# Analysing Wikipedia and Gold-Standard Corpora for NER Training

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

## Abstract

Named entity recognition (NER) for English typically involves one of three gold standards: MUC, CoNLL, or BBN, all created by costly manual annotation. Recent work has used Wikipedia to automatically create a massive corpus of named entity annotated text.

We present the first comprehensive cross-corpus evaluation of NER. We identify the causes of poor cross-corpus performance and demonstrate ways of making them more compatible. Using our process, we develop a Wikipedia corpus which outperforms gold standard corpora on cross-corpus evaluation by up to 11%.

## 1 Introduction

Named Entity Recognition (NER), the task of identifying and classifying the names of people, organisations and other entities within text, is central to many NLP systems. NER developed from information extraction in the Message Understanding Conferences (MUC) of the 1990s. By MUC 6 and 7, NER had become a distinct task: tagging proper names, and temporal and numerical expressions (Chinchor, 1998).

Statistical machine learning systems have proven successful for NER. These learn patterns associated with individual entity classes, making use of many contextual, orthographic, linguistic and external knowledge features. However, they rely heavily on large annotated training corpora. This need for costly expert annotation hinders the creation of more task-adaptable, high-performance named entity recognisers.

In acquiring new sources for annotated corpora, we require an analysis of training data as a variable in NER. This paper compares the three main gold-standard corpora. We found that tagging mod-els built on each corpus perform relatively poorly when tested on the others. We therefore present three methods for analysing internal and inter-corpus inconsistencies. Our analysis demonstrates that seemingly minor variations in the text itself, starting right from tokenisation can have a huge impact on practical NER performance.

We take this experience and apply it to a corpus created automatically using Wikipedia. This corpus was created following the method of Nothman et al. (2008). By training the C&C tagger (Curran and Clark, 2003) on the gold-standard corpora and our new Wikipedia-derived training data, we evaluate the usefulness of the latter and explore the nature of the training corpus as a variable in NER.

Our Wikipedia-derived corpora exceed the performance of non-corresponding training and test sets by up to 11% $F$-score, and can be engineered to automatically produce models consistent with various NE-annotation schema. We show that it is possible to automatically create large, free, named entity-annotated corpora for general or domain specific tasks.

## 2 NER and annotated corpora

Research into NER has rarely considered the impact of training corpora. The CoNLL evaluations focused on machine learning methods (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) while more recent work has often involved the use of external knowledge. Since many tagging systems utilise gazetteers 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 common entities present in lists or those caught by hand-built rules.

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

Less common 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 Nothman et al. (2008) in order to derive NE-annotated corpora in languages other than English. They classify Wikipedia articles in foreign languages by transferring knowledge from English Wikipedia via inter-language links. 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 when using MUC-style evaluation metrics.

## 2.1 Gold-standard corpora

We evaluate our Wikipedia-derived 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). We consider only the generic newswire NER task, although domain-specific annotated corpora have been developed for applications such as bio-text mining (Kim et al., 2003).

Stylistic and genre differences between the source texts affect compatibility for NER evaluation e.g., the CoNLL corpus formats headlines in all-caps, and includes non-sentential data such as tables of sports scores.

Each corpus uses a different set of entity labels. MUC marks locations (LOC), organisations (ORG) and personal names (PER), in addition to numerical and time information. The CoNLL NER shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) mark PER, ORG and LOC entities, as well as a broad miscellaneous class

| Corpus | # tags | Number of tokens | | |
|---|---|---|---|---|
| | | TRAIN | DEV | TEST |
| MUC-7 | 3 | 83601 | 18655 | 60436 |
| CoNLL-03 | 4 | 203621 | 51362 | 46435 |
| BBN | 54 | 901894 | 142218 | 129654 |

Table 1: Gold-standard NE-annotated corpora

(MISC; e.g. events, artworks and nationalities). BBN annotates the entire Penn Treebank corpus with 105 fine-grained tags (Brunstein, 2002): 54 corresponding to CoNLL entities; 21 for numerical and time data; and 30 for other classes. For our evaluation, BBN's tags were reduced to the equivalent CoNLL tags, with extra tags in the BBN and MUC data removed. Since no MISC tags are marked in MUC, they need to be removed from CoNLL, BBN and Wikipedia data for comparison.

We transformed all three corpora into a common format and annotated them with part-of-speech tags using the Penn Treebank-trained C&C POS tagger. We altered the default MUC tokenisation to attach periods to abbreviations when sentence-internal. While standard training (TRAIN), development (DEV) and final test (TEST) set divisions were available for CoNLL and MUC, the BBN corpus was split at our discretion: sections 03–21 for TRAIN, 00–02 for DEV and 22-24 for TEST. Corpus sizes are compared in Table 1.

## 2.2 Evaluating NER performance

One challenge for NER research is establishing an appropriate evaluation metric (Nadeau and Sekine, 2007). In particular, entities may be correctly delimited but mis-classified, or entity boundaries may be mismatched.

MUC (Chinchor, 1998) awarded equal score for matching *type*, where an entity's *class* is identified with at least one boundary matching, and *text*, where an entity's boundaries are precisely delimited, irrespective of the classification. This equal weighting is unrealistic, as some boundary errors are highly significant, while others are arbitrary.

CoNLL awarded exact (*type* and *text*) phrasal matches, ignoring boundary issues entirely and providing a lower-bound measure of NER performance. Manning (2006) argues that CoNLL-style evaluation is biased towards systems which leave entities with ambiguous boundaries untagged, since boundary errors amount simultaneously to false positives and false negatives. In both MUC and CoNLL, micro-averaged precision, recall and $F_1$ score summarise the results.

Tsai et al. (2006) compares a number of methods for relaxing boundary requirements: matching only the left or right boundary, any tag overlap, per-token measures, or more semantically-driven matching. ACE evaluations instead use a customizable evaluation metric with weights specified for different types of error (NIST-ACE, 2008).

## 3 Corpus and error analysis approaches

To evaluate the performance impact of a corpus we may analyse (a) the annotations themselves; or (b) the model built on those annotations and its performance. A corpus can be considered in isolation or by comparison with other corpora. We use three methods to explore intra- and inter-corpus consistency in MUC, CoNLL, and BBN in Section 4.

### 3.1 N-gram tag variation

Dickinson and Meurers (2003) present a clever method for finding inconsistencies within POS annotated corpora, which we apply to NER corpora. Their approach finds all n-grams in a corpus which appear multiple times, albeit with variant tags for some sub-sequence, the *nucleus* (see e.g. Table 3). To remove valid ambiguity, they suggest using (a) a minimum n-gram length; (b) a minimum margin of invariant terms around the nucleus.

For example, the BBN  TRAIN corpus includes eight occurrences of the 6-gram the San Francisco Bay area ,. Six instances of area are tagged as non-entities, but two instances are tagged as part of the LOC that precedes it. The other five tokens in this n-gram are consistently labelled.

### 3.2 Entity type frequency

An intuitive approach to finding discrepancies between corpora is to compare the distribution of entities within each corpus. To make this manageable, instances need to be grouped by more than their class labels. We used the following groups:

**POS sequences:** Types of candidate entities may often be distinguished by their POS tags, e.g. nationalities are often JJ or NNPS.

**Wordtypes:** Collins (2002) proposed *wordtypes* where all uppercase characters map to A, lowercase to a, and digits to 0. Adjacent characters in the same orthographic class were collapsed. However, we distinguish single from multiple characters by duplication. e.g. USS Nimitz (CVN-68) has wordtype AA Aaa (AA-00).

**Wordtype with functions:** We also map content words to wordtypes only—function words are retained, e.g. Bank of New England Corp. maps to Aaa of Aaa Aaa Aaa..

No approach provides sufficient discrimination alone: wordtype patterns are able to distinguish within common POS tags and vice versa. Each method can be further simplified by merging repeated tokens, NNP NNP becoming NNP.

By calculating the distribution of entities over these groupings, we can find anomalies between corpora. For instance, 4% of MUC's and 5.9% of BBN's PER entities have wordtype Aaa A. Aaa, e.g. David S. Black, while CoNLL has only 0.05% of PERs like this. Instead, CoNLL has many names of form A. Aaa, e.g. S. Waugh, while BBN and MUC have none. We can therefore predict incompatibilities between systems trained on BBN and evaluated on CoNLL or vice-versa.

### 3.3 Tag sequence confusion

A confusion matrix between predicted and correct classes is an effective method of error analysis. For phrasal sequence tagging, this can be applied to either exact boundary matches or on a per-token basis, ignoring entity bounds. We instead compile two matrices: C/P comparing correct entity classes against predicted tag sequences; and P/C comparing predicted classes to correct tag sequences.

C/P equates oversized boundaries to correct matches, and tabulates cases of undersized boundaries. For example, if [ORG Johnson and Johnson] is tagged [PER Johnson] and [PER Johnson], it is marked in matrix coordinates (ORG, PER O PER). P/C emphasises oversized boundaries: if gold-standard Mr. [PER Ross] is tagged PER, it is counted as confusion between PER and O PER. To further distinguish classes of error, the entity type groupings from Section 3.2 are also used.

This analysis is useful for both tagger evaluation and cross-corpus evaluation, e.g. BBN versus CoNLL on a BBN test set. This involves finding confusion matrix entries where BBN and CoNLL's performance differs significantly, identifying common errors related to difficult instances in the test corpus as well as errors in the NER model.

## 4 Comparing gold-standard corpora

We trained the C&C NER tagger (Curran and Clark, 2003) to build separate models for each gold-standard corpus. The C&C tagger utilises a number

| TRAIN | With MISC | | Without MISC | | |
|---|---|---|---|---|---|
| | CoNLL | BBN | MUC | CoNLL | BBN |
| MUC | — | — | *73.5* | 55.5 | 67.5 |
| CoNLL | *81.2* | 62.3 | 65.9 | *82.1* | 62.4 |
| BBN | 54.7 | *86.7* | 77.9 | 53.9 | *88.4* |

Table 2: Gold standard $F$-scores (exact-match)

of orthographic, contextual and in-document features, as well as gazetteers for personal names. Table 2 shows that each training set performs much better on corresponding (same corpus) test sets (italics) than on test sets from other sources, also identified by (Ciaramita and Altun, 2005). NER research typically deals with small improvements ($\sim$1% $F$-score). The 12-32% mismatch between training and test corpora suggests that an appropriate training corpus is a much greater concern. The exception is BBN on MUC, due to differing TEST and DEV subject matter. Here we analyse the variation within and between the gold standards.

Table 3 lists some n-gram tag variations for BBN and CoNLL (TRAIN + DEV). These include cases of schematic variations (e.g. the period in Co .) and tagging errors. Some n-grams have three variants, e.g. the Standard & Poor 's 500 which appears untagged, as the [ORG Standard & Poor] 's 500, or the [ORG Standard & Poor 's] 500. MUC is too small for this method. CoNLL only provides only a few examples, echoing BBN in the ambiguities of trailing periods and leading determiners or modifiers.

Wordtype distributions were also used to compare the three gold standards. We investigated all wordtypes which occur with at least twice the frequency in one corpus as in another, if that wordtype was sufficiently frequent. Among the differences recovered from this analysis are:
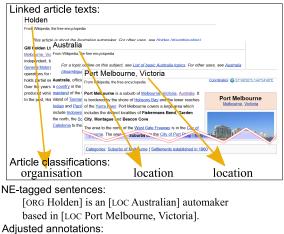
- CoNLL has an over-representation of uppercase words due to all-caps headlines.
- Since BBN also annotates common nouns, some have been mistakenly labelled as proper-noun entities.
- BBN tags text like Munich-based as LOC; CoNLL tags it as MISC; MUC separates the hyphen as a token.
- CoNLL is biased to sports and has many event names in the form of 1990 World Cup.
- BBN separates organisation names from their products as in [ORG Commodore] [MISC 64].
- CoNLL has few references to abbreviated US states.
- CoNLL marks conjunctions of people (e.g. Ruth and Edwin Brooks) as a single PER entity.
- CoNLL text has Co Ltd instead of Co. Ltd.

We analysed the tag sequence confusion when training with each corpus and testing on BBN DEV. While full confusion matrices are too large for this paper, Table 4 shows some examples where the

Wikipedia articles:



Sentences with links:

Holden|Holden is an Australian|Australia automaker based in Port_Melbourne,_Victoria|Port_Melbourne,_Victoria.

Linked article texts:



Article classifications:
organisation    location    location

NE-tagged sentences:

[ORG Holden] is an [LOC Australian] automaker based in [LOC Port Melbourne, Victoria].

Adjusted annotations:

[ORG Holden] is an [MISC Australian] automaker based in [LOC Port Melbourne], [LOC Victoria].

Figure 1: Deriving training data from Wikipedia

NER models disagree. MUC fails to correctly tag U.K. and U.S.. U.K. only appears once in MUC, and U.S. appears 22 times as ORG and 77 times as LOC. CoNLL has only three instances of Mr., so it often mis-labels Mr. as part of a PER entity. The MUC model also has trouble recognising ORG names ending with corporate abbreviations, and may fail to identify abbreviated US state names.

Our analysis demonstrates that seemingly minor orthographic variations in the text, tokenisation and annotation schemes can have a huge impact on practical NER performance.

## 5 From Wikipedia to NE-annotated text

Wikipedia is a collaborative, multilingual, online encyclopedia which includes over 2.3 million articles in English alone. Our baseline approach detailed in Nothman et al. (2008) exploits the hyperlinking between articles to derive a NE corpus.

Since $\sim$74% of Wikipedia articles describe topics covering entity classes, many of Wikipedia's links correspond to entity annotations in gold-standard NE corpora. We derive a NE-annotated corpus by the following steps:

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

| N-gram | Tag | # | Tag | # |
|---|---|---|---|---|
| Co **.** | - | 52 | ORG | 111 |
| Smith Barney **,** Harris Upham & Co. | - | 1 | ORG | 9 |
| the **Contra** rebels | MISC | 1 | ORG | 2 |
| in the **West** is | - | 1 | LOC | 1 |
| that **the** Constitution | MISC | 2 | - | 1 |
| Chancellor of the **Exchequer** Nigel Lawson | - | 11 | ORG | 2 |
| the **world** 's | - | 80 | LOC | 1 |
| **1993** BellSouth Classic | - | 1 | MISC | 1 |
| **Atlanta** Games | LOC | 1 | MISC | 1 |
| **Justice** Minister | - | 1 | ORG | 1 |
| **GOLF** - GERMAN OPEN | - | 2 | LOC | 1 |

Table 3: Examples of n-gram tag variations in BBN (top) and CoNLL (bottom). Nucleus is in bold.

| Tag sequence | | Grouping | # if trained on | | | Example |
|---|---|---|---|---|---|---|
| Correct | Pred. | | MUC | CoNLL | BBN | |
| LOC | LOC | A.A. | 101 | 349 | 343 | U.K. |
| - PER | PER | Aa. Aaa | 9 | 242 | 0 | Mr. Watson |
| - | LOC | Aa. | 16 | 109 | 0 | Mr. |
| ORG | ORG | Aaa Aaa. | 118 | 214 | 218 | Campeau Corp. |
| LOC | - | Aaa. | 20 | 0 | 3 | Calif. |

Table 4: Tag sequence confusion on BBN DEV when training on gold-standard corpora (no MISC)

In Figure 1, a sentence introducing Holden as an Australian car maker based in Port Melbourne has links to separate articles about each entity. Cues in the linked article about Holden indicate that it is an organisation, and the article on Port Melbourne is likewise classified as a location. The original sentence can then be automatically annotated with these facts. We thus extract millions of sentences from Wikipedia to form a new NER corpus.

We classify each article in a bootstrapping process using its category head nouns, definitional nouns from opening sentences, and title capitalisation. Each article is classified as one of: unknown; a member of a NE category (LOC, ORG, PER, MISC, as per CoNLL); a disambiguation page (these list possible referent articles for a given title); or a non-entity (NON). This classifier classifier achieves 89% $F$-score.

A sentence is selected for our corpus when all of its capitalised words are linked to articles with a known class. Exceptions are made for common titlecase words, e.g. I, Mr., June, and sentence-initial words. We also infer additional links — variant titles are collected for each Wikipedia topic and are marked up in articles which link to them — which Nothman et al. (2008) found increases coverage.

Transforming links into annotations that conform to a gold standard is far from trivial. Link boundaries need to be adjusted, e.g. to remove excess punctuation. Adjectival forms of entities (e.g. American, Islamic) generally link to nominal articles. However, they are treated by CoNLL and our

| N-gram | Tag | # | Tag | # |
|---|---|---|---|---|
| of **Batman** 's | MISC | 2 | PER | 5 |
| in **the** Netherlands | - | 58 | LOC | 4 |
| Chicago **,** Illinois | - | 8 | LOC | 3 |
| the **American** and | LOC | 1 | MISC | 2 |

Table 5: N-gram variations in the Wiki baseline

BBN mapping as MISC. POS tagging the corpus and relabelling entities ending with JJ as MISC solves this heuristically. Although they are capitalised in English, personal titles (e.g. Prime Minister) are not typically considered entities. Initially we assume that all links immediately preceding PER entities are titles and delete their entity classification.

## 6 Improving Wikipedia performance

The baseline system described above achieves only 58.9% and 62.3% on the CoNLL and BBN TEST sets (exact-match scoring) with 3.5-million training tokens. We apply methods proposed in Section 3 to to identify and minimise Wikipedia errors on the BBN DEV corpus.

We begin by considering Wikipedia's internal consistency using n-gram tag variation (Table 5). The breadth of Wikipedia leads to greater genuine ambiguity, e.g. Batman (a character or a comic strip). It also shares gold-standard inconsistencies like leading modifiers. Variations in American and Chicago, Illinois indicate errors in adjectival entity labels and in correcting link boundaries.

Some errors identified with tag sequence confusion are listed in Table 6. These correspond to re-

| Tag sequence | | Grouping | # if trained on | | Example |
| Correct | Pred. | | BBN | Wiki | |
|---|---|---|---|---|---|
| LOC | LOC | Aaa. | 103 | 14 | Calif. |
| LOC - LOC | ORG | Aaa , Aaa. | 0 | 15 | Norwalk , Conn. |
| LOC | LOC | Aaa-aa | 23 | 0 | Texas-based |
| - PER | PER | Aa. Aaa | 4 | 208 | Mr. Yamamoto |
| - PER | PER | Aaa Aaa | 1 | 49 | Judge Keenan |
| - | PER | Aaa | 7 | 58 | President |
| MISC | MISC | A. | 25 | 1 | R. |
| MISC | LOC | NNPS | 0 | 39 | Soviets |

Table 6: Tag sequence confusion on BBN DEV with training on BBN and the Wikipedia baseline

sults of an entity type frequency analysis and motivate many of our Wikipedia extensions presented below. In particular, personal titles are tagged as PER rather than unlabelled; plural nationalities are tagged LOC, not MISC; LOCs hyphenated to following words are not identified; nor are abbreviated US state names. Using R. to abbreviate Republican in BBN is also a high-frequency error.

## 6.1 Inference from disambiguation pages

Our baseline system infers extra links using a set of alternative titles identified for each article. We extract the alternatives from the article and redirect titles, the text of all links to the article, and the first and last word of the article title if it is labelled PER.

Our extension is to extract additional inferred titles from Wikipedia's disambiguation pages. Most disambiguation pages are structured as lists of articles that are often referred to by the title $D$ being disambiguated. For each link with target $A$ that appears at the start of a list item on $D$'s page, $D$ and its redirect aliases are added to the list of alternative titles for $A$.

Our new source of alternative titles includes acronyms and abbreviations (AMP links to AMP Limited and Ampere), and given or family names (Howard links to Howard Dean and John Howard).

## 6.2 Personal titles

Personal titles (e.g. Brig. Gen., Prime Minister-elect) are capitalised in English. Titles are sometimes linked in Wikipedia, but the target articles, e.g. U.S. President, are in Wikipedia categories like Presidents of the United States, causing their incorrect classification as PER.

Our initial implementation assumed that links immediately preceding PER entity links are titles. While this feature improved performance, it only captured one context for personal titles and failed to handle instances where the title was only a portion of the link text, such as Australian *Prime Minister-elect* or *Prime Minister* of Australia.

To handle titles more comprehensively, we compiled a list of the terms most frequently linked immediately prior to PER links. These were manually filtered, removing LOC or ORG mentions and complemented with abbreviated titles extracted from BBN, producing a list of 384 base title forms, 11 prefixes (e.g. Vice) and 3 suffixes (e.g. -elect). Using these gazetteers, titles are stripped of erroneous NE tags.

## 6.3 Adjectival forms

In English, capitalisation is retained in adjectival entity forms, such as American or Islamic. While these are not exactly entities, both CoNLL and BBN annotate them as MISC. Our baseline approach POS tagged the corpus and marked all adjectival entities as MISC. This missed instances where nationalities are used nominally, e.g. five Italians.

We extracted 339 frequent LOC and ORG references with POS tag JJ. Words from this list (e.g. Italian) are relabelled MISC, irrespective of POS tag or pluralisation (e.g. Italian/JJ, Italian/NNP, Italian/NNPS). This unfiltered list includes some errors from POS tagging, e.g. First, Emmy; and others where MISC is rarely the appropriate tag, e.g. the Democrats (an ORG).

## 6.4 Miscellaneous changes

**Entity-word aliases** Longest-string matching for inferred links often adds redundant words, e.g. both Australian and Australian people are redirects to Australia. We therefore exclude from inference titles of form $X Y$ where $X$ is an alias of the same article and $Y$ is lowercase.

**State abbreviations** A gold standard may use stylistic forms which are rare in Wikipedia. For instance, the Wall Street Journal (BBN) uses US state abbreviations, while Wikipedia nearly always refers to states in full. We boosted performance by substituting a random selection of US state names in Wikipedia with their abbreviations.

| TRAIN | With MISC | | No MISC | | |
|---|---|---|---|---|---|
| | CoN. | BBN | MUC | CoN. | BBN |
| MUC | — | — | *82.3* | 54.9 | 69.3 |
| CoNLL | *85.9* | 61.9 | 69.9 | *86.9* | 60.2 |
| BBN | 59.4 | *86.5* | 80.2 | 59.0 | *88.0* |
| WP0 – no inf. | 62.8 | 69.7 | 69.7 | 64.7 | 70.0 |
| WP1 | 67.2 | 73.4 | 75.3 | 67.7 | 73.6 |
| WP2 | **69.0** | **74.0** | 76.6 | 69.4 | **75.1** |
| WP3 | 68.9 | 73.5 | **77.2** | **69.5** | 73.7 |
| WP4 – all inf. | 66.2 | 72.3 | 75.6 | 67.3 | 73.3 |

Table 7: Exact-match DEV $F$-scores

| TRAIN | With MISC | | No MISC | | |
|---|---|---|---|---|---|
| | CoN. | BBN | MUC | CoN. | BBN |
| MUC | — | — | *89.0* | 68.2 | 79.2 |
| CoNLL | *91.0* | 75.1 | 81.4 | *90.9* | 72.6 |
| BBN | 72.7 | *91.1* | 87.6 | 71.8 | *91.5* |
| WP0 – no inf. | 71.0 | 79.3 | 76.3 | 71.1 | 78.7 |
| WP1 | 74.9 | 82.3 | 81.4 | 73.1 | 81.0 |
| WP2 | 76.1 | **82.7** | 81.6 | 74.5 | **81.9** |
| WP3 | **76.3** | 82.2 | **81.9** | **74.7** | 80.7 |
| WP4 – all inf. | 74.3 | 81.4 | 80.9 | 73.1 | 80.7 |

Table 8: MUC-style DEV $F$-scores

| Training corpus | DEV (MUC-style $F$) | | |
|---|---|---|---|
| | MUC | CoNLL | BBN |
| Corresponding TRAIN | 89.0 | 91.0 | 91.1 |
| TRAIN + WP2 | 90.6 | 91.7 | 91.2 |

Table 9: Wikipedia as additional training data

| TRAIN | With MISC | | No MISC | | |
|---|---|---|---|---|---|
| | CoN. | BBN | MUC | CoN. | BBN |
| MUC | — | — | *73.5* | 55.5 | 67.5 |
| CoNLL | *81.2* | 62.3 | 65.9 | *82.1* | 62.4 |
| BBN | 54.7 | *86.7* | 77.9 | 53.9 | *88.4* |
| WP2 | 60.9 | 69.3 | 76.9 | 61.5 | 69.9 |

Table 10: Exact-match TEST results for WP2

| TRAIN | With MISC | | No MISC | | |
|---|---|---|---|---|---|
| | CoN. | BBN | MUC | CoN. | BBN |
| MUC | — | — | *81.0* | 68.5 | 77.6 |
| CoNLL | *87.8* | 75.0 | 76.2 | *87.9* | 74.1 |
| BBN | 69.3 | *91.1* | 83.6 | 68.5 | *91.9* |
| WP2 | 70.2 | 79.1 | 81.3 | 68.6 | 77.3 |

Table 11: MUC-eval TEST results for WP2

**Removing rare cases** We explicitly removed sentences containing title abbreviations (e.g. Mr.) appearing in non-PER entities such as movie titles. Compared to newswire, these forms as personal titles are rare in Wikipedia, so their appearance in entities causes tagging errors. We used a similar approach to personal names including of, which also act as noise.

**Fixing tokenization** Hyphenation is a problem in tokenisation: should London-based be one token, two, or three? Both BBN and CoNLL treat it as one token, but BBN labels it a LOC and CoNLL a MISC. Our baseline had split hyphenated portions from entities. Fixing this to match the BBN approach improved performance significantly.

## 7 Experiments

We evaluated our annotation process by building separate NER models learned from Wikipedia-derived and gold-standard data. Our results are given as micro-averaged precision, recall and $F$-scores both in terms of MUC-style and CoNLL-style (exact-match) scoring. We evaluated all experiments with and without the MISC category.

Wikipedia's articles are freely available for download.[1] We have used data from the 2008 May 22 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 performed with a Wikipedia corpus of 3.5 million tokens. Although we had up to 294 million tokens available, we were limited by the RAM required by the C&C tagger training software.

## 8 Results

Tables 7 and 8 show $F$-scores on the MUC, CoNLL, and BBN development sets for CoNLL-style exact

match and MUC-style evaluations (which are typically a few percent higher). The cross-corpus gold standard experiments on the DEV sets are shown first in both tables. As in Table 2, the performance drops significantly when the training and test corpus are from different sources. The corresponding TEST set scores are given in Tables 9 and 10.

The second group of experiments in these tables show the performance of Wikipedia corpora with increasing levels of link inference (described in Section 6.1). Links inferred upon matching article titles (WP1) and disambiguation titles (WP2) consistently increase $F$-score by ~5%, while surnames for PER entities (WP3) and all link texts (WP4) tend to introduce error. A key result of our work is that the performance of non-corresponding gold standards is often significantly exceeded by our Wikipedia training data.

Our third group of experiments combined our Wikipedia corpora with gold-standard data to improve performance beyond traditional train-test pairs. Table 9 shows that this approach may lead

| Token | Corr. | Pred. | Count | Why? |
|-------|-------|-------|-------|------|
| . | ORG | - | 90 | Inconsistencies in BBN |
| House | ORG | LOC | 56 | Article White House is a LOC due to classification bootstrapping |
| Wall | - | LOC | 33 | Wall Street is ambiguously a location and a concept |
| Gulf | ORG | LOC | 29 | Georgia Gulf is common in BBN, but Gulf indicates LOC |
| , | ORG | - | 26 | A difficult NER ambiguity in e.g. Robertson , Stephens & Co. |
| 's | ORG | - | 25 | Unusually high frequency of ORGs ending 's in BBN |
| Senate | ORG | LOC | 20 | Classification bootstrapping identifies Senate as a house, i.e. LOC |
| S&P | - | MISC | 20 | Rare in Wikipedia, and inconsistently labelled in BBN |
| D. | MISC | PER | 14 | BBN uses D. to abbreviate Democrat |

Table 12: Tokens in BBN DEV that our Wikipedia model frequently mis-tagged

| Class | By exact phrase | | | By token | | |
|-------|-----|-----|-----|-----|-----|-----|
| | $P$ | $R$ | $F$ | $P$ | $R$ | $F$ |
| LOC | 66.7 | 87.9 | 75.9 | 64.4 | 89.8 | 75.0 |
| MISC | 48.8 | 58.7 | 53.3 | 46.5 | 61.6 | 53.0 |
| ORG | 76.9 | 56.5 | 65.1 | 88.9 | 68.1 | 77.1 |
| PER | 67.3 | 91.4 | 77.5 | 70.5 | 93.6 | 80.5 |
| All | 68.6 | 69.9 | 69.3 | 80.9 | 75.3 | 78.0 |

Table 13: Category results for WP2 on BBN TEST

to small $F$-score increases.

Our per-class Wikipedia results are shown in Table 13. 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, where Nothman et al. (2008) had reported results 13% higher than phrasal matching, reflecting a failure to correctly identify entity boundaries. We have reduced this difference to 9%. A BBN-trained model gives only 5% difference between phrasal and token $F$-score.

Among common tagging errors, we identified: 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; mis-labelling un-tagged title-case terms and tagged lowercase terms in the gold-standard.

The corpus analysis methods described in Section 3 show greater similarity between our Wikipedia-derived corpus and BBN after implementing our extensions. There is nonetheless much scope for further analysis and improvement. Notably, the most commonly mis-tagged tokens in BBN (see Table 12) relate more often to individual entities and stylistic differences than to a generalisable class of errors.

## 9   Conclusion

We have demonstrated the enormous variability in performance between using NER models trained and tested on the same corpus versus tested on other gold standards. This variability arises from not only mismatched annotation schemes but also stylistic conventions, tokenisation, and missing frequent lexical items. Therefore, NER corpora must be carefully matched to the target text for reasonable performance. We demonstrate three approaches for gauging corpus and annotation mismatch, and apply them to MUC, CoNLL and BBN, and our automatically-derived Wikipedia corpora.

There is much room for improving the results of our Wikipedia-based NE annotations. In particular, a more careful approach to link inference may further reduce incorrect boundaries of tagged entities. We plan to increase the largest training set the C&C tagger can support so that we can fully exploit the enormous Wikipedia corpus.

However, we have 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. Our results indicate that Wikipedia data can perform better (up to 11% for CoNLL on MUC) 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.

# 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 *Proc. 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 on Association for Computational Linguistics*, pages 489–496, Morristown, NJ, USA.

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.

Markus Dickinson and W. Detmar Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 107–114, Budapest, Hungary.

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.

Christopher Manning. 2006. Doing named entity recognition? Don't optimize for $F_1$. In *NLPers Blog*, 25 August. http://nlpers.blogspot.com.

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.

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

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

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.

NIST-ACE. 2008. Automatic content extraction 2008 evaluation plan (ACE08). NIST, April 7.

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, Hobart.

Alexander E. Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *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.

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*.

Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung, and Wen-Lian Hsu. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7:96–100.

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