Named Entity Recognition Using a Character-based Probabilistic Approach

Casey Whitelaw and **Jon Patrick**
Language Technology Research Group
Capital Markets Co-operative Research Centre
University of Sydney
{casey,jonpat}@it.usyd.edu.au

Abstract

We present a named entity recognition and classification system that uses only probabilistic character-level features. Classifications by multiple orthographic tries are combined in a hidden Markov model framework to incorporate both internal and contextual evidence. As part of the system, we perform a preprocessing stage in which capitalisation is restored to sentence-initial and all-caps words with high accuracy. We report f-values of 86.65 and 79.78 for English, and 50.62 and 54.43 for the German datasets.

1 Introduction

Language independent NER requires the development of a metalinguistic model that is sufficiently broad to accommodate all languages, yet can be trained to exploit the specific features of the target language. Our aim in this paper is to investigate the combination of a character-level model, orthographic tries, with a sentence-level hidden Markov model. The local model uses affix information from a word and its surrounds to classify each word independently, and relies on the sentence-level model to determine a correct state sequence.

Capitalisation is an often-used discriminator for NER, but can be misleading in sentence-initial or all-caps text. We choose to use a model that makes no assumptions about the capitalisation scheme, or indeed the character set, of the target language. We solve the problem of misleading case in a novel way by removing the effects of sentence-initial or all-caps capitalisation. This results in a simpler language model and easier recognition of named entities while remaining strongly language independent.

2 Probabilistic Classification using Orthographic Tries

Tries are an efficient data structure for capturing statistical differences between strings in different categories. In an orthographic trie, a path from the root through n nodes represents a string $a_1a_2\ldots a_n$. The n-th node in the path stores the occurrences (frequency) of the string $a_1a_2\ldots a_n$ in each word category. These frequencies can be used to calculate probability estimates $P(c \mid a_1a_2\ldots a_n)$ for each category c. Tries have previously been used in both supervised (Patrick et al., 2002) and unsupervised (Cucerzan and Yarowsky, 1999) named entity recognition.

Each node in an orthographic trie stores the cumulative frequency information for each category in which a given string of characters occurs. A heterogeneous node represents a string that occurs in more than one category, while a homogeneous node represents a string that occurs in only one category. If a string $a_1a_2\ldots a_n$ occurs in only one category, all longer strings $a_1a_2\ldots a_n\ldots a_{n+k}$ are also of the same category. This redundancy can be exploited when constructing a trie. We build minimum-depth MD-tries which have the condition that all nodes are heterogeneous, and all leaves are homogeneous. MD-tries are only as large as is necessary to capture the differences between categories, and can be built efficiently to large depths. MD-tries have been shown to give better performance than a standard trie with the same number of nodes (Whitelaw and Patrick, 2002).

Given a string $a_1a_2\ldots a_n$ and a category c an orthographic trie yields a set of relative probabilities $P(c \mid a_1)$, $P(c \mid a_1a_2)$, ..., $P(c \mid a_1a_2\ldots a_n)$. The probability that a string indicates a particular class is estimated along the whole trie path, which helps to smooth scores for rare strings. The contribution of each level in the trie is governed by a linear weighting function of the form

$$P(c \mid a_1 a_2 \ldots a_n) = \sum_{i=1}^{n} \lambda_i P(c \mid a_1 a_2 \ldots a_i)$$

$$\text{where } \lambda_i \in [0,1] \quad \text{and } \sum_{i=1}^{n} \lambda_i = 1$$

Tries are highly language independent. They make no assumptions about character set, or the relative importance of different parts of a word or its context. Tries use a progressive back-off and smoothing model that is well suited to the classification of previously unseen words. While each trie looks only at a single context, multiple tries can be used together to capture both word-internal and external contextual evidence of class membership.

3 Restoring Case Information

In European languages, named entities are often distinguished through their use of capitalisation. However, capitalisation commonly plays another role, that of marking the first word in a sentence. In addition, some sentences such as newspaper headlines are written in all-capitals for emphasis. In these environments, the case information that has traditionally been so useful to NER systems is lost.

Previous work in NER has been aware of this problem of dealing with words without accurate case information, and various workarounds have been exploited. Most commonly, feature-based classifiers use a set of capitalisation features and a sentence-initial feature (Bikel et al., 1997). Chieu and Ng used global information such as the occurrence of the same word with other capitalisation in the same document (Chieu and Ng, 2002a), and have also used a mixed-case classifier to teach a "weaker" classifier that did not use case information at all (Chieu and Ng, 2002b).

We propose a different solution to the problem of caseless words. Rather than noting their lack of case and treating them separately, we propose to restore the correct capitalisation as a preprocessing step, allowing all words to be treated in the same manner. If this process of case restoration is sufficiently accurate, capitalisation should be more correctly associated with entities, resulting in better recognition performance.

Restoring case information is not equivalent to distinguishing common nouns from proper nouns. This is particularly evident in German, where all types of nouns are written with an initial capital letter. The purpose of case restoration is simply to reveal the underlying capitalisation model of the language, allowing machine learners to learn more accurately from orthography.

We propose two methods, each of which requires a corpus with accurate case information. Such a corpus is easily obtained; any unannotated corpus can be used once

	Precision	Recall	$F_{\beta=1}$
lowercase	98.58%	96.58%	97.57
init-caps	89.76%	92.74%	91.22
allcaps	54.01%	92.33%	68.15
inner-caps	48.49%	80.00%	60.38

Table 1: Case restoration performance using an MD-trie, English.

sentence-initial words and allcaps sentences have been excluded. For both languages, the training corpus consisted of the raw data, training and test data combined.

The first method for case restoration is to replace a caseless word with its most frequent form. Word capitalisation frequencies can easily be computed for corpora of any size. The major weakness of this technique is that each word is classified individually without regard for its context. For instance, "new" will always be written in lowercase, even when it is part of a valid capitalised phrase such as "New York".

The second method uses an MD-trie which, if allowed to extend over word boundaries, can effectively capture the cases where a word has multiple possible forms. Since an MD-trie is only built as deep as is required to capture differences between categories, most paths will still be quite shallow. As in other word categorisation tasks, tries can robustly deal with unseen words by performing classification on the longest matchable prefix.

To test these recapitalisation methods, the raw, training, and development sets were used as the training set. From the second test set, only words with known case information were used for testing, resulting in corpora of 30484 and 39639 words for English and German respectively. Each word was classified as either lowercase ("new"), initial-caps ("New"), all-caps("U.S."), or inner-caps ("ex-English"). On this test set, the word-frequency method and the trie-based method achieved accuracies of 93.9% and 95.7% respectively for English, and 95.4% and 96.3% in German. Table 1 shows the trie performance for English in more detail. In practice, it is usually possible to train on the same corpus as is being recapitalised. This will give more accurate information for those words which appear in both known-case and unknown-case positions, and should yield higher accuracy.

This process of restoring case information is language independent and requires only an unannotated corpus in the target language. It is a pre-processing step that can be ignored for languages where case information is either not present or is not lost.

NER	Precision	Recall	$F_{\beta=1}$
English devel.	94.56%	91.31%	92.91
English test	91.48%	88.16%	89.79
German devel.	79.95%	45.02%	57.60
German test	79.16%	49.30%	60.76

Table 2: Recognition performance.

	NER		NEC	
	seen	unseen	seen	unseen
Eng devel.	99.1%	92.7%	95.0%	71.6%
Eng test	98.7%	89.5%	94.1%	70.5%
German devel.	96.7%	73.7%	95.4%	80.7%
German test	97.2%	80.8%	95.6%	85.7%

Table 3: Accuracy on seen and unseen tokens.

	word-based	trie-based
English devel.	+0.67	+0.92
English test	+1.29	+0.90
German devel.	+0.44	+0.78
German test	-0.12	+0.26

Table 4: Improvement in f-score through restoring case.

4 Classification Process

The training data was converted to use the IOB2 phrase model (Tjong Kim Sang and Veenstra, 1999). This phrase model was found to be more appropriate to the nature of NE phrases in both languages, in that the first word in the phrase may behave differently to consecutive words. MD-Tries were trained on the prefix and suffix of the current word, and the left and right surrounding contexts. Each trie T_x produces an independent probability estimate, $P_{T_x}(c \mid context)$. These probabilities are combined to produce a single estimate

$$P(c \mid context) = \prod_{i=0}^{n} P_{T_i}(c \mid context)$$

These probabilities are then used directly as observation probabilities in a hidden Markov model (HMM) framework. An HMM uses probability matrices Π, A, and B for the initial state, state transitions, and symbol emissions respectively (Manning and Schütze, 1999). We derive Π and A from the training set. Rather than explicitly defining B, trie-based probability estimates are used directly within the standard Viterbi algorithm, which exploits dynamic programming to efficiently search the entire space of state assignments. Illegal assignments, such as an I-PER without a preceding B-PER, cannot arise due to the restrictions of the transition matrix.

The datasets for both languages contained extra information including chunk and part-of-speech information, as well as lemmas for the German data. While these are rich sources of data, and may help especially in the recognition phase, our aim was to investigate the feasibility of a purely orthographic approach, and as such no extra information was used.

5 Results

Table 2 shows how the system performs in terms of recognition. There is a large discrepancy between recognition performance for English and German. For German, it appears that there is insufficient morphological information in a word and its immediate context to reliably discriminate between NEs and common nouns. Precision is markedly higher than recall across all tests. The most common error in English was the misclassification of a single-term entity as a non-entity, while multi-word entities were more successfully identified.

Table 3 shows the overall performance difference between words present in the tagged training corpus and those that only occurred in the test set. For previously seen words, both recognition and classification perform well, aided by the variable depth of MD-tries. The progressive back-off model of tries is quite effective in classifying new tokens, achieving up to 85% accuracy in classification unseen entities. It is interesting to note that, given a successful recognition phase, German NEs are more successfully classified than English NEs.

The effects of heuristically restoring case information can be seen in Table 4. The contribution of recapitalisation is limited by the proportion of entities in caseless positions. Both the word-based method and the trie-based method produced improvements. The higher accuracy of the trie-based approach gives better overall performance.

The final results for each language and dataset are given in Table 5. Both English datasets have the same performance profile: results for the PER and LOC categories were markedly better than the MISC and ORG categories. Since seen and unseen performance remained quite stable, the lower results for the second test set can be explained by a higher percentage of previously unseen words. While MISC is traditionally the worst-performing category, the lowest results were for ORG. This pattern of performance was different to that for German, in which MISC was consistently identified less well than the other categories.

6 Conclusion

We have presented a very simple system that uses only internal and contextual character-level evidence. This highly language-independent model performs well on both seen and unseen tokens despite using only the supervised training data. The incorporation of trie-based estimates into an HMM framework allows the optimal tag sequence to be found for each sentence.

We have also shown that case information can be restored with high accuracy using simple machine learning techniques, and that this restoration is beneficial to named entity recognition. We would expect most NER systems to benefit from this recapitalisation process, especially in fields without accurate case information, such as transcribed text or allcaps newswire.

Trie-based classification yields probability estimates that are highly suitable for use as features in a further machine learning process. This approach has the advantage of being highly language-independent, and requiring fewer features than traditional orthographic feature representations.

References

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of ANLP-97*, pages 194–201.

Hai Leong Chieu and Hwee Tou Ng. 2002a. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 190–196.

Hai Leong Chieu and Hwee Tou Ng. 2002b. Teaching a Weaker Classifier: Named Entity Recognition on Upper Case Text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 481–488.

S. Cucerzan and D. Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Jon Patrick, Casey Whitelaw, and Robert Munro. 2002. SLINERC: The Sydney Language-Independent Named Entity Recogniser and Classifier. In *Proceedings of CoNLL-2002*, pages 199–202. Taipei, Taiwan.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing Text Chunks. In *Proceedings of EACL'99*, pages 173–179. Bergen, Norway.

Casey Whitelaw and Jon Patrick. 2002. Orthographic tries in language independent named entity recognition. In *Proceedings of ANLP02*, pages 1–8. Centre for Language Technology, Macquarie University.

English devel.	Precision	Recall	$F_{\beta=1}$
LOC	91.35%	89.06%	90.19
MISC	88.09%	79.39%	83.51
ORG	79.18%	81.66%	80.40
PER	92.21%	86.70%	89.37
Overall	88.20%	85.16%	86.65

English test	Precision	Recall	$F_{\beta=1}$
LOC	82.40%	85.07%	83.72
MISC	75.93%	72.36%	74.11
ORG	76.10%	71.88%	73.93
PER	89.25%	79.59%	84.15
Overall	81.60%	78.05%	79.78

German devel.	Precision	Recall	$F_{\beta=1}$
LOC	68.95%	49.45%	57.59
MISC	75.34%	32.67%	45.58
ORG	63.58%	39.81%	48.96
PER	77.11%	35.83%	48.93
Overall	70.40%	39.52%	50.62

German test	Precision	Recall	$F_{\beta=1}$
LOC	64.46%	48.02%	55.04
MISC	64.86%	30.30%	41.30
ORG	65.64%	44.24%	52.86
PER	85.63%	48.37%	61.82
Overall	71.05%	44.11%	54.43

Table 5: Final results for English and German, development and test sets.