] Not necessarily all senses.

it is also used in the general layout of documents, indicating the start of sentences or dialogue, section headings, or instructions in block capitals. For this reason it is necessary to disambiguate capitalisation to determine whether a given word is normally capitalised in all contexts, or whether the capitalisation of a given word is context dependent. This disambiguation is referred to as *normalisation* in (Mikheev 00).

NERO exploits a method for normalisation that uses memory based learning (Daelemans *et al.* 01). Each instance of a capitalised word in the training data is associated with a vector of feature values and a binary classification (`NORMALLY_CAPITALISED` or `NOT_NORMALLY_CAPITALISED`). Features appearing in the vectors include

1. positional information,
2. the proportion of times the word is capitalised in the document,
3. the proportion of times the word is sentence initial in the document and in the BNC (Burnard 95),
4. whether the instance appears in a gazetteer of person names or, following (Mikheev 00), in a list of the top 100 most frequent sentence initial words in the BNC,
5. the part of speech of the word and the surrounding words,
6. agreement of the instance's grammatical number with the following verb *to be* or *to have.*

Grammatical information such as *part of speech* and *number* is available via the use of Conexor's FDG Parser (Tapanainen & Järvinen 97).

The training data contains 3168 instances. The method was evaluated using ten-fold cross validation. It obtained an overall precision of 98.63% and recall of 98.51% for `NORMALLY_CAPITALISED` instances and 100% precision and 98.31% recall for `NOT_NORMALLY_CAPITALISED` instances.

2.3 Classification of Named Entities

Classification of NEs exploits the derived typology. The typology is extended to mark the beginnings and ends of spans of NEs.

The approach described here assumes that normally capitalised sequences of words (Section 2.2) correspond to NEs. A NE is either identical to, or is a substring of, one of the set of capitalised sequences in the document. Each NE can be associated with a capitalised sequence that covers its position in the document.

If T is a type that subsumes hypernyms $\{t_1, ..., t_m\}$ and ϕ is a coefficient inversely proportional to the *height* of T, let w be a word that matches or is a substring of a capitalised sequence C. Further, let C have hypernyms $\{c_1, ..., c_n\}$ where each hypernym has a frequency fc_i.

The likelihood that w should be classified as T is given by:

$\sum_{j=1}^{m} \sum_{i=1}^{n} \phi.g(c_i, t_j).fc_i$

$g(c_i, t_j)$ is a function that maps to 1 when c_i is identical to t_j, and maps to 0 otherwise.

Having computed the likelihood for all types, w is classified as belonging to the one for which this measure is greatest. If no hypernym was obtained for the complete phrase C, then hypernyms of substrings of C that contain w are used instead.

3 The Test Corpus

The creation of evaluation data is very difficult for the non-expert due to the "open" nature of the NE typology. For the pilot study presented in this paper, just nine documents were hand annotated and an assessment of NERO's performance was made against this human annotated data.

Of the nine texts, eight were from the SEMCOR corpus (Landes *et al.* 98) and one was a technical document, *Windows Help File* (*win*). This was chosen in order to demonstrate system performance on a document containing many NEs of a type not found in the MUC-7 NE typology.

One point to be made about the documents taken from SEMCOR is that these are extracts from larger texts, and are thus incomplete. As will be noted in Section 4.1 this has some unfortunate consequences.

4 Evaluation

The system is evaluated with respect to its ability to identify NEs using text normalisation, to derive an appropriate typology of NEs for a given document, and to classify NEs in line with the derived typology.

Quantitative and automatic evaluation of the derived typology and the classification of NEs with respect to that typology relies on large

DOC	#WORDS	#NEs	GENRE
a01	1944	258	Legal
j53	1980	2	Psych.
j59	1959	55	Art
k09	2005	92	Lit.
k12	2020	129	Lit.
k22	2024	137	Lit.
k25	1985	29	Lit.
n05	2051	75	Lit.
win	2884	274	Tech.
TOT	18852	1051	-

Table 1: Characteristics of the documents used to test NERO

amounts of annotated data. Such data is difficult to produce because elements must be assigned types that are not known to annotators *a priori*. We undertook a small-scale manual evaluation of the system's performance.

4.1 Evaluating Normalisation

Some substantive evaluation has been performed for the capitalised word normalisation system (Section 2.2). That method was applied to the documents used to test NERO. Overall, for the capitalised words in the corpus, the system was able to identify normally capitalised words with a precision of 97.48% and recall of 88.39%. For words that are not normally capitalised, the figures were 100% and 97.11% respectively.

The performance of the normalisation system on the nine test documents was significantly worse than was suggested by ten-fold cross validation based on the training data. Part of the problem is that the literary documents in the test corpus contain a lot of dialogue. Here, new sentences are introduced using quotation marks without ending the introducing sentence with a full stop.

In addition, as noted in Section 3, the documents from SEMCOR (Landes *et al.* 98) are incomplete, and many PERSON NEs are referred to using only a surname or nickname. The full name may have been introduced earlier in the text, but this evidence is missing from the extract available in SEMCOR. This affected NERO's performance as it rendered the gazetteers used in Feature 4 redundant in many cases. While most systems for capitalised word normalisation are able to correctly classify nicknames and surnames when they appear, these successful classifications are usually facilitated by the appearance of the full names elsewhere in the document (Mikheev 00).

4.2 Evaluating Typology Derivation

Table 2 shows, for each file in the test corpus, the set of NE types that were manually annotated in each document (NE TYPES) and the set of NE types derived by NERO (DERIVED TYPOLOGY). Evaluation is problematic because manual annotation may require a high level of expertise within a given domain in order to classify new types of entity, e.g. menu items in software technical manuals, and non-experts will tend to assign general types such as `ARTIFACT` to these NEs. Hiring experts will be an expensive undertaking within the context of annotation in the open-domain. The typology induced by NERO may actually assist the non-expert annotator in selecting an appropriate type for a given NE. With respect to the NE TYPES, several match those that are used in MUC-7 but there are additional types used here. NAT_LAN covers NEs that refer to nationalities or languages. The titles of creative works such as paintings or books are marked CTVE_TTL. The names of menu items or buttons in computer software are marked OPT_BUT.

A quantitative assessment was made of system performance in the derivation of a typology. A large number of clusters results from the clustering phase, but the statistical classification method causes only a small subset of the total number of clusters to be evidenced in the output classification. For this crucial subset, the precision and recall of the clustering method was calculated. Precision is defined here as the ratio of the number of machine derived clusters that correlate well with a human derived type, to the total number of machine derived clusters. Recall is defined as the ratio of the number of machine derived clusters that correlate well with a human derived type to the total number of human derived types annotated in the key file. When calculating these figures, it was noted that in many cases (27 out of 66), the machine derived clusters were seen to correlate only partially with human derived types. In these cases, the machine derived cluster was counted as a good match if more than half of the senses in the cluster were felt to be indicative of the human derived type. Otherwise, it was not counted. The precision of the clustering method is poor, at 46.97%. Recall is mediocre, at 67.39%. The performance of the clustering method will limit precision and recall in the NE classification task.

Table 2 can be inspected in order to check the degree of correlation between machine derived clusters and human derived types.

It can be noted that the type BODY derived for text *a01* is too general, and incorporates hypernyms that distinguish two important types, ORG and LOC. The type CONCEPT derived from document *k09* highlights the problem of word sense ambiguity. Here the hypernym *character* has been merged with words that share its symbolic, rather than human sense. It would be necessary to perform word sense disambiguation (WSD) to solve this problem.

4.3 Evaluating NE Classification

Table 3 summarises the performance of the system. The column *NORM ACC* indicates the accuracy with which the system is able to identify the NEs in a document. *#NEsIDd* gives the exact number of capitalised words identified by NERO. It can be compared with the column *#NEs* in Table 1 which shows the actual number of words used in NE references in the document. The remaining columns in Table 3 show the number of instances classified correctly (CORRECT) or incorrectly (INCORRECT). In some cases, correctness of the classification is open to interpretation and these are marked *UNCERTAIN*. An example of this is the classification of the *Windows* program as `CONCEPT`, a very general type that does include several pertinent hypernyms such as `product` but also many irrelevant ones such as `privilege` or `right`. In many cases, the queries submitted to *google* are unable to indicate any potential hypernyms for a capitalised sequence. When this happens, the system uses any hypernyms that have been found for substrings of the sequence. When no potential hypernyms are available for any substrings, then the instance remains unclassified. The frequency of such occurrences in a document appears in the column *UNCLASSIFIED*. Note that these cases are included in the figures under INCORRECT.

5 Related Work

Research activity in IE and NER since the mid-90s has left a large literary footprint. In the first instance, readers are directed to the proceedings of MUC-7 (Chinchor 98b) for a description and evaluation of competing NER systems for English. Similar competitions have been held with respect to Japanese in the IREX (Sekine 99) conferences.

The continuing influence of the MUC competitions on the fields of IE and NER is significant. Papers such as (Zhou & Su 02), and the shared task at the CoNLL Workshop at ACL 2003 address the classification of NEs on the basis of the typology originally used in MUC-7.

The method recently reported in (Collins 02) was trained on data in which an extended set of NE types is annotated. The system was assessed on its ability to classify test data in line with a typology derived from the training corpus. The classification model identifies name types such as PERSON, ORGANIZATION, and LOCATION, but also brand names, scientific terms, and event titles, as well as others that are not described in the paper. Despite this, the approach is still scenario specific and is unlikely to be able to cater to other domains. The typology used in (Collins 02) is incomplete with respect to the open domain.

The development of an extended NE hierarchy, including more than 150 NE types, is described in (Sekine *et al.* 02). The researchers involved in this work have also developed a classification system for the extended set of NEs. However, the system is based on manually proposed rules specific to each type of entity. From this perspective, despite covering an extended typology, the classification process cannot yet be regarded as fully automatic, or robust in the open domain.

As mentioned in Section 2.2, the domain-independent task of text normalization has been addressed before, and with greater accuracy than the system reported in this paper, by (Mikheev 00). He used a Maximum Entropy classification model which also incorporated information about abbreviations and the occurrence of capitalised sequences in the document.

6 Conclusion

This paper has presented a framework for NER in the open domain, and has described an implemented system, *NERO*, that embodies this framework (Section 2). It includes automatic components for derivation of a typology of NEs (Section 2.1), for normalisation of capitalised words (Section 2.2), and for classification of NEs in a given document (Section 2.3). The paper has described these components and conducted a small-scale evaluation (Section 4). Related work has also been reviewed (Section 5).

The unsupervised nature of the approach, its

DOC	NE TYPES	DERIVED TYPOLOGY
a01	ORG, TIME, LOC PERS	UNKNOWN, BODY (University, organization, establishment, area, community, ...), ENTERPRISE (industry, business, agency), TIME_PERIOD, CODE#COMPUTER_CODE, PERSON, LEGAL_DOC, OCCUPATION
j53	NAT_LAN, PERS	PERSON, COLLECTION#ACCUMULATION (population, information)
j59	NAT_LAN, PERS, TIME, CTVE_TTL,	COGNITION#KNOWLEDGE (project, system, organization), CAUSAL_AGENT (artist, cubist, painter), PHENOMENON (effect, event), UTILITY (purpose, function), KNOWLEDGE_BASE (chemisty, classics, discipline, subject), PRODUCT (work, paper, journal, theme), DEVICE (instrument, man, bowl, style, tool), VISUAL_COMMUNICATION (artwork, art, movement, gesture)
k09	LOC, PERS, NAT_LAN, CTVE_TTL	CONCEPT (character, item, place), PERSON (breed, minister, author, help), LEGAL_DOC (title, language, article, book), PHENOMENON (product, event)
k12	LOC,PERS,TIME, OTHER, ABBR, ART, NAT_LAN	ENTITY#PHYSICAL_THING (setting, position, location, area), PERSON (agent, voter, source, officer), PART#COMPONENT (item, member), COGNITIVE_CONTENT#MENTAL_OBJECT (program, thought, idea, object), BUILDING (hospital, institution), DUTY (task, function) TRAIT (behaviour, action, initiative), ATTRIBUTE (power, quality, character)
k22	PERS, OTHER, ORG, LOC, TIME, ART, NAT_LAN, ABBR	PERSON, CREATION (dance, phrase, song), STATEMENT (point, thing, confession), PHENOMENON (influence, event), PROPERTY (species, type, form), ORGANIZATION (faith, charity, hospital), PRODUCT (reference, novel, product), ACTIVITY (location, place, function), POLITICAL_UNIT (power, nation, state) ADMINISTRATIVE_DIVISION (county, city, area),
k25	LOC, PERS, CTVE_TTL, NAT_LAN, TIME, ORG	LANGUAGE_UNIT (name, title, document, article), TEXTUAL_MATTER (headline, lyric, poem), MEDIUM (paper, newspaper, source), PERSON (voter, singer, person), RECREATION (festival, celebration, game), TIME_PERIOD (holiday, day, period), REGION (jurisdiction, county, city)
n05	PERS, LOC, ART	ORGANISM (member, scientist, person), ARTIFACT (web, material, song, product), MESSAGE#SUBJECT_MATTER (question, fabrication, book, source)
win	ART, LOC, PERS, OPT_BUT, ORG, CTVE_TTL, ABBR, OTHER	CONCEPT (component, thing, product), SYSTEM (internet, web, network), WORK (utility, function, task), USER (user, customer, client), COMPUTER_CODE (browser, application, software), COMMUNICATION (protocol, warranty, message), CHANGE (change, connection, transfer), SURFACE (port, window, interface), HUMAN_ACTION (consideration, condition, provision), EVIDENCE (file, proof, identification), REGION (county, community, neighborhood), MISC12 (disposition, property), SYMBOL (icon, item, indicator), TRANSMISSION (email, media, transmission), ARTIFACT (can, facility, source), MISC18 (business, object), SOCIAL_GROUP (band, company, organization)

Table 2: A comparison of manually annotated NE types and NE types obtained by clustering and cluster labelling applied to the test documents

independence from annotated training data, and the fact that it has addressed the open-domain are all strengths of the system. The fact that, to a certain extent, NERO is able to derive appropriate typologies of NEs for a variety of documents is also a favourable aspect of its performance.

Unfortunately, these strengths are overbalanced by numerous weaknesses.

Firstly, the patterns used to collect suitable hypernyms for capitalised sequences are vulnerable to data sparseness. In many cases, no suitable hypernyms were identified by the system. This problem can be addressed by formulating additional patterns, and by using alternative IR technology that recognises punctuation symbols in queries. Despite the rapid growth of the Internet, this problem of data sparseness is unlikely to be eliminated. Alternative approaches will be required in order to obtain the semantic type of entities for which the current hypernym collection method fails.

As noted in Section 4.2, word sense ambiguity is a major problem in the hypernym collection and clustering processes that NERO performs. In future work, it will be interesting to assess the feasibility of using a method for WSD in the hypernym collection and clustering phases. Soft clustering algorithms, in which elements may belong

DOC	NORM ACC	#NEsIDd	CORRECT	INCORRECT	UNCERTAIN	UNCLASSIFIED
a01	91.13	227	64 (28.19)	146 (64.32)	17 (7.49)	96(42.29)
j53	96.92	3	1 (33.33)	2 (66.67)	0 (0)	0 (0)
j59	89.91	42	21 (50.00)	15 (35.71)	6 (14.28)	2 (4.76)
k09	93.88	46	21 (45.65)	25 (54.35)	0 (0)	5 (10.87)
k12	87.29	101	15 (14.85)	84 (83.17)	2 (1.98)	39 (38.61)
k22	92.47	90	18 (20.00)	65 (72.22)	7 (7.78)	57 (63.33)
k25	88.51	22	7 (31.82)	10 (45.45)	5 (22.73)	7 (31.81)
n05	98.29	37	26 (70.27)	11 (29.73)	0 (0)	4 (10.81)
win	92.48	253	167 (66.01)	39 (15.41)	47 (18.58)	6 (2.37)
TOT	92.25	821	340 (41.41)	397 (48.35)	84 (10.23)	216 (26.31)

Table 3: Performance of NERO on named entity classification tasks

to more than one cluster, are another potential solution to the problem of word sense ambiguity.

It will be useful to experiment with different classification methods for the identified NEs. The weighting on different types can be adjusted not only with respect to the height of the label, but also with respect to the size of the cluster, or information from WSD.

The evaluation reported in this paper has been insufficient. In future work it may be useful to apply the system to the MUC-7 data in order to assess the effectiveness of the typologies derived from those documents. An alternative approach will be to incorporate NERO into different IE systems and obtain extrinsic evaluation results.

The overall value of the framework proposed in this paper remains an open question. The current performance by NERO of the clustering, normalisation, and classification tasks does leave much scope for improvement. These areas must be further explored, improved, and new assessments made before there are sufficient grounds for rejecting the approach suggested here.

References

(Burnard 95) Lou Burnard. *Users Reference Guide British National Corpus Version 1.0.* Oxford University Computing Services, UK, 1995.

(Chinchor 98a) Nancy Chinchor. Muc-7 information extraction task definition. Technical report, National Institute of Standards and Technology, 1998.

(Chinchor 98b) Nancy A. Chinchor, editor. *Message Understanding Conference Proceedings.* Science Applications International Corporation. US, 1998. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

(Collins 02) Michael Collins. Ranking algorithms for named-entity extraction: Boosting and voted perceptron. In *40th Annual Meeting of the Association for Computational Linguistics Proceedings of the Conference (ACL-02)*, pages 489–496, Pennsylvania, USA, July 2002.

(Daelemans *et al.* 01) Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. Timbl: Tilburg memory based learner, version 4.0, reference guide. Technical Report ILK Technical Report 01-04, Tilburg University, 2001.

(Fellbaum 98) Christiane Fellbaum, editor. *WordNet: An Eletronic Lexical Database.* The MIT Press, 1998.

(Grishman 03) Ralph Grishman. Information extraction. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics.* Oxford University Press, 2003.

(Hahn & Schnattinger 98) Udo Hahn and Klemens Schnattinger. Towards text knowledge engineering. In *The Fifteenth National Conference on Artificial Intelligence (AAAI/IAAI)*, pages 524 – 531, Wisconsin, US, July 1998.

(Hearst 92) Marti Hearst. Automatic acquisition of hyponyms from large text corpora. In *The 14th International Conference on Computational Linguistics (COLING1992)*, Nantes, France, July 1992.

(Landes *et al.* 98) Shari Landes, Claudia Leacock, and Randee I. Tengi. Building semantic concordances. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 199–216. MIT Press, 1998.

(Manning & Schütze 99) Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, 1999.

(Mikheev 00) Andrei Mikheev. Document centered approach to text normalization. In *Proceedings of SIGIR-2000*, pages 136–143, 2000.

(Sekine 99) Satoshi Sekine, editor. *Information Retrieval and Extraction Exercise*, Japan, 1999. CSL Sony. http://www.csl.sony.co.jp/person/sekine/IREX/.

(Sekine *et al.* 02) Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1818–1824, Las Palmas de Gran Canaria, Spain, May 2002.

(Tapanainen & Järvinen 97) Pasi Tapanainen and Timo Järvinen. Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the 5th Conference of Applied Natural Language Processing (ANLP'97)*, pages 64 – 71, Washington D.C., US, 31 March – 3 April 1997.

(Zhou & Su 02) GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *40th Annual Meeting of the Association for Computational Linguistics Proceedings of the Conference (ACL-02)*, pages 473–480, Pennsylvania, USA, July 2002.