Meta-Learning Orthographic and Contextual Models for Language Independent Named Entity Recognition

Robert Munro and **Daren Ler** and **Jon Patrick**
Language Technology Research Group
Capital Markets Co-operative Research Centre
University of Sydney
{rmunro,ler,jonpat}@it.usyd.edu.au

Abstract

This paper presents a named entity classification system that utilises both orthographic and contextual information. The random subspace method was employed to generate and refine attribute models. Supervised and unsupervised learning techniques used in the recombination of models to produce the final results.

1 Introduction

There are commonly considered to be two main tasks in named entity recognition, recognition (NER) and classification (NEC). As the features that best classify words according to the two tasks are somewhat disparate, the two are often separated. Attribute sets may be further divided into subsets through sub-grouping of attributes, sub-grouping of instances and/or the use of multiple classifying processes. While the use of multiple subsets can increase overall accuracy, the recombination of models has been shown to propagate errors (Carreras et al., 2002; Patrick et al., 2002). More importantly, the decision regarding the separation of attributes into various subsets is often a manual task. As it is reasonable to assume that the same attributes will have different relative levels of significance in different languages, using the same division of attributes across languages will be less than optimal, while a manual redistribution across different languages is limited by the users knowledge of those languages. In this paper, the division and subsequent recombination of subgroups is treated as a meta-learning task.

2 Feature Representation

It has been our intention to create linguistically driven model of named entity composition, and to search for the attribute representations of these linguistic phenomena that best suit inference by a machine learning algorithm.

It is important to note that a named entity is a label that has been consciously granted by some person or persons, and as these names are chosen rather than assigned randomly or evolved gradually, there are generalisations that may be inferred about the words that may be used for naming certain entity types (Allan, 2001; Kripke, 1972).

While generalisations relating to abstract connotations of a word may be difficult to infer, generalisations about the structure of the words are more emergent. As the use of a name stretches back in time, it stretches back to a different set of etymological constraints. It may also stem from another language, with a different orthographic structure, possibly representing a different underlying phonology. Foreign words are frequently named entities, especially in the domain of a newswire such as Reuters. In language in general it is also reasonable to assume that a foreign word is more likely to be an entity, as people are more likely to migrate between countries than prepositions. It is these generalisations of the structure of words that we have attempted to represent in the n-gram features.

Another emergent structural generalisation is that of capitalisation, as named entities are commonly expressed in title-case in European Languages. In this work it has been investigated as a preprocessing step.

The other features used were contextual features, such as observed trigger words, and the given part-of-speech and chunking tags.

In total, we selected 402 attributes from which the models were built.

2.1 Character N-Gram Modelling

The fundamental attribute of character n-gram modelling is the observed probability of a collocation of characters occurring as each of the category types. Individual n-grams, or aggregations of them, may be used as attributes in part of a larger data set for machine learning.

Modeling at the orthographic level has been shown to be a successful method of named entity recognition. Orthographic Tries (Cucerzan and Yarowsky, 1999; Whitelaw and Patrick, 2003; Whitelaw and Patrick, 2002) and character n-gram modelling (Patrick et al., 2002) are two methods for capturing orthographic features. While Tries give a rich representation of a word, they are fixed to one boundary of a word and cannot extend beyond unseen character sequences. As they are also a classifying tool in themselves, their integration with a machine learning algorithm is problematic, as evidenced by reduction of overall accuracy when processing a Trie output through a machine learner in Patrick et al. (2002). As such, Tries have not been used here. Although n-gram modelling has not always been successful as a lone method of classification (Burger et al., 2002), for the reasons outlined above it is a more flexible modelling technique than Tries.

To capture affixal information, we used N-Grams modelling to extract features for the suffixes and prefixes of all words for all categories.

For general orthographic information we used the average probability of all bi-grams occurring in a word for each category, and the value of the maximum and minimum probability of all bi-grams in a word for each category. To capture contextual information, these bigram attributes was also extracted across word boundaries, both pre/post and exclusive/inclusive of the current word, for different context windows.

All n-grams were extracted for the four entity types, location, person, organisation and miscellaneous, with the word level n-grams also extracted for NE recognition attributes using a IOE2 model.

The aggregate n-gram attributes (for example, the average probability of all the n-grams in a word belonging to a category), act as a memory based attribute, clustering forms with less then random variance. These most benefit agglutinative structures, such as the compound words common to German, as well as morphologically disparate forms, for example, 'Australia' and 'Australian'. Here, of all the n-grams, only the final one differs. While a stemming algorithm would also match the two words, stemming algorithms are usually based on language specific affixal rules and are therefore inappropriate for a language independent task. Furthermore, the difference between the words may be significant. The second of the two words, used adjectively, would most likely belong to the miscellaneous category, while the former is most likely to be a location.

2.2 Contextual Features

Other than the contextual n-gram attributes, contextual features used were: a bag of words, both pre and post an entity, the relative sentence position of the word, commonly observed forms, and observed collocational trigger words for each category.

2.3 Other Features

The part-of-speech and chunking tags were used with a context window of three, both before and after each word.

For the German data, an attribute indicating whether the word matched its lemma form or was unknown was also included.

An attribute indicating both the individual and sequential existence of a words in the gazetteer was included for both sets.

No external sources were used.

3 Normalising Case Information

As well as indicating a named entity, capitalisation may indicate phenomenon such as the start of a sentence, a title phrase or the start of reported speech. As orthographic measures such as n-grams are case sensitive, both in the building of the model and in classification, a preprocessing step to correctly reassign the case information was used to correct alternations caused by these phenomenon. To the best knowledge of the authors, the only other attempt to use computational inference methods for this task is Whitelaw and Patrick (2003). Here we assumed all words in the training and raw data sets that were not sentence initial, did not occur in a title sentence, and did not immediately follow punctuation were in the correct case. This amounted to approximately 10,000,000 words. From these, we extracted the observed probability of a word occurring as lowercase, all capitals, initial capital, or internal capital; the bi-gram distribution across these four categories; and the part-of-speech and chunking tags of the word. Using a decision graph (Patrick and Goyal, 2001), all words from the test and training sets were then either recapitalised or decapitalised according to the output. The results were 97.8% accurate, as indicated by the number of elements in the training set that were correctly re-assigned their original case.

The benefit of case-restoration for the English development set was $F_{\beta=1}$ 1.56. Case-restoration was not undertaken on the English test set or German sets. For consistency, the English development results reported in table 1 are for processing *without* case restoration. We leave a more thorough investigation of case restoration as future work.

4 Processing

In order to make classifications, we employ a meta-learning strategy that is a variant of stacking (Wolpert, 1992) and cascading (Gama and Brazdil, 2000) over an ensemble of classifiers. This classifier is described in two phases.

In the first phase, an ensemble of classifiers is produced by combining both the random subspace method (Ho, 1998) and bootstrap aggregation or bagging (Breiman, 1996).

In the random subspace method, subspaces of the feature space are formed, with each subspace trained to produce a classifier. Given that with n features, 2^n different subsets of features can be generated, not all possible subsets are created. Ho (1998) suggests that the random subspace method is best suited for problems with high dimensionality. Furthermore, he finds that the method works well where there exists a high degree of redundancy across attributes, and where the prior knowledge about the significance of various attributes is unknown. It is also a useful method for limiting the impact of attributes that may cause the learner to overfit the data. This is especially important in the domain of newswires where the division between training and test sets is temporal, as topic shift is likely to occur.

From a different prespective, bagging produces different subsets or bootstrap replicates by randomly drawing with replacement, m instances from the original training set. Once again, each bag is used to produce a different classifier.

Both techniques share the same fundamental idea of forming multiple training sets from a single original training set. An unweighted or weighted voting scheme is then typically adopted to make the ultimate classification. However, in this paper, as the second phase of our classifier, an additional level of learning is performed. For each training instance, the class or category probability distributions produced by the underlying ensemble is used in conjunction with the correct classification to train a new final classifier. The category probability distributions may be seen as meta-data that is used to train a meta-learner.

Specifically, given n features $A_1, A_2, ..., A_n$ and m training instances $I_1, I_2, ..., I_m$,we may then randomly form l different training subsets $S_1, S_2, ..., S_l$, with each S_i containing a random subset of both attributes and training instances. Given a learning algorithm L, each S_i is used to train L to produce l different classifiers $C_1, C_2, ..., C_l$. When tested, each C_i will produce a category probability distribution $C_i(D_1), C_i(D_2), ..., C_i(D_g)$ where g is the total number of categories. Then for each training instance h, the unified category probability distribution $\sum_{r=1}^{l} C_r(D_1), \sum_{r=1}^{l} C_r(D_2), ..., \sum_{r=1}^{l} C_r(D_g)$ in conjunction with the correct category for that instance CL_h is used to train L to produce the final classifier C'.

In our experimentation, we divided each data set into subsets containing approximately 65% of the original training set (with replication) and with 50 of the total 402 attributes. In total, the meta-learner utilised data generated from the combined output of 150 sub-classifiers. The choices regarding the number of subsets and their respective attribute and instance populations were made in consideration of both processing constraints and the minimum requirements in terms of the required original data. While increasing the number of subsets will generally increase the overall accuracy, obtaining an optimal subset size through automated experimentation would have been a preferable method, especially as the optimal size may differ between languages.

To eliminate subsets that were unlikely to produce accurate results across any language, we identified eight subtypes of attributes, and considered only those sets with at least one attribute from each. These were:

1. prefixal n-grams
2. suffixal n-grams
3. n-grams specifically modelling IOE2 categories
4. trigger forms occurring before a word
5. trigger forms occurring after a word
6. sentence and clausal positions
7. collocating common forms
8. the observed probability of the word and surrounding words belonging to each category type

To classify the various subsets generated as well as to train the final meta-learner, a boosted decision graph (Patrick and Goyal, 2001) is used.

5 Results

N-Grams, in various instances, were able to capture information about various structural phenomenon. For example, the bi-gram 'ae' occurred in an entity in approximately 96% of instances in the English training set and 91% in the German set, showing that the compulsion to not assimilate old forms of names 'Israel' and 'Michael' to something like 'Israil' and 'Michal' is more emergent than the constraint to maintain form. An example bi-gram indicating a word from a foreign language with a different phonology is 'cz', representing the voiced palatal fricative, which is not commonly used in English or German. The fact two characters were needed to represent one underlying phoneme in itself suggests this. Within English, the suffix 'gg' always indicates a named entity, with the exception of the word 'egg', which has retained both g's in accordance with the English constraint of content words being three or more letters long. All other word's with an etymological history of a 'gg' suffix such as 'beg' have assimilated to the shorter form.

The meta-learning strategy improved the German test set results by $F_{\beta=1}$ 9.06 over a vote across the classifiers. For English test set, this improvement was $F_{\beta=1}$ 0.40.

6 Discussion

The methodology employed was significantly more successful at identifying location and person entities (see table 1). The recalls for these values for English are especially high considering that precision is typically the higher value in named entity recognition. Although the lower value for miscellaneous entities was expected, due to the relatively smaller number of items and idiosyncrasies of the category membership, the significantly low values for organisations was surprising. There are three possible reasons for this: organisations are more likely than people or places to take their names from the contemporary lexicon, and are therefore less likely to contain orthographic structures able to be exploited by n-gram modelling; in the training set, organisations were relatively over represented in the errors made in the normalising of case information, most likely due to the previous reason; and organisations may be represented metonymically, creating ambiguity about the entity class.

As the difference that meta-learning made to German was very large, but to English very small (see Results), it is reasonable to assume that the individual English classifiers were much more homogeneous, indicating both that the attribute space for the individual classifiers for English were very successful, but only certain classifiers or combinations of them were beneficial for German. The flexibility of the strategy as a whole was successful when generalising across languages

English devel.	precision	recall	$F_{\beta=1}$
LOC	90.02%	92.76%	91.37
MISC	89.05%	78.52%	83.46
ORG	77.61%	82.48%	79.97
PER	88.48%	93.38%	90.86
Overall	86.49%	88.42%	87.44

English test	precision	recall	$F_{\beta=1}$
LOC	84.74%	88.25%	86.46
MISC	80.13%	70.66%	75.09
ORG	75.20%	79.77%	77.42
PER	82.98%	90.48%	86.57
Overall	80.87%	84.21%	82.50

German devel.	precision	recall	$F_{\beta=1}$
LOC	74.70%	68.76%	71.60
MISC	76.07%	59.80%	66.96
ORG	71.14%	63.17%	66.92
PER	76.87%	76.87%	76.87
Overall	74.75%	67.80%	71.11

German test	precision	recall	$F_{\beta=1}$
LOC	67.73%	69.76%	68.73
MISC	64.38%	57.46%	60.73
ORG	60.54%	54.98%	57.63
PER	78.95%	75.31%	77.09
Overall	69.37%	66.21%	67.75

Table 1: Results for English and German sets.

References

K. Allan. 2001. *Natural Language Semantics*. Blackwell Publishers, Oxford, UK.

L. Breiman. 1996. Bagging predictors. In *Machine Learning, 24(2)*, pages 123–140.

J. D. Burger, J. C. Henderson, and W. T. Morgan. 2002. Statistical Named Entity Recognizer Adaptation. In *Proceedings of CoNLL-2002*. Taipei, Taiwan.

X. Carreras, L. Marques, and L. Padro. 2002. Named Entity Extraction using AdaBoost. In *Proceedings of CoNLL-2002*. Taipei, Taiwan.

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

J. Gama and P. Brazdil. 2000. Cascade generalization. In *Machine Learning, 41(3)*, pages 315–343.

T. K. Ho. 1998. The Random Subspace Method for Constructing Decision Forests. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8)*.

S. Kripke. 1972. Naming and necessity. In *Semantics of Natural Language*, pages 253–355.

J. Patrick and I. Goyal. 2001. Boosted Decision Graphs for NLP Learning Tasks. In Walter Daelemans and Rémi Zajac, editors, *Proceedings of CoNLL-2001*, pages 58–60. Toulouse, France.

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

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

C. Whitelaw and J. Patrick. 2003. Named Entity Recognition Using a Character-based Probabilistic Approach. In *Proceedings of CoNLL-2003*. Edmonton, Canada.

D. Wolpert. 1992. Stacked generalization. In *Neural Networks, 5(2)*, pages 241–260.