Improved Text Categorisation for Wikipedia Named Entities

Sam Tardif and **James R. Curran** and **Tara Murphy**
School of Information Technologies
University of Sydney
NSW 2006, Australia
{star4245,james,tm}@it.usyd.edu.au

Abstract

The accuracy of named entity recognition systems relies heavily upon the volume and quality of available training data. Improving the process of automatically producing such training data is an important task, as manual acquisition is both time consuming and expensive. We explore the use of a variety of machine learning algorithms for categorising Wikipedia articles, an initial step in producing the named entity training data. We were able to achieve a categorisation accuracy of 95% F-score over six coarse categories, an improvement of up to 5% F-score over previous methods.

1 Introduction

Named Entity Recognition (NER) is the task of identifying proper nouns, such as location, organisation and personal names, in text. It emerged as a distinct type of information extraction during the sixth Message Understanding Conference (MUC) evaluation in 1995, and was further defined and explored in the CoNLL NER evaluations of 2002 and 2003.

A set of four broad categories became the standard scheme for marking named entities (NEs) in text: person (PER), organisation (ORG), location (LOC), and miscellaneous (MISC). This scheme remains the most common, despite the development of more complex hierarchical category schemes (e.g. Brunstein (2002); Sekine et al. (2002)). Domain-specific category schemes have also been developed in many areas, such as astroinformatics (Murphy et al., 2006), bioinformatics (Kim et al., 2003) and the travel industry (Vijayakrishna and Sobha, 2008). We also extend the broad scheme with a DAB category for Wikipedia "disambiguation" pages — pages used to group articles with identical titles.

NER systems that categorise NEs under these schemes require a large amount of highly accurate training data to perform well at the task. Expert annotation is time consuming and expensive, so there is an imperative to generate this data automatically. Wikipedia is emerging as a significant resource due to its immense size and rich structural information, such as its link structure.

Nothman et al. (2009) introduced a novel approach to exploiting Wikipedia's internal structure to produce training data for NER systems. Their process involved an initial step of categorising all Wikipedia articles using a simple heuristic-based bootstrapping algorithm. Potential NEs were then identified as the words in an article's text that served as links to other Wikipedia articles. To label a NE they then used the category assigned to the article that it linked to.

We have explored the use of Naïve Bayes (NB) and support vector machines (SVMs) as replacements for the text categorisation approach taken by Nothman. This involved the conversion of heuristics used by Nothman into features as well as the incorporation of a number of new features. We demonstrate the superiority of our approach, providing a comparison of the individual text categorisation step to both Nothman's system and other previous research. Our state-of-the-art text categorisation system for Wikipedia achieved an improvement of up to 5% F-score over previous approaches.

2 Background

Accurate classifications for Wikipedia articles are useful for a number of natural language processing (NLP) tasks, such as question answering and NER. To produce article classifications for generating NER training data, Nothman et al. (2009) used a heuristic-based text categorisation system. This involved extracting the first head noun after the copula, head nouns from an article's categories, and incoming link information. They reported an F-score of 89% when evaluating on a set of 1,300 hand-labelled articles.

Dakka and Cucerzan (2008) explored the use of NB and SVM classifiers for categorising Wikipedia. They expanded each article's bag-of-words representation with disambiguated surface forms, as well as terms extracted from its first paragraph, abstract, and any tables present. They also extracted a small amount of context surrounding links to other Wikipedia articles.

Dakka and Cucerzan (2008) expanded their set of 800 hand-labelled articles using a semi-supervised approach, extracting training samples from Wikipedia "List" pages — pages that group other articles by type. For each "List" page containing a link to an article from the hand-labelled set they used the hand-labelled article's category to classify other articles on the list. They neglected to report how many training instances this left them with, but noted that they maintained the original class distribution of the hand-labelled data. They achieved an F-score of 89.7% with an SVM classifier and the category set PER, LOC, ORG, MISC and COM (for common nouns) when classifying their full article set.

We experimented with a combination of the classification techniques used by Dakka and Cucerzan (2008) and the feature extraction methods used by Nothman et al. (2009) and others (Ponzetto and Strube, 2007; Hu et al., 2008; Biadsy et al., 2008), focusing on the extraction of features from Wikipedia's rich metadata.

3 Data

Our annotation and experiments were all run on a March 2009 dump of Wikipedia. The mwlib[1] library was used to parse the Mediawiki markup and perform tasks such as expanding Wikipedia templates and extracting article categories and links. Punkt (Kiss and Strunk, 2006) and the NLTK (Loper and Bird, 2002) were used to tokenise the corpus.

[1] http://code.pediapress.com

New category	Example
PER	
Fictional	Popeye
Animal	Chupacabra
ORG	
Band	Blink-182
LOC	
Geological	Himalayas
MISC	
Franchise	Star Wars
Product → Software	Python

Table 1: Extensions to the BBN categories with examples

3.1 Annotation scheme

Annotation was performed under a slightly modified BBN category hierarchy (Brunstein, 2002). During annotation we discovered the need for a number of additional categories due to the large number of articles Wikipedia contains relating to popular culture, for example the new categories $Organisation \rightarrow Band$ and $Misc \rightarrow Work\ of\ Art \rightarrow TVSeries$ were quite common. We map these categories back to the "Other" subcategory of their parent category to allow accurate comparison with the original BBN scheme. Table 1 lists some of our new categories and gives an example for each.

We also discovered a number of ambiguities in the original BBN scheme. A number of Wikipedia articles were border cases in the BBN scheme — they related to a number of categories, but did not fit perfectly into any single one. The category $Misc \rightarrow Franchise$ is an example of an additional category to label articles such as "Star Wars" and "Final Fantasy". We also noticed some unresolvable overlaps in categories, such as $Location \rightarrow Location \rightarrow Island$ and $Location \rightarrow$ GPE $\rightarrow State$ for articles such as "Tasmania" and "Hawaii".

3.2 Manual annotation

A list of Wikipedia articles was selected for annotation based on several criteria. Given the large number of stub articles that exist within Wikipedia and

the poor representation of categories that selecting random articles would achieve, our list of articles was primarily based on their popularity as detailed by Ringland et al. (2009). We took into consideration the number of different language versions of Wikipedia that the article existed in to try and maximise the usefulness of our annotated data for further multi-lingual NLP tasks. We took a list of the most popular articles from August 2008 and checked for an article's existence on that list. We also considered the number of incoming links an article attracted. Based on these three criteria we produced a list of 2,311 articles for annotation.

Our resulting set of articles was of much higher quality than one that a random article selection process would produce. Random article selection fails to achieve good coverage of some important article categories, such as $Location \rightarrow GPE \rightarrow Country$ which annotators are likely to never come across using a random selection method. Random selection also yields a high number of stub articles with fewer features for a machine learner to learn from.

Our final set of Wikipedia articles was double-annotated with an inter-annotator agreement of 99.7% using the fine-grained category scheme, and an agreement of 99.87% on the broad NER categories. The remaining classification discrepancies were due to fundamental conflicts in the category hierarchy that could not be resolved. This set of hand-labelled articles will be released after publication.

4 Features for text categorisation

Our baseline system used a simple bag-of-words including tokens from the entire article body and the article title. This did not include tokens that appear in templates used in the generation of an article.

We then experimented with a number of different feature extraction methods, focusing primarily on the document structure for identifying useful features. Tokens in the first paragraph were identified by Dakka and Cucerzan (2008) as useful features for a machine learner, an idea stemming from the fact that most human annotators will recognise an article's category after reading just the first paragraph. We extended this idea by also marking the first sentence and title tokens as separate from other tokens, as we found that often the first sentence was all that was required for a human annotator to classify an article. We ran experiments limiting the feature space to these smaller portions of the document.

Wikipedia articles often have a large amount of metadata that helps in identifying an article's category, in particular Wikipedia categories and templates. Wikipedia categories are informal user defined and applied categories, forming a "folksonomy" rather than a strict taxonomy suitable for classification tasks, but the terms in the category names are usually strong indicators of an article's class. We extracted the list of categories applied to each article, tokenised the category names and added each token to the bag-of-words representation of the article.

Using the same reasoning we also extracted a list of each article's templates, tokenised their names, and expanded the article's bag-of-words representation with these tokens. Furthermore, we expanded the templates "Infobox", "Sidebar" and "Taxobox" to extract tokens from their content. These templates often contain a condensed set of important facts relating to the article, and so are powerful additions to the bag-of-words representation of an article. Category, template and infobox features were marked with prefixes to distinguish them from each other and from features extracted from the article body.

We reduced our raw set of features using a stop list of frequent terms, and removing terms with frequency less than 20 in a set of 1,800,800 articles taken from a separate Wikipedia dump. The assumption is that the majority of low frequency tokens will be typographical errors, or otherwise statistically unreliable data.

5 Results

We compared our two classifiers against the heuristic-based system described by Nothman et al. (2009) and the classifiers described by Dakka and Cucerzan (2008). We also tested a baseline system that used a bag-of-words representation of Wikipedia articles with rich metadata excluded. All SVM experiments were run using LIB-SVM (Chang and Lin, 2001) using a linear kernel with parameter $C = 2$. For NB experiments we used the NLTK.

The text categorisation system developed by Nothman et al. (2009) was provided to us by the authors, and we evaluated it using our hand-labelled

training data. Direct comparison with this system was difficult, as it has the ability to mark an article as "unknown" or "conflict" and defer classification. Given that these classifications cannot be considered correct we marked them as classification errors.

There were also a number of complications when comparing our system with the system described by Dakka and Cucerzan (2008): they used a different, and substantially smaller, hand-labelled data set; they did not specify how they handled disambiguation pages; they provided no results for experiments using only hand-labelled data, instead incorporating training data produced via their semi-automated approach into the final results; and they neglected to report the final size of the training data produced by their semi-automated annotation. However, these two systems provided the closest benchmarks for comparison.

We found that across all experiments the NB classifier performed best when using a bag-of-words representation incorporating the first sentence of an article only, along with tokens extracted from categories, templates and infoboxes. Conversely, the SVM classifier performed best using a bag-of-words representation incorporating the entire body of an article, along with category, template and infobox tokens. All experiment results listed were run with these respective configurations.

We evaluated our system on two coarse-grained sets of data: the first containing all articles from our hand-labelled set, and the second containing only those articles that described NEs. Table 2 lists results from the top scoring configurations for both the NB and SVM classifiers. The SVM classifier performed significantly better than the NB classifier.

Limiting the categorisation scheme to NE-only classes improved the classification accuracy for both classifiers, as the difficult NON class was excluded. With this exclusion the NB classifier became much more competitive with the SVM classifier.

Table 3 is a comparison of precision, recall and F-scores between our baseline and final systems, and the systems produced by Nothman et al. (2009) and Dakka and Cucerzan (2008). The difference between results from Nothman's system, our baseline and our full feature classifier were all found to be statistically significant at the $p < 0.05$ level. We performed this significance test using a stratified sampling approach outlined by Chinchor (1992).

(a) Full coarse-grained task

	NB			SVM		
Class	P	R	F	P	R	F
PER	72	98	83	99	92	95
ORG	70	94	80	95	91	93
LOC	97	99	98	99	99	99
MISC	69	84	76	90	88	89
NON	98	57	72	91	96	93
DAB	87	90	88	98	99	98
Micro Avg.	**83**	**83**	**83**	**95**	**95**	**95**

(b) NE-only task

	NB			SVM		
Class	P	R	F	P	R	F
PER	88	98	93	99	94	96
ORG	88	93	90	97	93	95
LOC	99	99	99	99	99	99
MISC	95	85	90	91	97	94
Micro Avg.	**94**	**94**	**94**	**97**	**97**	**97**

Table 2: NB and SVM results on coarse-grained problems.

Classifier	F
Nothman	91
Dakka	90
BASELINE	94
BEST	95

Table 3: Comparison with previous systems.

6 Conclusion

We exploited Wikipedia's rich document structure and content, such as categories, templates and infoboxes, to classify its articles under a categorisation scheme using NB and SVM machine learners. Our system produced state-of-the-art results, achieving an F-score of 95%, an improvement of up to 5% over previous approaches. These high quality classifications are useful for a number of NLP tasks, in particular named entity recognition.

Acknowledgements

We would like to thank the Language Technology Research Group and the anonymous reviewers for their helpful feedback. This work was partially supported by the Capital Markets Cooperative Research Centre Limited.

References

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. An unsupervised approach to biography production using wikipedia. In *Proceedings of ACL-08: HLT*, pages 807–815, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Ada Brunstein. Annotation guidelines for answer types. LDC2005T33, Linguistic Data Consortium, Philadelphia, 2002.

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

N Chinchor. Statistical significance of muc-6 results. In *Proceedings, Fourth Message Understanding Conference (MUC-4)*, 1992.

W Dakka and S Cucerzan. Augmenting wikipedia with named entity tags. In *Proceedings of IJC-NLP 2008*, 2008.

Jian Hu, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. Enhancing text clustering by leveraging wikipedia semantics. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 179–186, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl. 1):i180–i182, 2003.

Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32:485–525, 2006.

Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, July 2002.

Tara Murphy, Tara McIntosh, and James R. Curran. Named entity recognition for astronomy literature. In *Proceedings of the Australian Language Technology Workshop*, pages 59–66, Sydney, Australia, 2006.

Joel Nothman, Tara Murphy, and James R. Curran. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 612–620, Athens, Greece, March 2009. Association for Computational Linguistics.

S P Ponzetto and M Strube. Deriving a large scale taxonomy from wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, pages 1440–1445, 2007.

Nicky Ringland, James R. Curran, and Tara Murphy. Classifying articles in english and german wikipedias. In *Submitted to ALTA*, 2009.

S Sekine, K Sudo, and C Nobata. Extended named entity hierarchy. In *Proceedings of the LREC-2002*, 2002.

R. Vijayakrishna and L. Sobha. Domain focused named entity recognizer for tamil using conditional random fields. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, pages 59–66, Hyderabad, India, January 2008.