# Named Entity Recognition in Wikipedia

**Dominic Balasuriya Nicky Ringland Joel Nothman Tara Murphy James R. Curran**
School of Information Technologies
University of Sydney
NSW 2006, Australia
{dbal7610,nicky,joel,tm,james}@it.usyd.edu.au

## Abstract

Named entity recognition (NER) is used in many domains beyond the newswire text that comprises current gold-standard corpora. Recent work has used Wikipedia's link structure to automatically generate near gold-standard annotations. Until now, these resources have only been evaluated on newswire corpora or themselves.

We present the first NER evaluation on a Wikipedia gold standard (WG) corpus. Our analysis of cross-corpus performance on WG shows that Wikipedia text may be a harder NER domain than newswire. We find that an automatic annotation of Wikipedia has high agreement with WG and, when used as training data, outperforms newswire models by up to 7.7%.

## 1 Introduction

Named Entity Recognition (NER) is the task of identifying and classifying people, organisations and other named entities (NE) within text. NER is central to many NLP systems, especially information extraction and question answering.

Machine learning approaches now dominate NER, learning patterns associated with individual entity classes from annotated training data. This training data, including English newswire from the MUC-6, MUC-7 (Chinchor, 1998), and CoNLL-03 (Tjong Kim Sang and De Meulder, 2003) competitive evaluation tasks, and the BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005), is critical to the success of these approaches.

This data dependence has impeded the adaptation or *porting* of existing NER systems to new domains, such as scientific or biomedical text, e.g. Nobata et al. (2000). Similar domain sensitivity is exhibited by most tasks across NLP, e.g. parsing (Gildea, 2001), and the adaptation penalty is still apparent even when the same set of named entity classes is used in text from similar domains (Ciaramita and Altun, 2005).

Wikipedia is an important corpus for information extraction, e.g. Bunescu and Paşca (2006) and Wu et al. (2008) because of its size, currency, rich semi-structured content, and its closer resemblance to web text than newswire. Recently, Wikipedia's markup has been exploited to automatically derive NE annotated text for training statistical models (Richman and Schone, 2008; Mika et al., 2008; Nothman et al., 2008).

However, without a gold standard, existing evaluations of these models were forced to compare against mismatched newswire corpora or the noisy Wikipedia-derived annotations themselves. Further, it was not possible to directly ascertain the accuracy of these automatic extraction methods.

We have manually annotated 39,007 tokens of Wikipedia with coarse-grained named entity tags (WG). We present the first evaluation of Wikipedia-trained models on Wikipedia: the C&C NER tagger (Curran and Clark, 2003b) trained on (a) automatically annotated Wikipedia text (WP2) extracted by Nothman et al. (2009); and (b) traditional newswire NER corpora (MUC, CoNLL and BBN). The WP2 model, though trained on noisy annotations, outperforms newswire models on WG by 7.7%. However, every model, including WP2, performs far worse on WG than on the newswire.

We examined the quality of WG, and found that our annotation strategy produced a high-quality, consistent corpus. Our analysis suggests that it is the form and distribution of NEs in Wikipedia that make it a difficult target domain.

Finally, we compared WG with the annotations extracted by Nothman et al. (2009), and found agreement comparable to our inter-annotator agreement, demonstrating that NE corpora can be derived very accurately from Wikipedia.



*Proceedings of the 2009 Workshop on the People's Web Meets NLP, ACL-IJCNLP 2009*, pages 10–18,
Suntec, Singapore, 7 August 2009. 

## 2 Background

Traditional evaluations of NER have considered the performance of a tagger on test data from the same source as its training data. Although the majority of annotated corpora available consist of newswire text, recent practical applications cover a far wider range of genres, including Wikipedia, blogs, RSS feeds, and other data sources. Ciaramita and Altun (2005) showed that even when moving a short distance, e.g. annotating WSJ text with the same scheme as CoNLL's Reuters, the performance was 26% worse than on the original text.

Similar differences are reported by Nothman et al. (2009) who compared MUC, CoNLL and BBN annotations reduced to a common tag-set. They found poor cross-corpus performance to be due to tokenisation and annotation scheme mismatch, missing frequent lexical items, and naming conventions. They then compared automatically-annotated Wikipedia text as training data and found it also differs in otherwise inconsequential ways from the newswire corpora, in particular lacking abbreviations necessary to tag news text.

### 2.1 Automatic Wikipedia annotation

Wikipedia, a collaboratively-written online encyclopedia, is readily exploited in NLP, because it is large, semi-structured and multilingual. Its articles often correspond to NEs, so it has been used for NE recognition (Kazama and Torisawa, 2007) and disambiguation (Bunescu and Paşca, 2006; Cucerzan, 2007). Wikipedia links often span NEs, which may be exploited to automatically create annotated NER training data by determining the entity class of the linked article and then labelling the link text with it.

Richman and Schone (2008) use article classification knowledge from English Wikipedia to produce NE-annotated corpora in other languages (evaluated against NE gold standards for French, Spanish, and Ukrainian). Mika et al. (2008) explored the use of tags from a CoNLL-trained tagger to seed the labelling of entities and evaluate the performance of a Wikipedia-trained model by hand.

We make use of an approach described by Nothman et al. (2009) which is engineered to perform well on BBN data with a reduced tag-set (LOC, MISC, ORG, PER). They derive an annotated corpus with the following steps:

1. Classify Wikipedia articles into entity classes
2. Split the articles into tokenised sentences
3. Label expanded links according to target NEs
4. Select sentences for inclusion in a corpus

To prepare the text, they use `mwlib` (PediaPress, 2007) to parse Wikipedia's native markup retaining only paragraph text with links, apply Punkt (Kiss and Strunk, 2006) estimated on Wikipedia text to perform sentence boundary detection, and tokenise the resulting text using regular expressions.

Nothman et al. (2009) infer additional NEs not provided by existing links, and apply rules to adjust link boundaries and classifications to closer match BBN annotations.

### 2.2 NER evaluation

Meaningful automatic evaluation of NER is difficult and a number of metrics have been proposed (Nadeau and Sekine, 2007). Ambiguity leads to entities correctly delimited but misclassified, or boundaries mismatched despite correct classification.

Although the MUC-7 evaluation (Chinchor, 1998) defined a metric which was less sensitive to often-meaningless boundary errors, we consider only exact entity matches as correct, following the standard CoNLL evaluation (Tjong Kim Sang, 2002). We report precision, recall and $F$-score for each entity type.

## 3 Creating the Wikipedia gold standard

We created a corpus by manually annotating the text of 149 articles from the May 22, 2008 dump of English Wikipedia. The articles were selected at random from all articles describing named entities, with a roughly equal proportion of article topics from each of the four CoNLL-03 classes (LOC, MISC, ORG, PER). We adopted Nothman et al.'s (2008) preprocessing described above to produce tokenised sentences for annotation.

Only body text was extracted from the chosen articles for inclusion in the corpus. Four articles were found not to have any usable text, consisting solely of tables, lists, templates and section headings, which we remove. Their exclusion leaves a corpus of 145 articles.

### 3.1 Annotation

Annotation was initially carried out using a fine-grained tag-set which was expanded by the an-

[COMPANY Aero Gare] was a kitplane manufacturer founded by [PERSON Gary LeGare] in [CITY Mojave] , [STATE California] to marketed the [PLANE Sea Hawker] amphibious aircraft .

(a) Fine-grained annotation

[ORG Aero Gare] was a kitplane manufacturer founded by [PER Gary LeGare] in [LOC Mojave] , [LOC California] to marketed the [MISC Sea Hawker] amphibious aircraft .

(b) Coarse-grained annotation

Figure 1: An example of coarse and fine-grained annotation of Wikipedia text.

notators as annotation progressed, and eventually contained 96 tags.

We created a mapping from these fine-grained tags to the four coarse-grained tags used in the CoNLL-03 data: PER, LOC, MISC and ORG. This enables evaluation with existing NER models. We believe this two-phase approach allowed annotators to defer difficult mapping decisions, (e.g. should an airport be classified as a LOC, ORG, or MISC?) which can then be made after discussion. The mapping could also be modified to suit a particular evaluation task.

Figure 1 shows an example of the use of fine and coarse-grained tags to annotate a sentence. Tags such as PERSON correspond directly to coarse-grained tags, while most map to a more general tag, such as STATE and CITY mapping to LOC. PLANE is an example of a fine-grained tag that cannot be mapped to LOC, ORG, or PER. These tags may be mapped to MISC; some are not considered entities under the CoNLL scheme and are left unlabelled in the coarse-grained annotation.

Three independent annotators were involved in the annotation process. Annotator 1 annotated all 145 articles using the fine-grained tags. Annotators 2 and 3 then re-annotated 19 of these articles (316 sentences or 8030 tokens), amounting to 21% of the corpus. Annotator 2 used the fine-grained tags described above, while Annotator 3 used the four coarse-grained CoNLL tags. To measure variation, all three annotations of this common portion were mapped down to the CoNLL tag-set and inter-annotator agreement was calculated.

We found that 202 tokens were disagreed upon by at least one annotator (2.5% of all tokens annotated), and these discrepancies were then discussed by the three annotators. The inter-annotator agreement will be analysed in more detail in Section 5.

Sentences containing grammatical and typographical errors were not corrected, so that the corpus would be as close as possible to the source text. Web text often contains errors, such as to marketed the Sea Hawker from the example in Figure 1, so any NER system must deal with these errors. Sentences with poor tokenisation or sentence boundary detection were identified and corrected manually, since these errors are introduced by our processing and annotation, and do not exist in the source text.

| Train | Test | $P$ | $R$ | $F$ |
|---|---|---|---|---|
| WP2 | WG | 66.5 | 67.4 | 66.9 |
| BBN | WG | 59.2 | 59.1 | 59.2 |
| CoNLL | WG | 54.3 | 57.2 | 55.7 |
| WP2 * | WG * | 75.1 | 67.7 | 71.2 |
| BBN * | WG * | 57.2 | 64.1 | 60.4 |
| CoNLL * | WG * | 53.1 | 62.7 | 57.5 |
| MUC * | WG * | 52.3 | 57.2 | 54.6 |
| WP2 | BBN | 73.4 | 74.6 | 74.0 |
| WP2 | CoNLL | 73.6 | 64.9 | 69.0 |
| WP2 * | MUC * | 86.2 | 68.9 | 76.6 |
| BBN | BBN | 85.7 | 87.3 | 86.5 |
| CoNLL | CoNLL | 85.3 | 86.5 | 85.9 |
| MUC | MUC | 81.0 | 83.6 | 82.3 |

Table 2: Tagger performance on various corpora. Asterisks indicate that MISC tags are ignored.

The final corpus was created by correcting annotation mistakes, with annotators 2 and 3 each correcting 50% of the corpus. The fine-grained tags were mapped to the four CoNLL tags before the final corrections were made. The final WG corpus consists of the body text of 145 Wikipedia articles tagged with the four CoNLL-03 tags.

## 4 NER on the Wikipedia gold-standard

Nothman et al. (2009) have previously shown that that an NER system trained on automatically annotated Wikipedia corpora performs reasonably well on non-Wikipedia text. Having created our WG corpus of gold-standard annotations, we are able to evaluate the performance of these models on Wikipedia text.

We compare the C&C NE maximum-entropy tagger (Curran and Clark, 2003b) trained on gold-standard newswire corpora (MUC-7, BBN and CoNLL-03) with the same tagger trained on automatically annotated Wikipedia text, WP2. WG is

| | WG | WP2 | BBN | | CoNLL-03 | | MUC-7 | |
|---|---|---|---|---|---|---|---|---|
| | Test | Train | Train | Test | Train | Test | Train | Test |
| Tokens | 39 007 | 3 500 032 | 901 849 | 129 654 | 203 621 | 46 435 | 83 601 | 60 436 |
| Sentences | 1 696 | 146 543 | 37 843 | 5 462 | 14 987 | 3 453 | 3 485 | 2 419 |
| Articles | 145 | — | 1 775 | 238 | 946 | 231 | 102 | 99 |
| NEs | 3 558 | 288 545 | 49 999 | 7 307 | 23 498 | 5 648 | 4 315 | 3 540 |

Table 1: Corpus sizes.

too small to train a reasonable NER model on gold-standard Wikipedia annotations. Part-of-speech tags are added to all corpora using the C&C POS tagger (Curran and Clark, 2003a) before training and testing. [1] We evaluate each model on traditional newswire evaluation corpora as well as WG. Table 1 gives the size of each corpus.

The results are shown in Table 2. The WP2 tagger performed substantially better on WG than taggers trained on newswire text, with a $7-11\%$ increase in $F$-score compared to BBN and CoNLL-03, and a 16% increase compared to MUC-7, when miscellaneous NEs in the corpus are not considered in the evaluation. The Wikipedia trained model thus outperforms newswire models on our new WG corpus even though the training annotations were automatically extracted.

The WP2 tagger performed worse on WG than on gold-standard news corpora (BBN and CoNLL), with a $2-7\%$ reduction in $F$-score. Further, the performance of WP2 on WG is $11-20\%$ $F$-score lower than same-source evaluation results, e.g. BBN on BBN, CoNLL on CoNLL. Therefore, despite WP2 showing an advantage in tagging WG due to their common source domain, we find that WG's annotations are harder to predict than the newswire test data commonly used for evaluation.

One possible explanation is that our WG corpus has been inconsistently annotated. When NEs of miscellaneous type are not considered in the evaluation (asterisks in Table 2), the performance of all taggers on WG improves, with WP2 demonstrating a 4% increase. This result suggests another partial explanation: that MISC NEs in Wikipedia are more difficult to annotate correctly, due to their poor definition and broad coverage. A third explanation is that the automatic conversion process proposed by Nothman et al. (2008) produces much lower quality training data than manual annotation. We explore these three possibilities below.

[1] Both taggers are available from `http://svn.ask.it.usyd.edu.au/trac/candc`.

| | Token | Exact | NE only |
|---|---|---|---|
| A1 and A2 | 0.95 | 0.99 | 0.88 |
| A1 and A3 | 0.91 | 0.95 | 0.81 |
| A2 and A3 | 0.91 | 0.96 | 0.79 |
| Fleiss' Kappa | 0.92 | 0.97 | 0.83 |

Table 3: Initial human inter-annotator agreement.

## 5 Quality of the Wikipedia gold standard

The low performance observed on WG may be due to the poor quality of its annotation. We ensure that this is not the case by measuring inter-annotator agreement. The WG annotation process produced three independent annotations of a subset of WG. These annotations were compared using Cohen's $\kappa$ (Fleiss and Cohen, 1973) between pairs of annotators, and Fleiss' $\kappa$ (Fleiss, 1971), which generalises Cohen's $\kappa$ to more than two concurrent annotations.

Table 3 shows the three types of $\kappa$ values calculated. *Token* is calculated on a per token basis, comparing the agreement of annotators on each token in the corpus; *NE only*, is calculated on the agreement between entities alone, excluding agreement in cases where all annotators agreed that a token was not a NE; *Exact* refers to the agreement between annotators where all annotators have agreed on the boundaries of a NE, but disagree on the type of NE.

Annotator 1 originally annotated the entire corpus, and Annotators 2 and 3 then corrected exactly half of the corpus each after a discussion between the three annotators to resolve ambiguities. Landis and Koch (1977) determine that a $\kappa$ value greater than 0.81 indicates almost perfect agreement. By this standard, our three annotators were in strong agreement prior to discussion, with our Fleiss' $\kappa$ values all greater than 0.81. Inconsistencies in the corpus due to annotation mistakes by Annotator 1 were corrected by Annotators 2 and 3.

Inter-annotator agreement for cases where the annotators agreed on NE boundaries was higher than agreement on each token, which suggests that many discrepancies resulted from NE bound-

| | LOC | MISC | ORG | PER | $H(C)$: With O | Without O | Total NEs | % NE tokens |
|---|---|---|---|---|---|---|---|---|
| WG | 28.5 | 20.0 | 25.2 | 26.3 | 0.98 | 2.0 | 3 558 | 17.1 |
| BBN | 22.4 | 9.8 | 46.4 | 21.3 | 0.61 | 1.7 | 49 999 | 9.6 |
| MUC | 33.3 | — | 40.7 | 26.1 | 0.52 | 1.5 | 4 315 | 8.1 |
| CoNLL | 30.4 | 14.6 | 26.9 | 28.1 | 0.98 | 1.9 | 23 498 | 17.1 |

Table 4: NE class distribution, tag entropy and NE density statistics for gold-standard corpora and WG.

ary ambiguities, or disagreement as to whether a phrase constituted a NE at all. Higher inter-annotator agreement between Annotators 1 and 2 leads us to believe that the two-phase annotation strategy, where an initially fine-grained tag-set is reduced, results in more consistent annotation.

Our analysis demonstrates that WG is annotated in a consistent and accurate manner and the small number of errors cannot alone explain the reduced performance figures.

## 6 Comparing gold-standard corpora

### 6.1 NE class distribution

Table 4 compares the distribution of different classes of NEs across different corpora on the four CoNLL categories. WG has a higher proportion of PER and MISC NEs and a lower proportion of ORG NEs than the BBN corpus. This is also found in the MUC corpus, although comparisons to MUC are affected by its lack of a MISC category. The CoNLL-03 corpus is most similar to WG in terms of the distribution of the NE classes, although CoNLL-03 has a smaller proportion of MISC NEs than WG. An analysis of the lengths of NEs in CoNLL shows, however, that they are very different to those in WG (see Table 8), perhaps explaining the difference in performance observed.

Tag entropy $H(C)$ was calculated for each corpus with respect to the 5 possible classes (4 NE classes, and the O tag, indicating non-entities). $H(C)$ is a measure of the amount of information required to represent the classification of each token in the corpus. Two calculations are made, including and excluding the frequent O tag. Our results (Table 4) suggest that WG's tags are least predictable, with a tag entropy of 2.0 bits (without the O class) compared to 1.7 and 1.9 bits for BBN and CoNLL respectively.

### 6.2 Fine-grained class distribution

While the CoNLL-03 and MUC evaluation corpora are marked up with only very coarse tags, the BBN corpus uses 29 coarse tags, many with specific subtypes, including NEs, descriptors of NEs and non-NEs, intended as answer types for question answering (Brunstein, 2002). Non-NE types include MONEY and TIME, which are also tagged in the MUC corpus, and others such as ANIMAL. When evaluating the performance of the taggers, each of BBN's 150 fine-grained tags was mapped to one of four coarse-grained classes or none, using a mapping described in Nothman (2008).

| Mapped BBN tag | WG | BBN |
|---|---|---|
| PERSON | 25.9 | 19.3 |
| ORGANIZATION:OTHER | **13.0** | 2.8 |
| ORGANIZATION:CORPORATION | 9.2 | **43.1** |
| GPE:CITY | 8.0 | 6.7 |
| WORK_OF_ART:SONG | **4.7** | 0.1 |
| NORP | 4.3 | 3.1 |
| WORK_OF_ART:OTHER | **4.1** | 1.3 |
| GPE:COUNTRY | 3.5 | 5.1 |
| ORGANIZATION:EDUCATIONAL | **3.0** | 0.9 |
| GPE:STATE_PROVINCE | 2.8 | 2.8 |
| ORGANIZATION:POLITICAL | **2.6** | 0.6 |
| EVENT:OTHER | **2.5** | 0.4 |
| ORGANIZATION:GOVERNMENT | 2.0 | **7.5** |
| WORK_OF_ART:BOOK | **1.6** | 0.4 |
| EVENT:WAR | **1.6** | 0.1 |
| FAC:OTHER | **1.4** | 0.2 |
| LOCATION:REGION | 1.3 | 0.8 |
| FAC:ATTRACTION | **1.2** | 0.0 |

Table 5: Distribution of some fine-grained tags

However, since the WG corpus was initially annotated using 96 distinct classes, we map these tags to the corresponding fine-grained BBN NE classes. In some cases, the tags map exactly (e.g. COUNTRY mapped to LOCATION:COUNTRY); in other cases, classes have to be merged or not mapped at all, where the BBN and WG annotations differ in granularity. Where possible, we map to fine-grained BBN categories.

We create mappings to a total of 36 BBN entity types, and apply them across the WG corpus. Table 5 shows the distribution of the most common tags, calculated as a percentage of all counts of the 36 selected tags across each corpus. Tags for which there is at least a two-fold difference in proportion between BBN and WG are marked in bold.

The comparison is dominated by the presence of a disproportionate number of ORG:CORPORATIONS in the BBN corpus com-

| | 1 | 2 | 3 | 4 | 5 | 6 | 7+ | # NEs |
|---|---|---|---|---|---|---|---|---|
| WG | 53.0 | 77.0 | 88.9 | 94.8 | 96.6 | 98.2 | 100 | 712 |
| BBN (train) | 75.0 | 91.0 | 95.4 | 97.2 | 98.2 | 98.7 | 100 | 4 913 |
| CoNLL (train) | 75.0 | 93.8 | 98.1 | 99.5 | 99.9 | 99.9 | 100 | 3 437 |

Table 6: Comparing MISC NE lengths (cumulative).

| Feature group | WG | BBN | CoNLL |
|---|---|---|---|
| Current token | 0.88 | 0.89 | 0.93 |
| Current POS | 0.43 | 0.57 | 0.48 |
| Current word-type | 0.42 | 0.49 | 0.48 |
| Previous token | 0.46 | 0.43 | 0.47 |
| Previous POS | 0.12 | 0.19 | 0.14 |
| Previous word-type | 0.07 | 0.14 | 0.12 |

Table 7: Feature-tag gain ratios.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7+ |
|---|---|---|---|---|---|---|---|
| WG | 49.9 | 81.7 | 93.1 | 97.4 | 98.6 | 99.4 | 100 |
| BBN (train) | 57.4 | 83.3 | 92.9 | 97.4 | 99.1 | 99.6 | 100 |
| CoNLL (train) | 63.1 | 94.5 | 98.4 | 99.4 | 99.8 | 99.9 | 100 |
| MUC (train) | 62.0 | 89.1 | 96.1 | 99.1 | 99.7 | 99.8 | 100 |

Table 8: Comparing all NE lengths (cumulative).

pared to WG. It also mentions many more governmental organisations. Prominent cases of tags found in higher proportions in WG are works of art, organisations of type OTHER (e.g. bands, sports teams, clubs), events and attractions.

This comparison demonstrates that there are observable differences in NE types between the news and Wikipedia domains. These differences are reflected in the distribution of both coarse and fine-grained types of NEs. The more complex entity distribution in Wikipedia is a likely cause for reduced NER performance on WG.

### 6.3 Feature-tag gain

Nobata et al. (2000) use *gain ratio* as an information-theoretic measure of corpus difficulty:

$$GR(C;F) = \frac{I(C;F)}{H(C)}$$

where $I(C;F) = H(C) - H(C|F)$ is the information gain of the NE tag distribution ($C$) with respect to a feature set $F$.

This gain ratio normalises the information gain over the tag entropy, which Nobata et al. (2000) suggest allows us to compare gain ratios between corpora. It also makes the impact of including the 'O' tag negligible for our calculations.

We apply this approach to measure the relative difficulty of tagging NEs in the WG corpus. Table 7 shows that WG tags seem generally harder to predict than those in newswire, on the basis of words, POS tags or orthographic word-types (like those used in the Curran and Clark (2003b) tagger as proposed by Collins (2002)).

In particular, POS tags are less indicative than in BBN and CoNLL, suggesting a wider variety of grammatical functions in NE names in Wikipedia – this might be expected with more band names, and song and movie titles. Alternatively, it may be an indication that the POS tagging is less reliable on Wikipedia using newswire-trained models.

The previous word's orthographic form also provides less information, which may relate to titles like Mr. and Mrs., strong indicators of PER entities, which are frequent in BBN and to a lesser extent CoNLL, but are almost absent in Wikipedia.

### 6.4 Lengths of named entities

The number of tokens in NEs is substantially different between WG and other gold-standard corpora. When compared with WG, other gold-standard corpora have a larger proportion of single-word NEs (between 7 and 13% more), as shown in Table 8. The distribution of NE lengths in BBN is most similar to WG, but it still differs significantly in the proportion of single-word NEs.

Additionally, WG has a larger number of long multi-word NEs than the other gold-standard corpora. Longer entities are more difficult to classify, since boundary resolution is more error prone and they typically contain lowercase words with a wider range of syntactic roles. This adds to the difficulty of correctly identifying NEs in WG.

The difference in entity lengths is most pronounced MISC NEs (Table 6), with Wikipedia having a substantially smaller number of single-word MISC NEs. The presence of a large number of long miscellaneous NEs, including song, film and book titles, and other works of art are a feature that characterises the nature of Wikipedia text in contrast to newswire text. Typically, longer MISC NEs in newswire text are laws and NORPs, which also appear in Wikipedia text.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7+ | # NEs |
|---|---|---|---|---|---|---|---|---|
| WG | 49.2 | 82.9 | 94.2 | 98.0 | 99.2 | 99.8 | 100 | 2 846 |
| BBN (train) | 55.4 | 82.4 | 92.6 | 97.4 | 99.2 | 99.7 | 100 | 45 086 |
| CoNLL (train) | 61.1 | 94.7 | 98.4 | 99.4 | 99.8 | 99.9 | 100 | 20 061 |
| MUC (train) | 62.0 | 89.1 | 96.1 | 99.1 | 99.7 | 99.8 | 100 | 4 315 |

Table 9: Comparing non-MISC NE lengths (cumulative).

| | # Sents | # with NEs | # NEs |
|---|---|---|---|
| WG | 1 696 | 1 341 | 3 558 |
| WG WP2-style | 571 | 298 | 569 |
| WG WP4-style | 698 | 425 | 831 |

Table 10: Size of WG and auto-annotated subsets.

## 7 Evaluation of automatic annotation

We compared the gold-standard annotations in our WG corpus to those sentences that were automatically annotated by Nothman et al. (2009). Their automatic annotation process does not retain all Wikipedia sentences. Rather, it selects sentences where, on the basis of capitalisation heuristics, it seems all named entities in the sentence have been tagged by the automatic process. We adopt this confidence criterion to produce automatically-annotated subsets of the WG corpus.

Two variants of their automatic annotation procedure were used: WP2 uses a few rules to infer tags for non-linked NEs in Wikipedia; WP4 has looser criteria for inferring additional links, and its over-generation typically reduced its performance as training data (Nothman et al., 2009).

A large proportion of sentences in our WG corpus cannot be automatically tagged with confidence. Sentence selection leaves 571 sentences (33.7%) after the WP2 process and 698 (41.2%) after the WP4 process (see Table 10). The use of the more permissive WP4 process may lead to the labelling of more NEs, but many may be spurious.

We use three approaches to compare automatic and manual annotations of WG text: (a) treat each corpus as test data and evaluate NER performance on each; (b) treat WP2 and WP4-style subsets as NER predictions on the WG corpus to calculate an $F$-score; and (c) treat the automatic annotations like human annotators and calculate $\kappa$ values.

We first evaluate the WP2 model on each corpus and find that performance is higher on automatically-annotated subsets of WG (Table 11). This is unsurprising given the common automatic annotation process and the effects of the selection criterion. However, Nothman (2008) provides an $F$-score for the WP2 model when evaluated on 10 folds of automatically-annotated (WP2-style) test data. This $F$-score is $8-10\%$ higher than WP2's performance on the WP2-style subset of WG, suggesting that WG's text is somewhat more difficult to annotate than typical portions of WP2-style text.

| TRAIN | TEST | $P$ | $R$ | $F$ |
|---|---|---|---|---|
| WP2 | WG manual | 66.5 | 67.4 | 66.9 |
| WP2 | WG WP2-style | 76.0 | 72.9 | 74.4 |
| WP2 | WG WP4-style | 75.5 | 71.4 | 73.4 |
| WP2 | WP2 ten folds | — | — | 83.6 |
| WP2 * | WG manual * | 75.1 | 67.7 | 71.2 |
| WP2 * | WG WP2-style * | 81.5 | 74.4 | 77.8 |
| WP2 * | WG WP4-style * | 81.9 | 74.6 | 78.1 |
| WP2 * | WP2 ten folds * | — | — | 86.1 |

Table 11: NER performance of the WP2-trained model on auto-annotated subsets of WG.

| | $\kappa$ | NE $\kappa$ | $P$ | $R$ | $F$ |
|---|---|---|---|---|---|
| WP2-style | 0.94 | 0.84 | 89.0 | 89.0 | 89.0 |
| WP4-style | 0.93 | 0.83 | 86.8 | 87.6 | 87.2 |

Table 12: Comparing WP2-style WG and WP4-style WG on WG. The automatically annotated data was treated as predicted annotations on WG.

We compare the annotations of WG text more directly by treating the automatic annotations as if they are the output from a tagger run on the 698 and 571 sentences that were confidently chosen. A reasonable agreement between the gold standard and automatic annotation is observed (Table 12), with $F$-scores of 87.2% and 89.0% achieved by WP2 and WP4.

Table 12 also shows inter-annotator agreement calculated between the automatically annotated subsets and the gold-standard annotations in WG, using Cohen's $\kappa$ in the same way as for human annotators. The agreement was very high: equal or better than the agreement between human annotators prior to discussion and correction.

## 8 Conclusion

We have presented the first evaluation of named entity recognition (NER) on a gold-standard evaluation of Wikipedia, a resource of increasing

importance in Computational Linguistics. We annotated a corpus of Wikipedia articles (WG) with gold-standard NE tags. Using this new resource as test data we have evaluated models trained on three gold-standard newswire corpora for NER, and compared them to a model trained on Wikipedia-derived NER annotations (Nothman et al., 2009). We found that this WP2 model outperformed models trained on MUC, CoNLL, and BBN data by more than 7.7% $F$-score.

However, we found that all four models performed significantly worse on the WG corpus than they did on news text, suggesting that Wikipedia as a textual domain is more difficult for NER. We initially suspected that annotation quality was responsible, but found that we had very high inter-annotator agreement even before further discussion and correction of the corpus. This also validates our approach of creating many fine-grained categories and then reducing them down to the four CoNLL types.

To further examine the difficulty of tagging WG, we compared the distribution of fine-grained entity types in WG and BBN, finding a more even distribution over a larger range of types in WG. We found that the standard NER features such as current and previous POS tags and words had lower predictive power on WG. We also compared the distribution of NEs lengths and showed that WG entities are longer on average (for instance song and book titles). This all suggests that Wikipedia is genuinely more difficult to automatically annotate with named entities than newswire.

Finally, we compared the common sentences between Nothman et al.'s (2009) automatic NE annotation of Wikipedia and WG, directly measuring the quality of automatically deriving NE annotations from Wikipedia.

We found that WP2 agreed with our final WG corpus to a high degree, demonstrating that Wikipedia is a viable source of automatically annotated NE annotated data, reducing our dependence on expensive manual annotation for training NER systems.

## Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback. This work was supported by the Australian Research Council under Discovery Project DP0665973. Dominic Balasuriya was supported by a University of Sydney Outstanding Achievement Scholarship. Nicky Ringland was supported by a Capital Markets CRC High Achievers Scholarship. Joel Nothman was supported by a Capital Markets CRC PhD Scholarship and a University of Sydney Vice-Chancellor's Research Scholarship.

## References

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

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16.

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 *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*.

Michael Collins. 2002. Ranking algorithms for named-entity extraction: boosting and the voted perceptron. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 489–496.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716.

James R. Curran and Stephen Clark. 2003a. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 91–98, Budapest, Hungary, 12–17 April.

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

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Daniel Gildea. 2001. Corpus variation and parser performance. In *2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Pittsburgh, PA.

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.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

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, Sep./Oct.):26–33.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26.

Chikashi Nobata, Nigel Collier, and Jun'ichi Tsuji. 2000. Comparison between tagged corpora for the named entity task. In *Proceedings of the Workshop on Comparing Corpora*, pages 20–27.

Joel Nothman, James R Curran, and Tara Murphy. 2008. Transforming Wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 124–132, Hobart, Australia, December.

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, Athens, Greece, March.

Joel Nothman. 2008. *Learning Named Entity Recognition from Wikipedia*. Honours Thesis. School of IT, University of Sydney.

PediaPress. 2007. `mwlib` MediaWiki parsing library. `http://code.pediapress.com`.

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.

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, Edmonton, Canada.

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, Taipei, Taiwan.

Ralph Weischedel and Ada Brunstein. 2005. *BBN Pronoun Coreference and Entity Type Corpus*. LDC2005T33, Linguistic Data Consortium, Philadelphia.

Fei Wu, Raphael Hoffmann, and Daniel S. Weld. 2008. Information extraction from Wikipedia: Moving down the long tail. In *Proceedings of the 14th International Conference on Knowledge Discovery & Data Mining*, Las Vegas, USA, August.