# **Classifying articles in English and German Wikipedia**

Nicky Ringland and Joel Nothman and Tara Murphy and James R. Curran

School of Information Technologies University of Sydney NSW 2006, Australia

{nicky, joel, tm, james}@it.usyd.edu.au

#### Abstract

Named Entity (NE) information is critical for Information Extraction (IE) tasks. However, the cost of manually annotating sufficient data for training purposes, especially for multiple languages, is prohibitive, meaning automated methods for developing resources are crucial. We investigate the automatic generation of NE annotated data in German from Wikipedia. By incorporating structural features of Wikipedia, we can develop a German corpus which accurately classifies Wikipedia articles into NE categories to within 1% *F*-score of the state-of-the-art process in English.

#### **1** Introduction

Machine Learning methods in Natural Language Processing (NLP) often require large annotated training corpora. Wikipedia can be used to automatically generate robust annotated corpora for tasks like Named Entity Recognition (NER), competitive with manual annotation (Nothman et al., The CoNLL-2002 shared task defined 2009). NER as the task of identifying and classifying the names of people (PER), organisations (ORG), places (LOC) and other entities (MISC) within text (Tjong Kim Sang, 2002). There has been extensive research into recognising NEs in newspaper text and domain-specific corpora, however most of this has been in English. The cost of producing sufficient NER annotated data required for training makes manual annotation unfeasible, and the generation of this data is even more important for languages other than English, where goldstandard corpora are harder to obtain.

German NER is especially challenging since various features used successfully in English NER, including proper noun capitalisation, do not apply to German language data, making NEs harder to detect and classify (Nothman et al., 2008). Furthermore, German has partially free word order which affects the reliability of contextual evidence, such as previous and next word features, for NE detection.

Nothman et al. (2008) devised a novel method of automatically generating English NE training data by utilising Wikipedia's internal structure. The approach involves classifying all articles in Wikipedia into classes using a featuresbased bootstrapping algorithm, and then creating a corpus of sentences containing links to articles identified and classified based on the link's target.

We extend the features used in Nothman et al. (2008) for use with German Wikipedia by creating new heuristics for classification. We endeavour to make these as language-independent as possible, and evaluate on English and German.

Our experiments show that we can accurately classify German Wikipedia articles at an F-score of 88%, and 91% for entity classes only, achieving results very close to the state-of-the-art method for English data by Nothman et al. (2008) who reported 89% on all and 92% on entities only. Nothman et al.'s (2009) NER training corpus created from these entity classifications outperforms the best cross-corpus results with gold standard training data by up to 12% F-score using CoNLL-2003-style evaluation. Thus, we show that it is possible to create free, high-coverage NE annotated German-language corpora from Wikipedia.

# 2 Background

The area of NER has developed considerably from the Message Understanding Conferences (MUC) of the 1990s where the task first emerged. MET, the Multilingual Entity Task associated with MUC introduced NER in languages other than English (Merchant et al., 1996) which had previously made up the majority of research in the area. The CONLL evaluations of 2002 and 2003 shifted the focus to Machine Language, and further multilingual NER research incorporated language-independent NER. CoNLL-2002 evaluating on Spanish and Dutch (Tjong Kim Sang, 2002) and CoNLL-2003 on English and German (Tjong Kim Sang and De Meulder, 2003).

The results of the CoNLL-2002 shared task showed that whilst choosing an appropriate machine learning technique affected performance, feature choice was also vital. All of the top 8 systems at CoNLL-2003 used lexical features, POS tags, affix information, previously predicted NE tags and orthographic features.

The best-performing CoNLL-2003 system achieved an *F*-score of 88.8% on English and 72.4% on German (Florian et al., 2003). It combined Maximum Entropy Models and robust risk minimisation with the use of external knowledge in the form of a small gazetteer of names. This was collected by manually browsing web pages for about two hours and was composed of 4500 first and last names, 4800 locations in Germany and 190 countries. Gazetteers are very costly to create and maintain, and so considerable research has gone into their automatic generation from online sources including Wikipedia (Toral and Muñoz, 2006).

The CoNLL-2003 results for German were considerably lower than for English, up to 25%difference in *F*-score (Tjong Kim Sang and De Meulder, 2003). The top performing systems all achieved *F*-scores on English more than 15 higher than on German.

## 2.1 German NER

German is a very challenging language for NER, because various features used in English do not apply. There is no distinction in the capitalisation of common and proper nouns so the number of word forms which must be considered as potential NEs is much larger than for languages such as English. German's partially free word order also means that surface cues, such as PER entities often preceding verbs of communication, are much weaker.

A final consideration is gender. The name Mark is likely to be on any list of German person names, but also makes up part of Germany's old currency, the Deutsche Mark, also known as D-Mark or just Mark (gender: female; *die*), and also has the meaning 'marrow' (gender: neuter; *das*). Whilst gender can sometimes disambiguate word senses, in more complicated sentence construction, gender distinctions reflected on articles and adjectives can change or be lost when a noun is used in different cases.

## 2.2 Cross-language Wikipedia

The cross-lingual link structure of Wikipedia represents a valuable resource which can be exploited for inter-language NLP applications. Sorg and Cimiano (2008) developed a method to automatically induce new inter-language links by classifying pairs of articles of two different langauges as connected by a inter-language link. They use a classifier utilising various text and graph-based features including edit distance between the title of articles and link patterns. They find that since the fraction of bidirectional links (cases where the English article e.g. Dog is linked to the German article Hund which is linked to the original English article) is around 95% for German and English, they can be used in a bootstrapping manner to find new inter-language links. The consistency and accuracy of the links was also found to vary, with roughly 50% of German language articles being linked to their English equivalents, and only 14% from English to German.

Richman and Schone (2008) proposed a system in which English Wikipedia article classifications are used to produce NE-annotated corpora in other languages, achieving an F-score of up to 84.7% on French language data, evaluated against human-annotated corpora with the MUC evaluation metric. So far there has been very little research into directly classifying articles in non-English Wikipedias.

#### 2.3 Learning NER

Machine learning approaches to NER are flexible due to their statistical data-driven approach, but training data is key to their performance (Nothman et al., 2009). The size and topical coverage of Wikipedia makes its text appropriate for training general NLP systems.

The method of Nothman et al. (2008) for transforming Wikipedia into an NE annotated corpus relies on the fact that links between Wikipedia articles often correspond to NEs. By using structural features to classify an article, links to it can be labelled with an NE class.

The process of deriving a corpus of NE annotated sentences from Wikipedia consists of two main sub-tasks: (1) selecting sentences to include in the corpus; and (2) classifying articles linked in those sentences into NE classes. By relying on redundancy, articles that are difficult to classify with confidence may simply be discarded.

This method of processing Wikipedia enables the creation of free, much larger NE-annotated corpora than have previously been available, with wider domain applicability and up-to-date, copyright free text. We focus on the first phase of this process: accurate classification of articles.

NLP tasks in languages other than English are disadvantaged by the lack of available data for training and testing. Developing more automated methods of language-resource generation which is independent of existing data sets is an important and challenging goal. We work towards generating high-coverage training corpora which can be used for a range of German NLP.

## 3 Data

To learn a classification of German Wikipedia articles, we labelled a corpus of English Wikipedia articles. Wikipedia's inter-language links allow us to then develop classifiers for all articles in English and German (or other language) Wikipedias. We use XML dumps of Wikipedia from March 2009 for both languages.

#### 3.1 Article selection

Both Nothman et al. (2008) and Dakka and Cucerzan (2008) have labelled collections of Wikipedia articles with gold standard classifica-

Rank	Article	Pageviews
1	2008 Summer Olympics	4 4 37 2 51
2	Wiki	4 030 068
3	Sarah Palin	4 004 853
4	Michael Phelps	3 476 803
5	YouTube	2685316
6	Bernie Mac	2013775
7	Olympic Games	2003678
8	Joe Biden	1 966 877
9	Georgia (country)	1 757 967
10	The Dark Knight (film)	1 427 277

Table 1: Most frequently viewed Wikipedia articles from August 2008, retrieved from http://stats.grok.se

Rank	Title	Inlinks
1	United States	543 995
2	Australia	344 969
3	Wikipedia	272 073
4	Association Football	241 514
5	France	227 464

Table 2: Most linked-to articles of English Wikipedia.

tions. Both of these consist of randomly selected articles, Dakka and Cucerzan's consisting of a random set of 800 pages, expanded by list co-occurrence. Nothman et al.'s data set initially consisted of 1100 randomly sampled articles from among all Wikipedia articles. This biased the sample towards entity types that are frequent in Wikipedia, such as authors and albums, but poorly represented countries, for example, which are important but are only a small proportion of Wikipedia's articles. A high number of the selected articles were stubs or other pages which were comparatively underdeveloped in structure and text. As a result, the data set was augmented with a further 200 articles, randomly sampled from among articles with at least 700 incoming links (in-links).

We took a more complex approach to choosing articles for inclusion in our data set, to ensure greater utility for multilingual Wikipedia tasks. We selected  $\sim$ 2300 articles from:

- the top 1000 most frequently viewed, based on August 2008 statistics (see Table 1), and
- the most linked-to articles (see Table 2),

with the constraint that they appear in at least the top 10 largest language Wikipedias (Table 3).

Wikipedia	Articles
English	3 500 000
German	950 000
French	850 000
Polish	650 000
Japanese	650 000
Italian	600 000
Dutch	550 000
Spanish	500 000
Portuguese	500 000
Russian	450 000

Table 3: Top ten Wikipedia languages by number of articles (nearest 50 000) as at September 2009.

Dataset	# articles	Paras	Sents	Cats
English 0805	1 2 9 6	3.0	36.4	4.6
English 0903	2 2 6 9	8.8	122.4	6.4
German 0903	2 2 6 9	4.8	84.1	3.3

Table 4: Average size (in paragraphs, sentences and categories) of Nothman et al.'s  $\sim$ 1300 labelled articles from 2008 and our  $\sim$ 2300 articles from March 2009.

We experimented with selecting the articles with the most inter-language links, but results were not meaningful; languages such as Volapuk may have fewer than 30 speakers, but more than 100,000 articles, most of which are stubs created and edited automatically. Reducing the languages of interest to 10 allowed us to focus on selecting more meaningful articles, using the criteria above. Although we deemed these criteria appropriate, they skew the corpus to events relevant in August 2008; the Summer 2008 Olympics, upcoming American Presidential Election and conflict between Russia and Georgia were prominent in the data.

In Table 4, we show that we succeed in selecting articles which are more substantial than the random sample of Nothman et al. (2008). Our method largely avoided the "long tail" of more obscure articles, such as old songs, sports players or archaeological finds, whose representation in a random sample is disproportionate to their utility.

#### 3.2 Annotation

Our corpus was created by manually classifying approximately 2300 English articles which had German equivalents, selected as described in section 3.1 using a custom annotation tool described

PER	LOC	ORG	MISC	NON	DIS	Total
271	648	229	392	650	79	2 2 6 9
12%	29%	10%	17%	29%	3%	100%

Table 5: Breakdown of manual classifications: People, Locations, Organisations, Miscellaneous, Common and Disambiguation.

in Tardiff et al. (2009). It allowed for an arbitrary number of annotators, and for multiple annotations to be compared.

Annotation was carried out using a hierarchical fine-grained tag-set based upon guidelines from the BBN Technologies' 150 answer types (Brunstein, 2002). Categories were able to be added into the hierarchy and either 'grown' or 'shrunk' to better fit the data as the annotators saw it. The ability to add categories is especially important when annotating Wikipedia because many categories such as types of works of art or products are not adequately covered in BBN.

The corpus was annotated using fine-grained categories, adding more information for use in future work, and enabling easier annotation, as they allow an annotator to classify a topic into a well-defined sub-category, which can then be uniformly mapped to a coarse-grained category. For example, all hotels can be classified as HOTEL, which then can be mapped to either ORG or LOC as decided after annotation.

The annotation process allowed for a high level of feedback to annotators, with statistics including inter-annotator agreement and a list of articles not uniformly classified available during annotation. This allowed annotators to quickly and easily identify digressions from one another.

All articles were double-annotated. After tagging the first 78 articles, we discussed conflicts and refined the annotation scheme. The two annotators then both classified a further 1100 articles each, achieving inter-annotator agreement of 97.5% on fine-grained tags, and 99.5% on coarsegrained tags. A further discussion and annotation round of the remaining  $\sim$ 1100 followed, and the final inter-annotator agreement was 99.7% on fine-grained tags and 99.9% on coarse-grained tags, creating a highly accurate corpus which we plan to release upon publication. Coarse-grained class distribution is given in Table 5.



Figure 1: A bootstrapping approach to article classification

## 4 Classification

Classification of Wikipedia's articles into semantic groupings is useful for applications such as named entity recognition (Kazama and Torisawa, 2007; Nothman et al., 2008) and ontology construction (Suchanek et al., 2007). The Wikipedia category hierarchy is a folksonomy and not directly suitable for NLP tasks. Instead, rule-based (Toral and Muñoz, 2006; Richman and Schone, 2008), semi-supervised (Nothman et al., 2008; Mika et al., 2008) and supervised (Dakka and Cucerzan, 2008) article classifiers have derived coarse-grained entity groupings or taxonomies.

Features used in classification are varied. Suchanek et al. (2007) used Wikipedia categories to map articles to WordNet, but noted that conceptual categories in English usually have plural head nouns (e.g. COASTAL CITIES IN AUSTRALIA) which describe the nature of member articles, as opposed to thematic categories like JAMES BOND. Richman and Schone (2008) scanned the hierarchy of categories for known phrases to classify articles into named entity categories.

Since an article's topic is usually defined in its first sentence, Toral and Muñoz (2006) try to match words from the opening sentence to a related class through the WordNet taxonomy. The specific use of the predicative head noun following a copula (is, were, etc.) in the first sentence was suggested by Kazama and Torisawa (2007) as a single feature by which articles may be grouped.

Other approaches utilise the co-occurrence of entities in lists (Watanabe et al., 2007; Bhole et al., 2007; Dakka and Cucerzan, 2008); presence of entities in particular fields of *infobox* templates which summarise the properties and relations of article topics (Mika et al., 2008); and bag-ofwords SVM classification (Dakka and Cucerzan, 2008; Bhole et al., 2007).

Although using different data sets, both Nothman et al. (2008) and Dakka and Cucerzan (2008) have reported F-scores of approximately 90% for classification into CONLL style entity categories.

## 4.1 Classifying Wikipedia articles

Nothman et al. (2008)'s bootstrapping classifier works as follows (see Figure 1): By initially associating features of each training instance with its gold-standard class label, an initial classification of all articles in Wikipedia is produced. Features that are consistently associated with a particular predicted class are then mapped to that class, including those not present in the hand-labelled data. These classification and mapping stages are then repeated, increasing feature coverage until the classifications are generally stable. Such an approach allows for high recall over sparse multivalued features like the Wikipedia category membership of each article. We extend their approach to German Wikipedia.

#### 4.2 Increasing non-entity recall

The following rules help to determine whether an article describes a named entity, or a non-entity topic (NON).

**Capitalisation** In English, all named entities are proper nouns, which are conventionally capitalised. This can be utilised by observing the capitalisation of all incoming links, with basic features allowing for determiners and nonconventional orthographies such as gzip or iPod.

In German, since all nouns are capitalised, this distinction is lost. Furthermore, adjectival forms



Figure 2: A portion of the Wikipedia article on the University of Sydney with some useful features marked.

including NEs of countries (eg. "Australian") are not capitalised in German ("australisch"), which means even a basic heuristic to check whether a link is a noun is not feasible.

**List identification** If the English article title begins List of or German, Liste, we mark it as NON.

**Disambiguation identification** Wikipedia's disambiguation articles list candidate referents for a particular title. The German page Mark lists amongst others, the name, substance (marrow), river, saints and various British tanks from WW1 of the same name. Most disambiguation pages are children of a category of DISAMBIGUATION, many have the word Disambiguation or Begriffsklärung in the title, and further information is available in the form of disambiguation templates.

#### 4.3 Bootstrapped features

For general classification, we extracted features from articles, which may each be mapped to an entity class. These mappings are produced by the bootstrapping process.

**Category nouns** Head nouns from Wikipedia category titles for both English and German, extracted using C&C tools (Curran and Clark, 2003) in English and the Tree-Tagger in German (Schmid, 1995) to POS-tag and chunk the sentences. In English, the category feature only applied to plural head nouns (and bigrams thereof) following Suchanek et

al.'s (2007) suggestion that these best represent ontology. Differences in both language and the structure of the German Wikipedia project invalidate this approach in German: conceputal categories are not plural, and forms that are bigrams in English are generally compound nouns. Hence we experimented with ASV toolbox (Chris Biemann and Holz, 2008) to extract a head morpheme. This allows PREMIERMINISTER (Prime Minister) and WISSENSCHAFTSMINISTER (Science Minister) to both be interpreted as MINIS-TER, and KERNBRENNSTOFFAUFBEREITUNGSAN-LAGE (nuclear fuel treatment facility) to become AN-LAGE (facility).

**Definition nouns** We term a *definition noun* to be the first noun following a copula in the first sentences, such as university in Figure 2. Definition nouns are extracted using POS-tagging and chunking as per category nouns, from articles which had been split into sentences and tokenised according to the method described in Nothman et al. (2008).

For each article, the number of category nouns mapped to each class is counted, and the most frequent class is used to label the article. If this is inconclusive, or the highest class leads by only one category, the definition noun is used to decide the class. Where there are no mapped category nouns or definition nouns available, or no winning class can be decided, the article class is marked as unknown (UNK).

An article classification is considered confident for use in bootstrapping if it is not labelled UNK, and if none of the features disagree (i.e. all category and definition features available map to the same class).

Itor	German 0903			Eng	lish (	903	English 0805		
nei	P	R	F	P	R	F	P	R	F
0	93	79	85	95	88	92	93	73	82
1	93	83	88	97	91	94	93	79	85
2	93	84	88	95	90	93	93	80	86
3	93	84	88	95	90	93	95	84	89

Table 7: Results of bootstrapping iterations on the held-out test set of German and English, compared to English 0805, as reported in Nothman (2008).

## 5 Results

We present results for our German Wikipedia classifier, exploring the effect of bootstrapping and feature variants in comparison to Nothman et al.'s (2008) English Wikipedia classifier.

We were able to achieve 88% *F*-score for German classification on a held-out test-set of 15% of the data (Table 6). These results are comparable to those presented by Nothman on English, but slightly lower than those using our larger annotated training corpus on a 2009 dump of English Wikipedia.

**Data Validation** The inter-language links between the German and English Wikipedias were checked and found to be reliable, with only two errors in links from English to German pages from the test set used for experimentation, which is consistent with the findings of Sorg and Cimiano (2008). Both of these were articles pointing to disambiguation pages: Nikon (describing the company) and lsosceles (describing the type of triangle). In the held-out training set, similar, though few, mis-links were found: the English article on French playwright Molière linking to Molière (1978), a French film depicting his life, and the German article Ryan Vikedal, a former member of the band Nickelback, links to the English article of the same name, which itself redirects to Nickelback. It should be noted that all of these examples were correctly classified by our process. When these errors were corrected, F-score improved by under 0.1%, showing that even with occasional noise, inter-wiki language links can be used to produced good-quality data. The results we present use a uncorrected test set.

**Bootstrapping** Bootstrapping was found to be less effective than in Nothman (2008) (see Table 7), where it was more needed to increase recall given less manually-labelled seed data. With the larger seed, bootstrapping proved more important on the German data than English, with recall increasing 5% compared to 2%, still falling short of the 11% increase found by Nothman. In our experiments, we found that the results were unchanging after the second feedback stage.

**Feature Analysis** In Table 6, we examine the effects of removing some classification features, and compare against the same process on English Wikipedia. In English, the capitalisation feature improves recall slightly, as opposed to the substantial increase found in Nothman's work; we might expect German, in which capitalisation is not used, to be disadvantaged by a similar amount.

Category nouns are seen to be by far the most important feature, especially in German. Our experiments to extract the morphology-based head from each category noun were an attempt to increase recall. We observed a slightly higher recall in the seed classification, but the bootstrapping process – also designed to improve recall – was more effective with the finer granularity of whole category noun features. This ultimately led to slightly reduced recall, leading us to use whole category nouns in our remaining experiments.

Definition nouns gave mixed results. In German they improved recall but had little effect on precision, while in English they improved precision and recall.

**Cross-validation** The results of ten-fold cross-validation are shown in Table 8, with a class breakdown. Our system left 8% of German and 6% of English Wikipedia articles unclassified (UNK). Nothman (2008) reports that 10% of articles were left unclassified. Our present work was able to classify a greater proportion due to our selection of more, higher-quality seed articles.

The German system performs very well on LOC and on MISC, which is known to be difficult to classify, achieving almost equivalent scores to English. The system also achieves a high F-score on PER. All of the false negatives when classifying people were on articles describing fictional characters such as Luzifer, Godzilla and Hermaphroditos. The error analysis of ORG also shows that we fail to correctly classify articles which the annotators also were unsure of, such as eBay and amazon.com, and Jedi. MISC often appeared incorrectly classified as ORG, showing the often blurred distinction between a product and the organisation which produces it (eg: Jeep and Airbus A380). The BBN guidelines also proved difficult for the classifier to adhere to, with 'attractions' such as the Nürburgring being classified as LOC not MISC.

Table 9 compares the precision, recall and F-score of English and German overall and on entity classes only. We also report the standard deviation of performance over the ten folds of cross-validation. The larger gap between all-class and entity class results in German reflects the low NON recall (76% as opposed to 90% in English), likely due to no available capitalisation feature.

Classification features	German 0903		English 0903			English 0805			
	P	R	F	P	R	F	P	R	F
All features	-	-	-	95	90	93	95	84	89
<ul> <li>Capitalisation</li> </ul>	93	84	88	96	89	92	92	80	85
+ Category morphology	93	83	88	-	-	-	-	-	-
<ul> <li>Definition nouns</li> </ul>	93	81	87	93	88	91	95	80	87
<ul> <li>Category nouns</li> </ul>	48	7	12	76	28	41	48	13	21

Table 6: Subtractive feature analysis on the held-out test set, comparing German Wikipedia with English (0903) performance, and the results reported by Nothman (2008) (English 0805).

Wikiped	P	R	F	
English 0903	All	94 ±2	$89 \pm 1$	$91 \pm 1$
	Entities	$98 \pm 1$	$89 \pm 2$	$93 \pm 1$
Common 0002	All	91 ±3	$84 \pm 3$	$88\pm 2$
German 0903	Entities	$97 \pm 2$	$87 \pm \! 4$	$92\pm\!\!3$

Table 9: Classification performance (average and standard deviation) over ten-fold cross-validation.

Class	07	Ger	man (	903	Engl	lish 09	903
	%	P	R	F	P	R	F
NON	29	85	76	80	86	90	88
DAB	3	92	99	96	100	90	95
LOC	29	98	95	96	99	97	98
MISC	17	89	71	78	97	67	79
ORG	10	93	85	89	97	91	94
PER	12	92	94	93	96	98	97

Table 8: Class distribution of manually classified articles and average results of ten-fold cross-validation.

## 6 Conclusion

Our work develops a semi-supervised classifier assigning German Wikipedia articles to named entity tags, in comparison to English Wikipedia. In doing so, we labelled a large corpus (2269 articles) of English Wikipedia pages, and validated the use of Wikipedia's inter-language links to transfer those training classifications to the smaller German encyclopedia.

In distinction from previous annotations of Wikipedia data, we produced a corpus with finegrained classes, extending on BBN's 150 answer types (Brunstein, 2002), and consisting of only articles which satisfy popularity criteria.

The classifier we have produced for German Wikipedia achieves very high precision (97%) and recall (87%) on entity classes. Due to differences between English and German language, orthography and Wikipedia editorial style, we had to modify the semantic and structural features previously used to classify English Wikipedia articles (Nothman et al., 2008). Our use of bootstrapping to spread this semantic knowledge to features unseen in training greatly improves performance in German, in which capitali-

sation features cannot be easily applied to distinguish NEs from non-entities, and in which there are fewer features available for classification, due to a smaller, less-developed Wikipedia.

We intend to improve the classifier by exploring further features, as well as the integrity of article resolution and inter-language links.

The results we have presented in German are only 3% F-score lower than on English articles and 1% F-score lower when only evaluating on NEs. The CoNLL-2003 shared task presented a 12% minimum reduction in performance for German NER when compared to English (Tjong Kim Sang and De Meulder, 2003). This substantial difference is due either to the difficulty of the NER task in German, or to the paucity of training data available in the CoNLL-2003 shared task, where the German training data marked only half as many NEs as the English corpus. By transforming the links in Wikipedia into entity annotations, we intend to generate large NE-annotated corpora, and to evaluate their use for learning German NER. Our highaccuracy classifier therefore reduces the need for expensive manual annotation in languages other than English where resources tend to be scarce.

#### Acknowledgments

We would like to thank members of 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 (CMCRC) and by Australian Research Council Discovery Project DP0665973. Joel Nothman was supported by a University of Sydney Vice-Chancellor's Research Scholarship and a CMCRC PhD Scholarship.

### References

- Abhijit Bhole, Blaž Fortuna, Marko Grobelnik, and Dunja Mladenić. 2007. Extracting named entities and relating them over time based on Wikipedia. *Informatica*, 31:463–468.
- Ada Brunstein. 2002. Annotation guidelines for answer types. LDC2005T33, Linguistic Data Consortium, Philadelphia.
- Gerhard Heyer Chris Biemann, Uwe Quasthoff and Florian Holz. 2008. Asv toolbox: a modular collection of language exploration tools. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 1760– 1767, Marrakech, Morocco, May. European Language Resources Association (ELRA).
- James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural Language Learning*, pages 164–167, Morristown, NJ, USA.
- Wisam Dakka and Silviu Cucerzan. 2008. Augmenting Wikipedia with named entity tags. *Proceedings* of the 3rd International Joint Conference on Natural Language Processing, pages 545–552.
- Radu Florian, Abe Ittycheriah, Hongyang Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 168–171.
- Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the* 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 698–707.
- Roberta Merchant, Mary Ellen Okurowski, and Nancy Chinchor. 1996. The multilingual entity task (MET) overview. In *Proceedings of the Tipster Text Program Phase II*, pages 445–447, Vienna, Virginia, May.
- Peter Mika, Massimiliano Ciaramita, Hugo Zaragoza, and Jordi Atserias. 2008. Learning to tag and tagging to learn: A case study on Wikipedia. *IEEE Intelligent Systems*, 23(5):26–33.
- Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. In *Proceedings of the Australian Language Technology Workshop*, pages 124–132.
- Joel Nothman, Tara Murphy, and James R. Curran. 2009. 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, March.

- Joel Nothman. 2008. *Learning Named Entity Recognition from Wikipedia*. School of Information Technologies, University of Sydney, Honours Thesis.
- Alexander E. Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1– 9.
- Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL-SIGDAT Workshop*, March.
- Philipp Sorg and Philipp Cimiano. 2008. Enriching the crosslingual link structure of wikipedia classification-based approach. *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence.*
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge — unifying WordNet and Wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, Banff, Alberta, Canada.
- Sam Tardiff, James R. Curran, and Tara Murphy. 2009. Improved text categorisation for Wikipedia named entities. In *Proceedings of the Australian Language Technology Workshop*.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147.
- Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 1–4.
- Antonio Toral and Rafael Muñoz. 2006. A proposal to automatically build and maintain gazetteers for named entity recognition using Wikipedia. In *Proceedings of the Workshop on NEW TEXT, 11th Conference of the European Chapter of the Association for Computational Linguistics.*
- Yotaro Watanabe, Masayuki Asahara, and Yuji Matsumoto. 2007. A graph-based approach to named entity categorization in Wikipedia using conditional random fields. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 649–657.