Transforming Wikipedia into Named Entity Training Data

Joel Nothman and **James R. Curran** and **Tara Murphy**
School of Information Technologies
University of Sydney
NSW 2006, Australia
{jnot4610,james,tm}@it.usyd.edu.au

Abstract

Statistical named entity recognisers require costly hand-labelled training data and, as a result, most existing corpora are small. We exploit Wikipedia to create a massive corpus of named entity annotated text. We transform Wikipedia's links into named entity annotations by classifying the target articles into common entity types (e.g. person, organisation and location). Comparing to MUC, CONLL and BBN corpora, Wikipedia generally performs better than other cross-corpus train/test pairs.

1 Introduction

Named Entity Recognition (NER), the task of identifying and classifying the names of people, organisations, locations and other entities within text, is central to many NLP tasks. The task developed from information extraction in the Message Understanding Conferences (MUC) of the 1990s. By the final two MUC evaluations, NER had become a distinct task: tagging the aforementioned proper names and some temporal and numerical expressions (Chinchor, 1998).

The CoNLL NER evaluations of 2002 and 2003 (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) focused on determining superior machine learning algorithms and feature models for multilingual NER, marking tags for person (PER), organisation (ORG), location (LOC) and miscellaneous (MISC; broadly including e.g. events, artworks and nationalities). Brunstein (2002) and Sekine et al. (2002) expanded this into fine-grained categorical hierarchies; others have utilised the WordNet noun hierarchy (Miller, 1998) in a similar manner (e.g. Toral et al. (2008)). For some applications, such as biotextmining (Kim et al., 2003) or astroinformatics (Murphy et al., 2006), domain-specific entity classification schemes are more appropriate.

Statistical machine learning systems have proved successful for NER. These learn terms and patterns commonly associated with particular entity classes, making use of many contextual, orthographic, linguistic and external knowledge features. They rely on annotated training corpora of newswire text, each typically smaller than a million words. The need for costly, low-yield, expert annotation therefore hinders the creation of more task-adaptable, high-performance named entity (NE) taggers.

This paper presents the use of Wikipedia[1]—an enormous and growing, multilingual, free resource—to create NE-annotated corpora. We transform links between encyclopaedia articles into named entity annotations (see Figure 1). Each new term or name mentioned in a Wikipedia article is often linked to an appropriate article. A sentence introducing Ian Fleming's novel Thunderball about the character James Bond may thus have links to separate articles about each entity. Cues in the linked article about Ian Fleming indicate that it is about a person, and the article on Thunderball states that it is a novel. The original sentence can then be automatically annotated with these facts. Millions of sentences may similarly be extracted from Wikipedia to form an enormous corpus for NER training.

[1] http://www.wikipedia.org

Having produced annotated text in this manner, it can be used to train an existing NER system. By training the C&C tagger (Curran and Clark, 2003) on standard annotated corpora and Wikipedia-derived training data, we have evaluated the usefulness of the latter. We have used three gold-standard data sets, and have found that tagging models built on each perform relatively poorly on the others. Our Wikipedia-derived corpora are usually able to exceed the performance of non-corresponding training and test sets, by up to 8.7% F-score. Wikipedia-derived training data may also be more appropriate than newswire for many purposes.

In a similar manner, free, large named entity-annotated corpora can be flexibly engineered for general or domain-specific tasks, allowing for NER without any manual annotation of text.

2 NER and Wikipedia

Following the CoNLL evaluations which focused on machine learning methods (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), work to improve NER performance has often involved the use of external knowledge. Since many tagging systems utilise categorised lists of known entities, some research has focused on their automatic extraction from the web (Etzioni et al., 2005) or Wikipedia (Toral et al., 2008), although Mikheev et al. (1999) and others have shown that larger NE lists do not necessarily correspond to increased NER performance. Nadeau et al. (2006) use such lists in an unsupervised NE recogniser, outperforming some entrants of the MUC Named Entity Task. Unlike statistical approaches which learn patterns associated with a particular type of entity, these unsupervised approaches are limited to identifying only common entities present in lists or those identifiable by hand-built rules.

External knowledge has also been used to augment supervised NER approaches. Kazama and Torisawa (2007) produced an F-score increase of 3% by including a Wikipedia-based feature in their NER system. Such approaches are nonetheless limited by the gold-standard data already available.

A less-common approach is the automatic creation of training data. An et al. (2003) extracted sentences containing listed entities from the web, and produced a 1.8 million word Korean corpus that gave similar results to manually-annotated training data. Richman and Schone (2008) used a method similar to that presented here in order to derive NE-annotated corpora in languages other than English. Their approach involves classifying English Wikipedia articles and using Wikipedia's inter-language links to infer classifications in other languages' articles. With these classifications they automatically annotate entire articles for NER training, and suggest that their results with a 340k-word Spanish corpus are comparable to 20k-40k words of gold-standard training data.

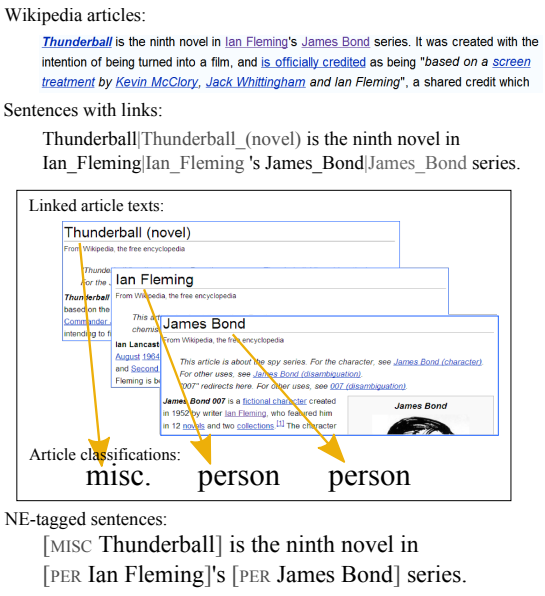

Figure 1: Deriving training sentences from Wikipedia text: sentences are extracted from articles; links to other articles are then translated to NE categories.

3 From Wikipedia to NE corpora

Wikipedia is a multilingual online encyclopedia written by many thousands of its users, and includes over 2.3 million articles in English alone. We take advantage of Wikipedia's links between articles to derive a NE-annotated corpus. Since around 74% of Wikipedia articles (see Table 2) describe topics falling under traditional entity classes, many of Wikipedia's links correspond to entity annotations in gold-standard NER training corpora. These links also disambiguate their referent, distinguishing David Jones, a department store, from David Jones, a British poet. In sum-

mary, an entity-tagged corpus may be derived by the following steps (see Figure 1):

1. Classify all articles into entity classes
2. Split Wikipedia articles into sentences
3. Label NEs according to link targets
4. Select sentences for inclusion in a corpus

This same approach could be applied to multiple languages, or different granularities or domains of NE categories. We use the standard CoNLL categories (LOC, ORG, PER, MISC) in order to facilitate evaluation.

4 Classifying Wikipedia articles

In order to label links according to their targets, we first must classify Wikipedia's articles into a fixed set of entity categories.

Many researchers have already tackled the task of classifying Wikipedia articles, often into named entity categories. The current state-of-the-art results (90% F-score) were achieved using bag-of-words with an SVM learner (Dakka and Cucerzan, 2008). They also make use of entities co-occurring in lists as a classification heuristic. While Wikipedia provides its own categorisation hierarchy, it has been described as a *folksonomy*, comparable to other collaborative online tagging. Suchanek et al. (2007) divide Wikipedia categories into conceptual (Sydney is a coastal city in Australia), relational (Ian Fleming had a 1908 birth), thematic (James Bond has theme James Bond) and administrative (the Sydney article is available as a spoken article). Conceptual categories, which are most useful for categorisation, often have plural head nouns (e.g. cities of Coastal cities in Australia) which describe the nature of member articles.

We use a bootstrapping approach to classification (see Figure 2), with heuristics based primarily on category head nouns and definitional opening sentences of articles. This approach reflects our intuitions that: (a) it is difficult to manually design rules with high coverage; (b) bag-of-words methods lose structural and linguistic information; (c) a semi-supervised approach is able to learn heuristics from unlabelled data; (d) we may easily leave an article's class undecided and ignore it in future processing.

Each article is classified as one of: unknown (UNK; not a target category for evaluation, like O in IOB tagging); a member of a NE category; a disambiguation page (DAB; these list possible referent articles for a given title); or a non-entity (NON). We identify these final two categories largely on the basis of specialised static heuristics, while entity classes are assigned through mappings learnt in the bootstrapping process.

4.1 Non-entity heuristics

Because of the diversity of non-entity articles, and the ease of identifying large portions of them, we first attempt to classify each article as a non-entity. We generally assume that articles whose incoming links are largely lowercase are non-entities, and also classify as NON all articles with a title beginning List of , together finding 32% of NON articles in our hand-labelled data. We also separately identify DAB articles on the basis of their title and categories.

4.2 Bootstrapped heuristics

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

Category nouns Using the C&C tools, we POS tagged and chunked (shallow phrasal parsing) all category titles in Wikipedia in order to determine their head nouns, the last word of the first noun phrase chunk. If the POS tagger identified this head as plural, we assume that the category is *conceptual* by Suchanek et al.'s (2007) designation. Thus institutions would be extracted as the head of category Educational institutions established in 1850 and might identify the category constituents as belonging to ORG. Each such phrasal head, or bigram collocation (differentiating radio stations from railway stations), is considered a feature which may be mapped to an entity class. The most frequent class present among an article's categories is then assigned to the article. If no conceptual categories with mappings are identified, or multiple classes tie maximum, UNK is tentatively assigned.

Definition nouns Where category nouns are inconclusive, or the maximum category class only

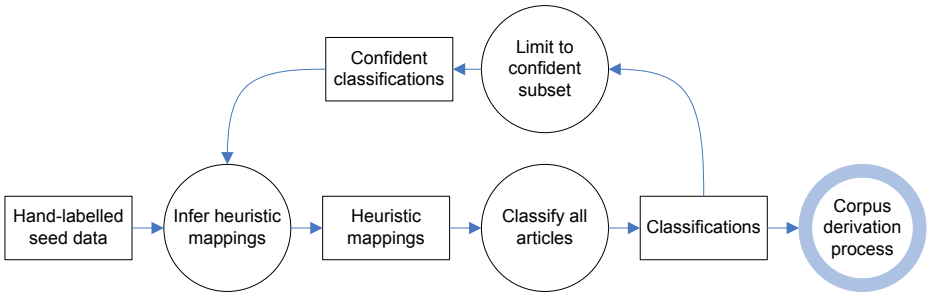

Figure 2: A bootstrapping approach to article classification

leads by 1, we resort to a secondary heuristic feature. Kazama and Torisawa (2007) make the assumption that many articles' opening sentences are in the form of definitions, and hence the noun phrase following a copula (is, are, was, were) is indicative of the article's category. Thus the article on Sydney opens Sydney is the most populous city in Australia..., wherein the word city best identifies the class of the article as LOC. We extract only one such definition noun (again, a unigram or bigram collocation) by POS tagging and chunking the first sentence of each article. Like with category nouns, this may be mapped to an entity class which relabels articles previously marked UNK, or in case of contradicting heuristics may revert a classification to UNK.

4.3 Bootstrapping

As shown in Figure 2, bootstrapping in our classification process involves using hand-labelled data to initialise mappings from category and definition nouns to entity classes, and having a feedback loop in which we use the confident results of one classification to produce heuristic mappings for the next. Since most articles fall within multiple categories, each subsequent bootstrap produces new mappings (until convergence), and therefore allows the confident classification of a larger portion of Wikipedia.

We infer mappings as follows: Given a set of articles and their classes, we can count the number of times each feature occurs with the class. For each candidate noun N (unigram or bigram), the class k with which it is most often associated is determined. If n classified articles support the mapping $N \rightarrow k$ and m articles contradict it, then we accept the mapping if $n \geq t$ and $\frac{m}{n+m} < p$, for some constant thresholds t and p. We have used $p = .25$, with values for t given in Table 1.

t	Category NNs	Definition NNs
Seed inference	1	2
Feedback	2	4

Table 1: Varying threshold values for inference.

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

A single annotator manually labelled 1300 Wikipedia articles with over 60 fine-grained category labels based on Brunstein (2002), which were reduced for evaluation to {LOC,PER,ORG,MISC,NON,DAB}, apart from 7 ambiguous entities which were left unlabelled. Initially 1100 random articles were labelled, but we found that the data poorly represented popular entity types with few instances, such as countries. Hence an additional 200 articles were randomly selected from among the set of articles with over 700 incoming links. The distribution of classes in the data is shown in Table 2. Of this set, we used 15% in a held-out test set for development, and performed final evaluation with ten-fold cross-validation.

4.4 Classification results

We found that evaluation statistics stabilised for our held-out test set after three bootstrap loops. Although even initial mappings (1050 category nouns; 132 definition nouns) produced a micro-averaged F-score of 84%, this reflects the high proportion of easily-identifiable PER articles, as the macro-average was much lower (63%). After the third bootstrap, with 8890 category nouns and 26976 definition nouns, this had increased to 91% micro- and 90% macro-average F-score. 12% of

Class	%	P	R	F
LOC	19	95	94	95
PER	24	96	98	97
ORG	14	92	80	85
MISC	18	93	72	80
DAB	5	100	93	96
NON	20	89	69	78
All		94	84	89
Entities only		96	88	92

Table 2: The class distribution within our manual article labels and average results of a ten-fold cross-validation. Overall results are micro-averaged.

the test data was labelled UNK.

Per-class and overall results of cross-validation are shown in Table 2. Our largest failures are in recall for MISC and NON, by far the broadest classes and hence difficult to capture completely.

5 Extracting and selecting sentences

Wikipedia's articles are composed using a structural markup language specific to its software. While marked-up data is available, it requires cleaning, separation into sentences and tokenisation in order to be transformed into a NER training corpus. We produce a parse tree of the markup using `mwlib`[2], remove most non-sentential data and all markup other than inter-article links, and split article texts into sentences using Punkt (Kiss and Strunk, 2006)—an unsupervised algorithm for sentence boundary detection, trained here on Wikipedia data—before tokenising.

We need to select sentences for inclusion in our training corpus for which we are confident of having correctly labelled all named entities. For the generic NER task in English, this depends highly on capitalisation information. For instance, we simply might accept only sentences where all capitalised words have links to articles of known classification (not UNK or DAB). This criterion is overly restrictive: (a) it provides a low recall of sentences per article; (b) it is biased towards short sentences; and (c) since each entity name is often linked only on its first appearance in an article, it is more likely to include fully-qualified names than shorter referential forms (surnames, acronyms, etc.) found later in the article. We are also challenged by many words that are capitalised by English convention but do not correspond to entities. These and related problems are tackled in the following sub-sections.

5.1 Inferring additional links

In order to increase our coverage of Wikipedia sentences, we attempt to infer additional links. In particular, since Wikipedia style dictates that only the first mention of an entity should be linked in each article, we try to identify other mentions of that entity in the same article. We begin by compiling a list of alternative titles for each article. Then for any article in which we are attempting to infer links we produce a trie containing the alternative titles of all outgoing links. When a word with an uppercase letter is found, we find the longest matching string within the trie and assign its class to the matching text.

Alternative titles for an article A include:

Type 1 The title of A and those of redirects[3] to A (with expressions following a comma or within parentheses removed);

Type 2 The first or last word of A's title if A is of class PER;

Type 3 The text of all links whose target is A.

We have switched use of each type of inferred titles on and off in our experiments below.

5.2 Conventional capitalisation

As well as proper names, first words of sentences, pronouns (I in English), dates, adjectival forms of names (e.g. nationalities), personal titles and acronyms are capitalised in English. As an exception to our general sentence selection criterion, we include such capitalised words in our corpus.

First words If a word beginning a sentence or following some punctuation (semicolon, left-quote, etc.) is capitalised and unlinked, it may be difficult to determine whether it should be labelled as belonging to an entity. Unless an entity link can be inferred, a first word is ignored if it is found on a list of 1520 words (compiled from Wikipedia data) including collocational frequent sentence starters (Kiss and Strunk, 2006),

[2] A Python-based parser for MediaWiki markup. `http://code.pediapress.com`

[3] Redirect pages make articles accessible through alternative titles.

and words which are commonly both sentence-initial and lowercase when sentence-internal.

Dates Names of months and days of the week are identified by regular expressions.

Personal titles Personal titles (e.g. Brig. Gen., Prime Minister-elect) are conventionally capitalised in English. In some cases, such titles are linked to relevant articles, but e.g. U.S. President is in categories like Presidents of the United States, causing its incorrect classification as PER. We have implemented a trivial solution to catch some titles: if a link appears immediately before a link to a PER target, we assume that it is a title and may be included in the corpus without a NE tag. Note that this fails to handle the same titles when linked freely in text. We use the BBN corpus (Weischedel and Brunstein, 2005) to compile a list of titles that are rarely linked (e.g. Mr.).

Adjectival forms Adjectival forms of entity names, such as American or Islamic, are capitalised in English. While these are not technically entities, both the CoNLL and BBN gold-standard corpora (see section 6.1) tag them. Our rudimentary solution for this involves POS tagging the potential corpus text and relabelling entities as MISC if their final word is tagged as an adjective. This does not cover nationalities used as nouns, for which we currently retain the incorrect label.

5.3 Anomalous capitalisation

Capitalised non-entity links and all-lowercase entity links may be problematic. The former often results from mis-classification, or includes an NE in its title, e.g. Greek alphabet or Jim Crow laws, in which case it would be incorrect to leave the reference untagged. Lowercase entity links result from non-proper noun references to entities, e.g. in In the Ukraine, anarchists fought in the civil war ..., the text civil war links to Russian Civil War. Sentences with capitalised NON links or lowercase entity links are therefore discarded, excepting entities like gzip where Wikipedia marks the article as having a lowercase title.

5.4 Adjusting link boundaries

Link text sometimes incorporates more than just the entity name, such the possessive 's at the end of a name, or the linking of Sydney, Australia which should be treated as two separate entities. Hence we unlink the following strings when found at the end of link text: parenthesised expressions; text following a comma for LOC, ORG and PER; possessive 's; or other punctuation.

6 Evaluation

We evaluate our corpora by training the C&C tagger[4] to build separate models (a) when trained with Wikipedia data; (b) when trained with hand-annotated training data; (c) when trained with both combined, and comparing the tagging results on gold-standard test data. We use the C&C Maximum Entropy NER tagger with default orthographic, contextual, in-document and first name gazetteer features (Curran and Clark, 2003). Our results are given as per-category and micro-averaged phrasal precision, recall and F_1-score.

6.1 Gold-standard corpora

We evaluate our generated corpora against three sets of manually-annotated data from (a) the MUC-7 Named Entity Task (MUC, 2001); (b) the English CoNLL-03 Shared Task (Tjong Kim Sang and De Meulder, 2003); (c) the BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005). Stylistic and genre differences between the source texts affect compatibility for NER, e.g. the CoNLL corpus formats headlines in all-caps, and includes much non-sentential data, such as tables of sports scores.

Each corpus uses a different set of entity labels, with MUC marking locations, organisations and personal names in addition to numerical and time information. CoNLL labels only proper names but adds a MISC category for all entities not otherwise tagged. BBN ambitiously annotate the entire Penn Treebank corpus with 105 fine-grained tags: 54 corresponding to CoNLL entities; 21 for numerical and time data; and 30 for other classes of terms. For the present evaluation, BBN's tags were reduced to the equivalent CoNLL tags, with non-CoNLL tags in the BBN and MUC data removed. Since no MISC entities are marked in MUC, such labels need to be removed from CoNLL, BBN and Wikipedia data for comparison.

[4] http://svn.ask.it.usyd.edu.au/trac/candc

Corpus	# tags	Number of tokens		
		TRAIN	DEV	TEST
MUC-7	3	84051	18764	60872
CONLL-03	4	203621	51362	46435
BBN	54	901894	142218	129654

Table 3: Corpora used for evaluation

All corpora were transformed into a common format and tagged with parts of speech using the Penn Treebank-trained (sections 2-21) C&C POS tagger. While standard training (TRAIN), development (DEV) and final test (TEST) set divisions were available for the CONLL and MUC data, the BBN corpus was split at our discretion: sections 03–21 for TRAIN, 00–02 for DEV and 22-24 for TEST. The corpus sizes are compared in Table 3.

6.2 Wikipedia data and experiments

Wikipedia's article text is made freely available for download.[5] We have used data from the 22 May 2008 dump of English Wikipedia which includes 2.3 million articles. Splitting this into sentences and tokenising produced 32 million sentences each containing an average of 24 tokens.

Our experiments were mostly performed with Wikipedia-derived corpora of 150,000 sentences. Despite having 14 million sentences available, we were limited by time and memory for training.

We report results from four groups of experiments: (a) how does a Wikipedia-trained tagger compare to gold-standard data? (b) what is the effect of training a tagger with both gold-standard and Wikipedia-derived data? (c) how does the number of sentences in the Wikipedia corpus affect performance? (d) to what extent does the inference of additional links (see section 5.1) affect results? Levels of link inference are differentiated between corpora WP0 (no inference), WP1 (type 1 alternative titles), WP2 (types 1–2) and WP3 (types 1–3). The 150k-sentence corpora contain 3.5 million tokens on average.

7 Results and discussion

As shown in Tables 4 and 5, each set of gold-standard training data performs much better on corresponding evaluation sets (italicised) than on test sets from other sources. The exception is for BBN on MUC TEST, due to differing TEST and DEV subject matter. The 12-33% mismatch between training and evaluation data suggests that the training corpus is an important performance factor (see also Ciaramita and Altun (2005)).

[5] http://download.wikimedia.org/

Training corpus	DEV overall F-score		
	MUC	CONLL	BBN
MUC	*83.4*	54.8	59.7
CONLL	64.5	*86.9*	60.2
BBN	75.0	58.0	*88.0*
WP0 – no inference	63.4	63.6	56.6
WP1	65.3	65.4	58.6
WP2	**68.1**	67.0	**60.6**
WP3 – all inference	64.4	**68.5**	58.0

Table 4: DEV results without MISC.

TRAIN	With MISC		No MISC		
	CONLL	BBN	MUC	CONLL	BBN
MUC	—	—	*74.4*	51.7	54.8
CONLL	*81.2*	62.3	58.8	*82.1*	62.4
BBN	54.7	*86.7*	75.7	53.9	*88.4*
WP2	58.9	62.3	67.5	60.4	58.8

Table 5: TEST results for WP2.

A key result of our work is that the performance of non-corresponding hand-annotated corpora is often exceeded by Wikipedia-trained models.

We also assess using our Wikipedia corpora together with gold-standard data on traditional train-test pairs. Table 6 shows that this approach leads to only marginal variations in performance.

Table 4 also illustrates the effectiveness of link inference, which is able to increase F-score by 5%. Performance increases as inference types 1 (identifying article titles) and 2 (single words of PER titles) are added, but 3 (all incoming link labels) degrades performance on the MUC and BBN corpora, since it likely over-generates alternative titles. Matching the longest string may also falsely include additional words, e.g. tagging Australian citizen rather than Australian, to which inference level 3 is most susceptible.

In Figure 3, we illustrate the effect of varying the size of the Wikipedia-derived training data. Increased training data tends to improve performance to a point (around 25k sentences for MUC and 125k for BBN) after which improvements are marginal and results may degrade. The late stabilising of BBN performance is possibly caused by the size and breadth of its evaluation data sets.

To analyse overall error, our per-class results

Training corpus	TEST overall F-score		
	MUC	CONLL	BBN
Corresponding TRAIN	74.4	82.1	88.4
TRAIN + WP2	76.8	81.8	87.7

Table 6: Wikipedia as additional training data

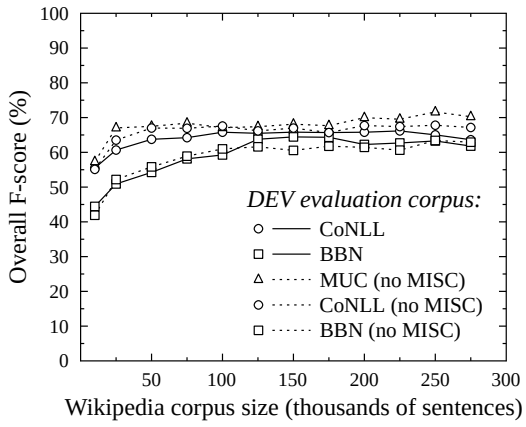

Figure 3: The effect of varying WP2 corpus size.

are shown in Table 7. LOC and PER entities are relatively easy to identify, although a low precision for PER suggests that many other entities have been marked erroneously as people, unlike the high precision and low recall of ORG. As an ill-defined category, with uncertain mapping between BBN and CONLL classes, MISC precision is unsurprisingly low. We also show results evaluating the correct labelling of each token, and the much higher results (13%) reflects a failure to correctly identify entity boundaries. This is common in NER, but a BBN-trained model only gives 5% difference between phrasal and token F-score. We believe this reflects Wikipedia links often including other tokens along with proper names.

Among common tagging errors we have identified, we find: tags continuing over additional words as in New York-based Loews Corp. all being marked as a single ORG; nationalities marked as LOC rather than MISC; White House a LOC rather than ORG, as with many sports teams; single-word ORG entities marked as PER; titles such as Dr. included in PER tags; untagged title-case terms and tagged lowercase terms in the gold-standard.

Our results suggest many avenues for improving corpus derivation, but highlight Wikipedia as a source of competitive training data.

Class	By phrase			By token		
	P	R	F	P	R	F
LOC	62.0	76.8	68.7	61.0	82.4	70.1
MISC	43.5	55.7	48.8	42.5	59.3	49.5
ORG	76.8	53.4	63.0	87.9	65.3	74.9
PER	48.0	81.0	60.3	56.3	95.5	70.8
All	60.9	63.8	62.3	76.7	72.5	74.6

Table 7: Results for each entity category when the WP2 model was evaluated on the BBN TEST set.

8 Conclusion and future work

There is much room for improving the results of our Wikipedia-based NE annotations. A more careful approach to link inference may reduce incorrect boundaries of tagged entities. Adding disambiguation pages as another source of alternative article titles will make the labelling of acronyms more common. Better labelling of personal titles and adjectival entity names may also provide great gain, as did our simple approaches.

Since the training corpus is not often a variable in NER research, we need to explore ways to fairly evaluate corpus-based experiments: the number of articles, sentences, tagged entities or tokens may all be chosen as variables or invariants in experiments; and differing genres or annotation schemes make results difficult to compare.

We have nonetheless shown that Wikipedia can be used a source of free annotated data for training NER systems. Although such corpora need to be engineered specifically to a desired application, Wikipedia's breadth may permit the production of large corpora even within specific domains. Focusing on this flexibility, we intend to experiment with finer-grained NE hierarchies, domain-specific annotation schema, and multilingual NER. Our results indicate that Wikipedia data can perform better (up to 8.7% for MUC on CONLL) than training data that is not matched to the evaluation, and hence is widely applicable. Transforming Wikipedia into training data thus provides a free and high-yield alternative to the laborious manual annotation required for NER.

Acknowledgments

We would like to thank members of the Language Technology Research Group and the anonymous reviewers for their helpful feedback. Curran and Murphy were funded under ARC Discovery grant

DP0665973 and Nothman was supported by a University of Sydney Honours Scholarship.

References

Joohui An, Seungwoo Lee, and Gary Geunbae Lee. 2003. Automatic acquisition of named entity tagged corpus from world wide web. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 165–168.

Ada Brunstein. 2002. Annotation guidelines for answer types. LDC2005T33.

Nancy Chinchor. 1998. Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference*.

Massimiliano Ciaramita and Yasemin Altun. 2005. Named-entity recognition in novel domains with external lexical knowledge. In *NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*.

James R. Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 164–167.

Wisam Dakka and Silviu Cucerzan. 2008. Augmenting Wikipedia with named entity tags. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 545–552, Hyderabad, India.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

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.

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

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

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–8, Bergen, Norway.

George A. Miller. 1998. Nouns in WordNet. In *WordNet: An Electronic Lexical Database*, chapter 1. MIT Press.

2001. *Message Understanding Conference (MUC) 7*. Linguistic Data Consortium, Philadelphia.

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

David Nadeau, Peter D. Turney, and Stan Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Proceedings of the 19th Canadian Conference on Artificial Intelligence*, volume 4013 of *LNCS*, pages 266–277.

Alexander E. Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1–9, Columbus, Ohio.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1818–1824.

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.

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, Rafael Muñoz, and Monica Monachini. 2008. Named entity WordNet. In *Proceedings of the 6th International Language Resources and Evaluation Conference*.

Ralph Weischedel and Ada Brunstein. 2005. *BBN Pronoun Coreference and Entity Type Corpus*. Linguistic Data Consortium, Philadelphia.