Language Independent NER using a Maximum Entropy Tagger

James R. Curran and **Stephen Clark**
School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh. EH8 9LW
{jamesc,stephenc}@cogsci.ed.ac.uk

Abstract

Named Entity Recognition (NER) systems need to integrate a wide variety of information for optimal performance. This paper demonstrates that a maximum entropy tagger can effectively encode such information and identify named entities with very high accuracy. The tagger uses features which can be obtained for a variety of languages and works effectively not only for English, but also for other languages such as German and Dutch.

1 Introduction

Named Entity Recognition[1] (NER) can be treated as a tagging problem where each word in a sentence is assigned a label indicating whether it is part of a named entity and the entity type. Thus methods used for part of speech (POS) tagging and chunking can also be used for NER. The papers from the CoNLL-2002 shared task which used such methods (e.g. Malouf (2002), Burger et al. (2002)) reported results significantly lower than the best system (Carreras et al., 2002). However, Zhou and Su (2002) have reported state of the art results on the MUC-6 and MUC-7 data using a HMM-based tagger.

Zhou and Su (2002) used a wide variety of features, which suggests that the relatively poor performance of the taggers used in CoNLL-2002 was largely due to the feature sets used rather than the machine learning method. We demonstrate this to be the case by improving on the best Dutch results from CoNLL-2002 using a maximum entropy (ME) tagger. We report reasonable precision and recall (84.9 F-score) for the CoNLL-2003 English test data, and an F-score of 68.4 for the CoNLL-2003 German test data.

[1]We assume that NER involves assigning the correct label to an entity as well as identifying its boundaries.

Incorporating a diverse set of overlapping features in a HMM-based tagger is difficult and complicates the smoothing typically used for such taggers. In contrast, a ME tagger can easily deal with diverse, overlapping features. We also use a Gaussian prior on the parameters for effective smoothing over the large feature space.

2 The ME Tagger

The ME tagger is based on Ratnaparkhi (1996)'s POS tagger and is described in Curran and Clark (2003) . The tagger uses models of the form:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(x, y)\right) \tag{1}$$

where y is the tag, x is the context and the $f_i(x, y)$ are the *features* with associated weights λ_i. The probability of a tag sequence $y_1 \ldots y_n$ given a sentence $w_1 \ldots w_n$ is approximated as follows:

$$p(y_1 \ldots y_n | w_1 \ldots w_n) \approx \prod_{i=1}^{n} p(y_i | x_i) \tag{2}$$

where x_i is the context for word w_i. The tagger uses beam search to find the most probable sequence given the sentence.

The features are binary valued functions which pair a tag with various elements of the context; for example:

$$f_j(x, y) = \begin{cases} 1 & \text{if } \mathtt{word}(x) = \mathtt{Moody} \ \& \ y = \mathtt{I\text{-}PER} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$\mathtt{word}(x) = \mathtt{Moody}$ is an example of a *contextual predicate*.

Generalised Iterative Scaling (GIS) is used to estimate the values of the weights. The tagger uses a Gaussian prior over the weights (Chen et al., 1999) which allows a large number of rare, but informative, features to be used without overfitting.

Condition	Contextual predicate
$freq(w_i) < 5$	X is prefix of w_i, $\|X\| \leq 4$ X is suffix of w_i, $\|X\| \leq 4$ w_i contains a digit w_i contains uppercase character w_i contains a hyphen
$\forall w_i$	$w_i = X$ $w_{i-1} = X, w_{i-2} = X$ $w_{i+1} = X, w_{i+2} = X$
$\forall w_i$	$\text{POS}_i = X$ $\text{POS}_{i-1} = X$, $\text{POS}_{i-2} = X$ $\text{POS}_{i+1} = X$, $\text{POS}_{i+2} = X$
$\forall w_i$	$\text{NE}_{i-1} = X$ $\text{NE}_{i-2}\text{NE}_{i-1} = XY$

Table 1: Contextual predicates in baseline system

3 The Data

We used three data sets: the English and German data for the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003) and the Dutch data for the CoNLL-2002 shared task (Tjong Kim Sang, 2002). Each word in the data sets is annotated with a named entity tag plus POS tag, and the words in the German and English data also have a chunk tag. Our system does not currently exploit the chunk tags.

There are 4 types of entities to be recognised: persons, locations, organisations, and miscellaneous entities not belonging to the other three classes. The 2002 data uses the IOB-2 format in which a B-XXX tag indicates the first word of an entity of type XXX and I-XXX is used for subsequent words in an entity of type XXX. The tag O indicates words outside of a named entity. The 2003 data uses a variant of IOB-2, IOB-1, in which I-XXX is used for all words in an entity, including the first word, unless the first word separates contiguous entities of the same type, in which case B-XXX is used.

4 The Feature Set

Table 1 lists the contextual predicates used in our baseline system, which are based on those used in the Curran and Clark (2003) CCG supertagger. The first set of features apply to *rare* words, i.e. those which appear less than 5 times in the training data. The first two kinds of features encode prefixes and suffixes less than length 5, and the remaining rare word features encode other morphological characteristics. These features are important for tagging unknown and rare words. The remaining features are the word, POS tag, and NE tag history features, using a window size of 2. Note that the $\text{NE}_{i-2}\text{NE}_{i-1}$ feature is a composite feature of both the previous and previous-previous NE tags.

Condition	Contextual predicate
$freq(w_i) < 5$	w_i contains period w_i contains punctuation w_i is only digits w_i is a number w_i is {upper,lower,title,mixed} case w_i is alphanumeric length of w_i w_i has only Roman numerals w_i is an initial (X.) w_i is an acronym (ABC, A.B.C.)
$\forall w_i$	memory NE tag for w_i unigram tag of w_{i+1} unigram tag of w_{i+2}
$\forall w_i$	w_i in a gazetteer w_{i-1} in a gazetteer w_{i+1} in a gazetteer
$\forall w_i$	w_i not lowercase and $f_{\text{lc}} > f_{\text{uc}}$
$\forall w_i$	unigrams of word type bigrams of word types trigrams of word types

Table 2: Contextual predicates in final system

Table 2 lists the extra features used in our final system. These features have been shown to be useful in other NER systems. The additional orthographic features have proved useful in other systems, for example Carreras et al. (2002), Borthwick (1999) and Zhou and Su (2002). Some of the rows in Table 2 describe sets of contextual predicates. The w_i *is only digits* predicates apply to words consisting of all digits. They encode the length of the digit string with separate predicates for lengths 1–4 and a single predicate for lengths greater than 4. Titlecase applies to words with an initial uppercase letter followed by all lowercase (e.g. `Mr`). Mixedcase applies to words with mixed lower- and uppercase (e.g. `CityBank`). The length predicates encode the number of characters in the word from 1 to 15, with a single predicate for lengths greater than 15.

The next set of contextual predicates encode extra information about NE tags in the current context. The memory NE tag predicate (see e.g. Malouf (2002)) records the NE tag that was most recently assigned to the current word. The use of beam-search tagging means that tags can only be recorded from previous sentences. This memory is cleared at the beginning of each document. The unigram predicates (see e.g. Tsukamoto et al. (2002)) encode the most probable tag for the next words in the window. The unigram probabilities are relative frequencies obtained from the training data. This feature enables us to know something about the likely NE tag of the next word before reaching it.

Most systems use gazetteers to encode information about personal and organisation names, locations and trigger words. There is considerable variation in the size of the gazetteers used. Some studies found that gazetteers did not improve performance (e.g. Malouf (2002)) whilst others gained significant improvement using gazetteers and triggers (e.g. Carreras et al. (2002)). Our system incorporates only English and Dutch first name and last name gazetteers as shown in Table 6. These gazetteers are used for predicates applied to the current, previous and next word in the window.

Collins (2002) includes a number of interesting contextual predicates for NER. One feature we have adapted encodes whether the current word is more frequently seen lowercase than uppercase in a large external corpus. This feature is useful for disambiguating beginning of sentence capitalisation and tagging sentences which are all capitalised. The frequency counts have been obtained from 1 billion words of English newspaper text collected by Curran and Osborne (2002).

Collins (2002) also describes a mapping from words to *word types* which groups words with similar orthographic forms into classes. This involves mapping characters to classes and merging adjacent characters of the same type. For example, `Moody` becomes `Aa`, `A.B.C.` becomes `A.A.A.` and `1,345.05` becomes `0,0.0`. The classes are used to define unigram, bigram and trigram contextual predicates over the window.

We have also defined additional composite features which are a combination of atomic features; for example, a feature which is active for mid-sentence titlecase words seen more frequently as lowercase than uppercase in a large external corpus.

5 Results

The baseline development results for English using the supertagger features only are given in Table 3. The full system results for the English development data are given in Table 7. Clearly the additional features have a significant impact on both precision and recall scores across all entities. We have found that the word type features are particularly useful, as is the memory feature. The performance of the final system drops by 1.97% if these features are removed. The performance of the system if the gazetteer features are removed is given in Table 4. The sizes of our gazetteers are given in Table 6. We have experimented with removing the other contextual predicates but each time performance was reduced, except for the next-next unigram tag feature which was switched off for all final experiments.

The results for the Dutch test data are given in Table 5. These improve upon the scores of the best performing system at CoNLL-2002 (Carreras et al., 2002).

English DEV	PRECISION	RECALL	$F_{\beta=1}$
LOCATION	90.78%	90.58%	90.68
MISC	85.80%	81.24%	83.45
ORGANISATION	82.24%	80.09%	81.15
PERSON	92.02%	92.67%	92.35
OVERALL	**88.53%**	**87.41%**	**87.97**

Table 3: Baseline C&C results for English DEV data

English DEV	PRECISION	RECALL	$F_{\beta=1}$
LOCATION	91.69%	93.14%	92.41
MISC	88.15%	83.08%	85.54
ORGANISATION	83.48%	85.53%	84.49
PERSON	94.40%	95.11%	94.75
OVERALL	**90.13%**	**90.47%**	**90.30**

Table 4: No external resources results for Eng. DEV data

Dutch TEST	PRECISION	RECALL	$F_{\beta=1}$
LOCATION	84.42%	81.91%	83.15
MISC	78.46%	74.89%	76.64
ORGANISATION	77.35%	68.93%	72.90
PERSON	80.13%	90.71%	85.09
OVERALL	**79.91%**	**79.35%**	**79.63**

Table 5: Results for the Dutch TEST data

Gazetteer	ENTRIES
FIRST NAME	6,673
LAST NAME	89,836
$freq_{LC} > freq_{UC}$ LIST	778,791

Table 6: Size of Gazetteers

The final results for the English test data are given in Table 7. These are significantly lower than the results for the development data. The results for the German development and test sets are given in Table 7. For the German NER we removed the lowercase more frequent than uppercase feature. Apart from this change, the system was identical. We did not add any extra gazetteer information for German.

6 Conclusion

Our NER system demonstrates that using a large variety of features produces good performance. These features can be defined and extracted in a language independent manner, as our results for German, Dutch and English show. Maximum entropy models are an effective way of incorporating diverse and overlapping features. Our maximum entropy tagger employs Gaussian smoothing which allows a large number of sparse, but informative, features to be used without overfitting.

Using a wider context window than 2 words may improve performance; a reranking phase using global features may also improve performance (Collins, 2002).

Acknowledgements

We would like to thank Jochen Leidner for help collecting the Gazetteers. This research was supported by a Commonwealth scholarship and a Sydney University Travelling scholarship to the first author, and EPSRC grant GR/M96889.

English devel.	Precision	Recall	$F_{\beta=1}$
LOC	91.75%	93.20%	92.47
MISC	88.34%	82.97%	85.57
ORG	83.54%	85.53%	84.52
PER	94.26%	95.39%	94.82
Overall	90.15%	90.56%	90.35

English test	Precision	Recall	$F_{\beta=1}$
LOC	84.97%	90.53%	87.66
MISC	76.77%	75.78%	76.27
ORG	79.60%	79.41%	79.51
PER	91.64%	90.79%	91.21
Overall	84.29%	85.50%	84.89

German devel.	Precision	Recall	$F_{\beta=1}$
LOC	67.59%	70.11%	68.83
MISC	71.87%	48.81%	58.14
ORG	71.85%	50.60%	59.39
PER	81.69%	64.03%	71.79
Overall	73.29%	58.89%	65.31

German test	Precision	Recall	$F_{\beta=1}$
LOC	70.91%	71.11%	71.01
MISC	68.51%	46.12%	55.13
ORG	68.43%	50.19%	57.91
PER	88.04%	72.05%	79.25
Overall	75.61%	62.46%	68.41

Table 7: Full system results for the English and German development and test data sets.

References

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University.

John D. Burger, John C. Henderson, and William T. Morgan. 2002. Statistical Named Entity Recogizer Adaptation. In *Proceedings of the 2002 CoNLL Workshop*, pages 163–166, Taipei, Taiwan.

Xavier Carreras, Lluis Màrquez, and Lluis Padró. 2002. Named Entity Recognition using AdaBoost. In *Proceedings of the 2002 CoNLL Workshop*, pages 167–170, Taipei, Taiwan.

John Chen, Srinivas Bangalore, and K. Vijay-Shanker. 1999. New Models for Improving Supertag Disambiguation. In *Proceedings of the 9th Meeting of EACL*, Bergen, Norway.

Michael Collins. 2002. Ranking Algorithms for Named-Entity Extraction: Boosting and the Voted Perceptron. In *Proceedings of the 40th Meeting of the ACL*, pages 489–496, Philadelphia, PA.

James R. Curran and Stephen Clark. 2003. Investigating GIS and Smoothing for Maximum Entropy Taggers. In *Proceedings of the 11th Meeting of the European Chapter of the ACL*, Budapest, Hungary.

James R. Curran and Miles Osborne. 2002. A very very large corpus doesn't always yield reliable estimates. In *Proceedings of the 2002 CoNLL Workshop*, pages 126–131, Taipei, Taiwan.

Robert Malouf. 2002. Markov models for language-independent named entity recognition. In *Proceedings of the 2002 CoNLL Workshop*, pages 187–190, Taipei, Taiwan.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the EMNLP Conference*, pages 133–142, Philadelphia, PA.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, Edmonton, Canada.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2002*, Taipei, Taiwan.

Koji Tsukamoto, Yutaka Mitsuishi, and Manabu Sassano. 2002. Learning with Multiple Stacking for Named Entity Recognition. In *Proceedings of the 2002 CoNLL Workshop*, pages 191–194, Taipei, Taiwan.

GuoDong Zhou and Jian Su. 2002. Named Entity Recognition using an HMM-based Chunk Tagger. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 473–480, Philadelphia, PA.