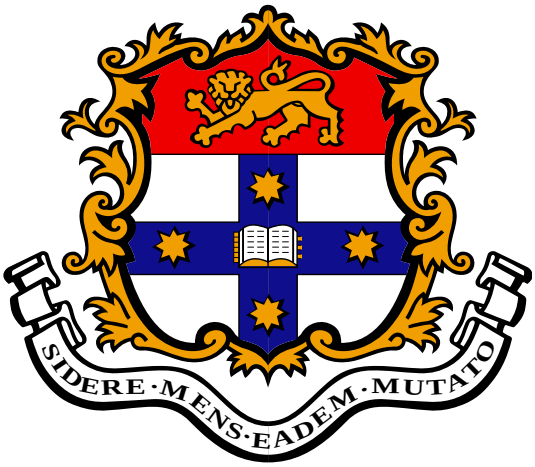
Learning Named Entity Recognition from Wikipedia

JOEL NOTHMAN
SID: 200319377



Supervisor: James Curran and Tara Murphy

This thesis is submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Science (Honours)

School of Information Technologies
The University of Sydney
Australia

7 November 2008

Abstract

We present a method to produce free, enormous corpora to train taggers for Named Entity Recognition (NER), the task of identifying and classifying names in text, often solved by statistical learning systems. Our approach utilises the text of Wikipedia, a free online encyclopedia, transforming links between Wikipedia articles into entity annotations. Having derived a baseline corpus, we found that altering Wikipedia's links and identifying classes of capitalised non-entity terms would enable the corpus to conform more closely to gold-standard annotations, increasing performance by up to 32% F score.

The evaluation of our method is novel since the training corpus is not usually a variable in NER experimentation. We therefore develop a number of methods for analysing and comparing training corpora. Gold-standard training corpora for NER perform poorly (F score up to 32% lower) when evaluated on test data from a different gold-standard corpus. Our Wikipedia-derived data can outperform manually-annotated corpora on this cross-corpus evaluation task by up to 7% on held-out test data. These experimental results show that Wikipedia is viable as a source of automatically-annotated training corpora, which have wide domain coverage applicable to a broad range of NLP applications.

Acknowledgements

Firstly, I would like to thank my supervisors Dr. James Curran and Dr. Tara Murphy. James first welcomed me to the world of Computational Linguistics and named entities in 2004, and has kept me in tow ever since. Thanks to Tara for allowing James to take me on as an honours student despite an agreed quota. Both were wonderful at encouraging and inspiring my work throughout the project (and handling frequent, verbose email correspondence), and were often able to complement each others' supervision to my work's advantage. Thank you.

As well as my supervisors, I thank James Haggerty and Adam Hotz for volunteering to proofread my work, giving me valuable outsiders' perspectives.

Also to Jon Patrick, for challenging my work from the outset and bringing me back to the bigger issues and the ever untamed world of semantics.

I must thank the women of SIT 4E: Susan Howlett, Katie Bell, Nicky Ringland, Mae Kitvitee and Tara McIntosh who supplied essential company, critique, chocolate and tea throughout the year.

I also acknowledge the contribution of the Sydney University Honours Scholarship to easing my lifestyle if only a little.

Finally, to my family for being constantly supportive of my leech-like lifestyle, and being encouraging and patient even while not understanding my work.

And to Galina, perhaps the most patient of all, who waited before forcing her way into my life, which has only been for the better.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

Named Entity Recognition (NER) is the task of processing text to identify and classify names, an important component in many Natural Language Processing (NLP) applications, enabling the extraction of useful information from documents. NER is often performed using a statistical tagger which learns patterns for the recognition of names from manually-annotated textual corpora. A few corpora have been constructed as *gold standard*s—i.e. they define correct annotations for NER by example. They are commonly used to train statistical machine learners but are limited in scope due to the cost of manual annotation. This is a problem because others have shown that more training data leads to higher accuracy language models (Banko and Brill, 2001). The automatic construction of large training corpora would enable the training of high-accuracy, wide-coverage named entity (NE) taggers.

We present a method for automatically obtaining such a corpus from Wikipedia[1], an enormous, free online encyclopedia, taking advantage of its internal links and their role in disambiguating term reference. Having produced such a corpus, we evaluate it by training the C&C NER tagger (Curran and Clark, 2003b) and comparing the performance of the model produced with models learnt from standard manually-annotated data. This corpus derivation and evaluation process is summarised in Figure 1.1.

1.1 Named Entity Recognition

In evaluations at the Message Understanding Conferences of the 1990s, it became clear that in order to reasonably extract information from documents, it is useful to first identify certain classes of information referred to in the text. They therefore established the Named Entity Task, where systems attempted to identify dates, times, numerical information and names (Chinchor and Robinson, 1997). Named Entity Recognition has remained an essential component of Information Extraction and related NLP tasks.

[1] `http://en.wikipedia.org`

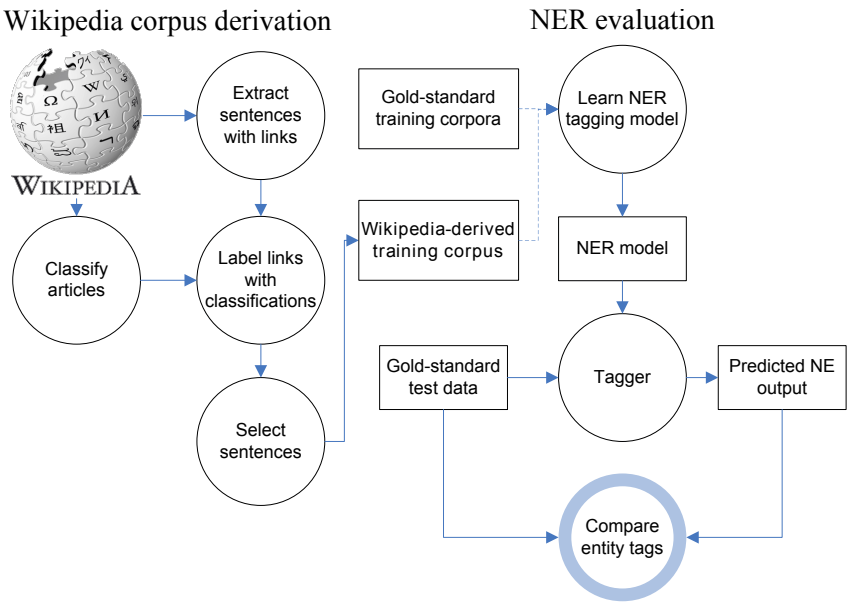


FIGURE 1.1: Data flow overview of our corpus derivation and evaluation

While time and number data can often be found using regular expressions, the identification and classification of names often involves challenging ambiguity. Consider the sentence Paris Hilton visited the Paris Hilton. This example illustrates the fact that Paris may refer to a personal name or a place, apart from numerous other referents including film titles or a genus of plants. Similarly, Hilton is a first or a last name, a hotel corporation or any one of the corporation's individual hotels. A named entity (NE) tagging scheme may define the correct annotation of the above sentence as:[2]

[PER Paris Hilton] visited the [LOC Paris] [ORG Hilton].

A computer tagging this sentence cannot inherently understand the ambiguity between a personal name and an organisation with its location, but it can use knowledge in the form of rules (e.g. names before visited are people; those following the are not[3]) or statistics to make tag boundary and class decisions.

While rule-based systems were initially popular for NER, statistical machine learning techniques allow the incorporation of many more linguistic properties in tagging decisions (Borthwick, 1999), and may be reapplied to new texts, domains or languages without having to laboriously reconstruct a list of rules. Machine learning approaches to NER require the availability of training corpora whose gold-standard

[2]Note that this is not the only possible annotation of this sentence, considering (a) ambiguity of Hilton as an organisation or a particular location; (b) use of the city Paris adjectivally; (c) nested entities, e.g. [FAC [LOC Paris] [ORG Hilton]], may be more accurate. While the ACE evaluations (Doddington et al., 2004) consider nested entities, the task traditionally considers non-overlapping entities, and provides schema (e.g. Chinchor et al. (1999)) to describe their annotation.

[3]Some counter-examples: [ORG Microsoft] visited [LOC Washington]; . . . the [PER Barack Obama] supporters.

annotations can be analysed statistically to produce a predictive model. Because training texts are traditionally annotated manually by linguistic experts, they are costly to produce and generally small in size (up to 1.2 million tokens[4] in BBN's annotation of Wall Street Journal text (Weischedel and Brunstein, 2005)). These corpora also train models in the styles and topics of their particular source texts, and may include only dated knowledge (i.e. it may be difficult to identify the name Barack Obama with training data from 1989 news), limiting their broad reapplication to various contemporary real-world tasks. For these reasons, a NER system trained on data from one manually-annotated source may perform very poorly on another manually-annotated corpus (Ciaramita and Altun, 2005).

With the ability to automatically add accurate entity annotations to a large body of text, preferably incorporating various genres and up-to-date knowledge, we may train a NER system with much wider applicability than traditionally-trained systems, and with more data from which to learn patterns for recognising entity names. Moreover, the corpus need not rely on copyrighted texts and annotations.

1.2 Wikipedia

Wikipedia[5], an online collaboratively-authored encyclopedia, has many properties which make it a suitable source of huge NE-annotated corpora. In English, it covers over 2.6 million articles and is constantly being updated. Unlike other online resources, Wikipedia data is *semi-structured*, incorporating both sentential natural language portions and structural features such as links between articles, *infoboxes* containing key facts and topic categorisation as illustrated in Figure 1.2. These aid its use as a source of knowledge for NLP applications. Controversy regarding the collaborative encyclopedia's factuality and use as a reference work (Giles, 2005; Encyclopædia Britannica, Inc., 2006) does not undermine its use in such applications. Wikipedia has therefore recently become a popular repository for mining natural language and ontological knowledge.

Our own method to transform Wikipedia into a named entity-annotated corpus relies on the fact that links between Wikipedia articles often correspond to entity annotations. By using structural features to classify Wikipedia's articles into entity classes, we can use an article's class to label links which target it. This application of Wikipedia data relies only on the integrity of features used for classification; links targeting appropriate articles; and text having a sentential form that is usable as NER training data. This enables the creation of free NE-annotated corpora much larger than have previously been available.

[4] Tokens are generally words or punctuation.

[5] `http://www.wikipedia.org`

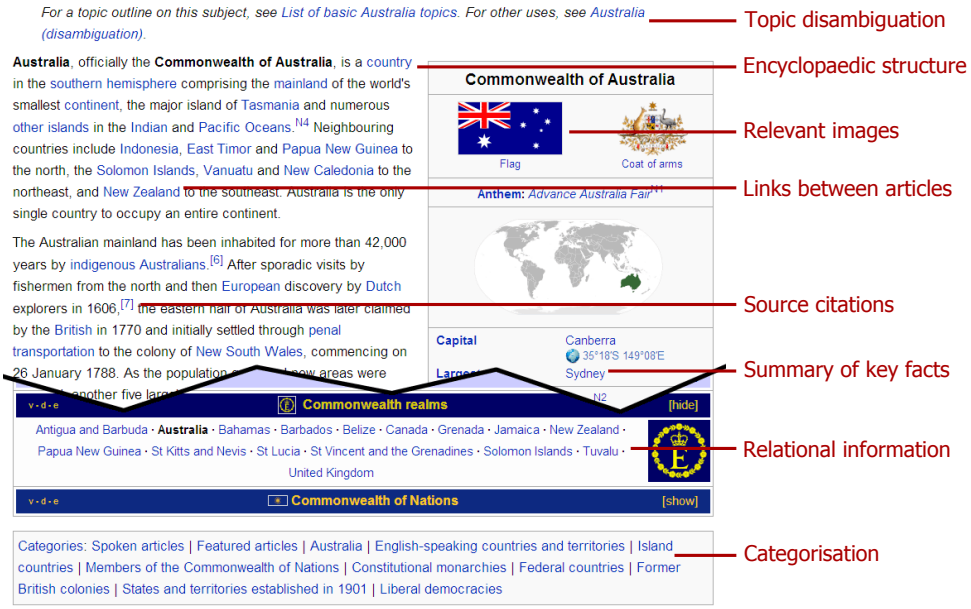


FIGURE 1.2: Useful structural features of the Australia Wikipedia article

1.3 Our contributions

Our work to improve NER performance by providing a readily-updated training resource appears in the context of recent research into advancing NER and using Wikipedia as a computational resource (see Chapter 2). Unlike most of the NER literature, we consider the training corpus as an experimental variable, which has led to the development of tools to analyse and compare training corpora, also providing insight into the poor performance of gold-standard corpora when evaluated on each other (Chapter 3). We have found that our basic approach to transforming Wikipedia links into NE annotations described in chapters 4 and 5 is greatly improved by adjusting Wikipedia links to conform to gold-standard corpus annotations among other extensions (Chapter 6). Our final corpus then generally outperforms systems trained on one gold-standard corpus and tested on another (by up to 7% on final test data). Although this work leaves many opportunities for further experimentation and development of evaluation techniques (Chapter 8), our experimental results demonstrate the viability of automatically deriving huge, free and competitive-performance named entity-annotated corpora from Wikipedia.

Preliminary results from our work will be published in the *Proceedings of the Australian Language Technology Workshop*, December 2008.

CHAPTER 2

Background

2.1 Named Entity Recognition

Named Entity Recognition (NER) involves *identifying* names within text and *classifying* each instance. This processing has become a standard component of NLP systems, such as Information Extraction, Information Retrieval, Summarisation and Question Answering. The task originally developed from within Information Extraction (IE) at the DARPA-funded Message Understanding Conferences (MUC) of the 1990s. For the final two MUC conferences (MUC-6, 1995; MUC-7, 1998), NER was defined as a separate evaluation task (Chinchor and Robinson, 1997), along with template-based IE and Coreference tasks. For the MUC evaluations, systems competed to tag portions of texts that name people, locations or organisations (together ENAMEX), numerical information (NUMEX) or date-time reference (TIMEX). For example, they would specify the SGML markup of "Murdoch's News Corp." as:

```
<B_ENAMEX TYPE="PERSON">Murdoch</B_ENAMEX>'s <B_ENAMEX TYPE="ORGANIZATION">
News Corp.</B_ENAMEX>
```

Throughout this thesis we will use a conceptual representation independent of storage format:

[PER Murdoch]'s [ORG News Corp.]

While NER as a word sequence classification task requires that the number of named entity categories is fixed, later evaluations considered a broader range of named entity categories than those defined for MUC. The Conference on Computational Natural Language Learning (CoNLL) 2002 and 2003 shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) distinguished the problem of finding names from that of identifying numerical and temporal information. They focused on four named-entity classes: PER, LOC, ORG and MISC, the latter very broadly including everything from nationalities (*Italian*) to events (*Sydney 2000 Olympic Games*).

The Automatic Content Extraction (ACE) Evaluation added "geopolitical entity" to handle the ambiguity of entities like countries which have semantic properties of both locations and organisatons, and also subdivided broad categories (LDC, 2008). Fine-grained hierarchies, including many common noun "entities" (colours, animals, etc.), are useful in many information extraction tasks such as Question Answering. BBN presents a collection of 29 basic "answer types" (Brunstein, 2002), while Li and Roth (2006) classify entity-related answers into 31 total types.[1] Sekine et al. (2002) have systematically derived their *Extended Named Entity Hierarchy* which includes 150 types altogether, grouped into 11 classes apart from numerical types.

Although NER was originally developed to consider names of generic entity types in newswire text, the task is often applied to domain-specific texts, where it may be useful to identify mentions of proteins and viruses (Kim et al., 2003) or stars and satellites (Murphy et al., 2006).

Others (Evans, 2003; Suchanek et al., 2007) have widened the domain of classification to the *lexicalised concepts*, or *synsets*, contained in WordNet (Fellbaum, 1998), a machine-readable thesaurus. Its 117,798 nouns form 82,115 concepts[2] in a hierarchy under 25 basic noun-types (Miller, 1998), and since it associates related common nouns, it is useful for categorising terms or labelling lexical clusters. Rather than classifying entity references, the related task of *Named Entity Disambiguation* attempts to identify the particular individual specified by a given name in context from a set of candidates. Recent research has promoted Wikipedia for this task (Bunescu and Paşca, 2006; Cucerzan, 2007; Mihalcea and Csomai, 2007), which will be discussed further in section 2.3.3.

The appropriate granularity of classification is essentially a compromise between a useful level of discrimination for a given task and what a NER system is able to reliably identify.

Named Entity Recognition includes many categories of ambiguity that make it difficult to establish an appropriate evaluation metric (Nadeau and Sekine, 2007). As illustrated in Figure 2.1, one entity may have many plausible correct annotations, which differ due to *boundary* and *classification* ambiguities. Although MUC attempted to design a scheme for resolving annotation ambiguities (Chinchor et al., 1999), they allowed for the inherent discrepancy in the task when designing an evaluation metric.

. . . sanctioned by the [LOC U.S] .
. . . sanctioned by the [LOC U.S .]
. . . sanctioned by [LOC the U.S] .
. . . sanctioned by the [ORG U.S] .

FIGURE 2.1: Possibly valid entity annotations for "sanctioned by the U.S .". The first three examples illustrate boundary ambiguity, while the last considers U.S. as referring to an organisation, i.e. a government, an example of classification ambiguity.

Scoring method	Matches	
	Boundaries	Class
TOKEN	Each token	Yes
TYPE	Left or right	Yes
TEXT	Left and right	No
EXACT	Left and right	Yes
MUCEVAL	Average of TYPE and TEXT	

TABLE 2.1: Common scoring methods for NER

2.1.1 Evaluating NER

MUC (Chinchor, 1998) equally awarded 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 assigned. This equal weighting is unrealistic, as some boundary errors are highly significant, while others are arbitrary.

CONLL simply awarded exact phrasal matches, ignoring boundary issues entirely. Manning (2006) argues that the CONLL evaluation is biased towards systems which leave entities with ambiguous boundaries untagged, since boundary errors amount simultaneously to false positives (an entity is tagged which is not in the gold-standard annotation) and false negatives (the gold-standard entity is not matched). The simplicity of the CONLL evaluation nonetheless makes it appealing and its popularity is aided by a free PERL evaluation script.[3] It also provides a lower bound on other possible scoring methods.

Tsai et al. (2006) compare a number of approaches to relaxing entity boundary requirements and their application to biological NER: matching only the left or right boundary, any tag overlap, per-token measures, or more semantically-driven matching. Common scoring methods are summarised in Table 2.1.

[1]This total excludes 13 numerical question types and another six classes of non-entities such as abbreviations and definitions.

[2]See `http://wordnet.princeton.edu/man/wnstats.7WN`.

[3]`http://nextens.uvt.nl/~conll/software.html`

For each of the scoring methods discussed, true positives (correctly predicted), false positives (incorrectly predicted) and false negatives (unpredicted) are counted, such that precision ($P = \frac{tp}{tp+fp}$) and recall ($R = \frac{tp}{tp+fn}$) can be calculated for each class and overall (micro-average). Whether precision or recall is a greater priority depends on the application to which NER is applied. The harmonic mean of precision and recall known as F measure (or F-score) provides a single metric which balances the two equally:

$$F = \frac{2PR}{P+R} \tag{2.1}$$

Note that on the basis of his critiques cited above, Manning (2006) suggests that optimising NER systems for F-score may be detrimental, and proposes counting labelling errors, boundary errors and label-boundary errors in addition to exact-match tp, fp, fn and tn; no metric is suggested which combines these raw counts. Over two thirds of the errors of his own NER system belonged to these three additional categories. Manning's method requires segmenting a document into candidate strings which can each be categorised and counted, but there are some problems with his method such as the validity of his proposed segmentation being dubious and biasing results towards certain types of prediction error.[4]

Acknowledging the significance of some errors over others, ACE evaluations avoid traditional F-based scoring by using a customisable evaluation metric with weights specified for different types of error (NIST-ACE, 2008).

The difficulties in evaluating NER are unresolved, and results reported using different evaluation methods are incomparable.

2.1.2 Methods of NE identification

A number of cues are used to identify named entities. McDonald (1996) introduced the concepts of *internal* evidence (e.g. Ltd. within ORG entities) and *external* or contextual evidence (e.g. CEO or Dr. before PER entities) by which many may be recognised. Most early systems consisted primarily of manually-built lists of such cues. A majority of teams involved in the MUC evaluations utilised hand-crafted rules, and while the best teams competing were able to achieve F-scores close to human annotators (see Table 2.2; Marsh and Perzanowski (1998)), this method of laboriously constructing rules produces a largely language- and domain-dependent result.[5]

[4] See my comments on the Manning (2006) blog post.

[5] Overall results are inflated in comparison to other evaluations which do not include numerical and time entities whose patterns are more predictable.

Statistic	MUC-6	MUC-7
Median	92.7	84.7
Maximum	96.4	93.4
Human annotator 1	96.7	97.6
Human annotator 2	93.2	97.0

TABLE 2.2: MUCEVAL F-score results for the systems in MUC-6 and MUC-7 NER compared to human annotation. (Marsh and Perzanowski, 1998)

The primary alternative approach uses statistical machine learning (ML) in which a system learns patterns from an annotated training corpus, allowing it to predict the most likely NE in a given context. Assuming the availability of appropriate training texts, a single machine-learning system may easily be applied to varying languages, domains or classification schemes.

Two of the top four entrants in MUC-7 used machine learning approaches: Among the early adopters of a ML approach, Bikel et al. (1999) used a series of class-specific Hidden Markov Models in their commercially-successful IdentiFinder to build a model of the language associated with each entity type. Since HMMs rely on having previously seen patterns, their approach uses a number of backoff strategies. Maximum entropy modelling, as used by Borthwick et al. (1998), allowed for many features to be incorporated without a backoff scheme, and their best results were achieved by using the output of multiple high-performance rule-based systems in addition to linguistic features.

The machine learning focus of the CONLL-2002 evaluation encouraged various statistical techniques, and allowed for cross-linguistic application and evaluation that was not as feasible with manual rule construction (Tjong Kim Sang, 2002).[6] Models included Support Vector Machines (SVM), AdaBoost, transformation-based learning and maximum entropy modelling. The top system at CONLL-2003 combined the classification decisions of a number of machine learners (Florian et al., 2003). In addition to the applicability to new languages and domains, Borthwick (1999) emphasises the fact that statistical systems are able to take advantage of a diverse range of knowledge sources in predicting NE annotations, and are not as subject to the human bias present in manual rule construction.

One result of the CONLL-2002 shared task was the realisation that while choosing an appropriate machine learning technique affected performance, "the choice of features is at least as important." (Tjong Kim Sang and De Meulder, 2003) Their overview of entrants in the CONLL-2003 evaluation compares

[6] While MUC included evaluations in languages other than English, few teams competed, and only four entrants presented systems for multiple languages.

Statistic	Spanish	Dutch	English	German
Baseline	62.5	57.6	59.6	30.3
Median	74.9	71.0	84.2	67.9
Maximum	81.4	77.1	88.8	72.4

TABLE 2.3: Baseline, median and maximum EXACT-match F-scores for the 12 entrants in CONLL-2002 (Spanish and Dutch), and the 16 entrants in the CONLL-2003 (English and German) shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). The 2002 baselines are adjusted according to footnote 5 in (Tjong Kim Sang, 2002).

the types of features used in each competing system. They could each be classed into one of three groups:

Local context: By far the most universal class of features, this includes text tokens (unigrams, bigrams, etc.), affixes, orthographic cues (e.g. capitalisation), part-of-speech (POS) tags and phrasal chunks.

Global context: Particularly prevalent in Chieu and Ng (2003), these features mark the reappearance of terms in the target document that appeared elsewhere in training, noting features such as capitalisation, possible expansions of acronyms, and the NE class assigned to previous occurrences of the term, among others.

External knowledge: Many systems rely on the lookup of terms in external lists of names, while others such as Curran and Clark (2003b) used unannotated data to obtain knowledge for capitalisation-based features.

Other entrants showed some success with primarily character-based models (Klein et al., 2003; Whitelaw and Patrick, 2003), following on from the work of Cucerzan and Yarowsky (1999).

The results of the CONLL-2003 shared task showed that a number of systems could achieve uniformly high performance (see Table 2.3); results in German were much lower (but further improved from the baseline), possibly because all nouns are capitalised, providing less distinction for NEs, and also because of fewer available knowledge sources. The CONLL tasks unfortunately did not publish a measure of human inter-annotator agreement; MUC-7 reported 3-5% human error for ORG, 1% for PER and 1-3% for LOC[7] (Marsh and Perzanowski, 1998). Since a number of successful features for machine-learning NER have been identified, the challenge may now lie in providing them with sufficient and high-quality training material from which to learn.

[7]Their best entrant had 13%, 5% and 10% error on these categories, respectively.

Just as statistical systems avoid the need for manual rule construction to identify named entities, some recent NER systems remove the need for annotated training data through unsupervised approaches. One unsupervised approach combines NE lists with disambiguation rules. A state-of-the-art system (Etzioni et al., 2005) for generating lists of named entities for a class X searches the web for phrases like "X, such as [Y]" in attempting to find lists of items Y. Nadeau et al. (2006) uses such automatically-acquired lists to mark entities in texts, along with unsupervised means of disambiguating entity-noun ambiguity, entity-entity ambiguity, and entity boundaries. Using only their web-derived lists and some language-independent algorithms, their system outperformed the MUC-7 baseline, but could not compete with its top entrants.

A converse approach is taken by Evans (2003), who identifies named entity strings within a document (using capitalisation features) and uses the web to find common nouns X that describe such an entity Y, by searching for phrases like "[X,] such as Y". Given lists of common nouns related to each name in a document, they can be clustered over the WordNet hierarchy, and each cluster can be labelled with its minimal bounding synset. Given the clustering-based approach, these labels may be arbitrary in their granularity and so are not necessarily appropriate for many tasks, although a classification-based variant may provide an approach to NER. Evans's (2003) method ignores the significant ambiguity issue, and since it acquires online knowledge during processing is relatively inefficient.

These unsupervised approaches have in common the ability to classify entities within various domains of categorisation, largely cross-linguistically, and an independence from training corpora that are difficult to develop. The former system (Nadeau et al., 2006) is nonetheless far from the best-performing statistical NER systems, and the latter in its current form is very difficult to evaluate in terms of conventional NER. Neither is able to take advantage of extensive linguistic cues for recognising NEs.

In comparison to rule-based and unsupervised approaches, statistical NER best handles unseen or ambiguous names, while being adaptable to new languages and domains, given appropriate training data.

2.2 Knowledge sources for NER

Named Entity Recognition systems are only as reliable as their training sources. Rule-based, statistical and unsupervised systems alike may make use of lists of names categorised into entity types, often

referred to as *gazetteers*.[8] Gazetteer-based approaches require additional methods to resolve ambiguity and unknown names. Machine learning approaches to NER are able to take advantage of learned patterns, and such knowledge is contingent on the availability of training data. This suggests that additional sources of annotated training data are able to benefit statistical NER.

2.2.1 Gazetteers

It has often been assumed that reasonable NER performance can be achieved merely by list-lookup for familiar names (Cucchiarelli et al., 1998). With the assumption that larger categorised lists of names may improve system recall, a number of approaches have been implemented to automatically acquire such lists from the web often using context patterns and bootstrapping (Riloff and Jones, 1999; Thelen and Riloff, 2002; Etzioni et al., 2005; Talukdar et al., 2006), or from Wikipedia (Toral and Muñoz, 2006). Mikheev et al. (1999) bring extensive arguments against the assumption that larger gazetteers aid NER:

- such lists need to be enormous and cover naming variations;
- there is ambiguity with common nouns and between entities (in this vein, Stevenson and Gaizauskas (2000) report a perfect-recall list lookup approach with only 88% precision); and
- linguistic data sparseness means no list can approach completeness.

Moreover, Mikheev et al. (1999) bring evidence from a leading MUC-7 competitor (Krupka and Hausman, 1998) that found little performance loss in reducing 25,000 gazetteer entries to 9,000, but had dramatic improvement with a careful selection of forty-two.

Mikheev et al.'s own NER system (1999) was tested without a gazetteer and gave only small increases in error for ORG and PER classes, but significant performance losses (from 6 to 48% error) for LOC, which were largely alleviated with a short list of common locations. This particular dependence on geographic gazetteers seems to be system-specific, though: their implementation relies initially on lists of cues (e.g. *Mrs.*, *Ltd.*, *Inc.*) that are less available for location identification. Machine learning techniques have been able to produce high accuracy for LOC without gazetteer information, and some authors choose to use only personal name gazetteers (e.g. Curran and Clark (2003b)). Stevenson and Gaizauskas (2000) confirm that gazetteer *size* is not key, and that lists extracted from the web are most effective when filtered. Evaluations from CONLL-2003 nonetheless reported up to 22% error reduction for the English

[8] Historically the term gazetteer has referred to exhaustive lists of geographic names with associated information; here the term is applied more generally to extensive lists of names of any class.

corpus and 15% for German when gazetteer data was incorporated (Tjong Kim Sang and De Meulder, 2003), although one of the best performers in both languages used none at all (Klein et al., 2003). It seems that selectivity in the use of lists can provide greater performance value than large gazetteers.

In a novel extension to the use of lists, Kazama and Torisawa (2008) note that for statistical systems, gazetteers do not need to group entities into the target entity classes. Any knowledge source which can be used to attach the same label to semantically similar entities may be added as a feature for machine learning. They improved NER performance by 1.6% F-score with a feature based on a cluster labels for entities. While this approach may have advantages over traditional list methods in resolving ambiguity, it is still only able to provide an advantage for known entities.

2.2.2 Training corpora

Data-driven statistical approaches are popular in contemporary computational linguistics, although the time and monetary costs of manually producing training corpora are prohibitive. For NER training, the only data widely available are corpora used in conference evaluations of named entity technology (MUC, IEER, ACE and CONLL), or for specific domains such as biomedicine, and many require purchase, relying on copyrighted materials. While these are useful for evaluating and comparing NER systems, they are not necessarily sufficient training data to produce systems capable of high-accuracy real-world NER. For instance, the top-performing system in CONLL-2003 made auxiliary use of two classifiers trained on a private data collection. Training corpora provide patterns and context that NER systems can learn, unlike gazetteers which although easily generated do not provide sufficient information for machine learners. Hence it is appealing to find low-cost ways to generate new corpora.

One approach involves extracting sentences from the web. An et al. (2003) used a simple approach of searching the web for a given unambiguous named entity (in Korean), extracting sentences that contain it, and tagging the known entity for use in a training corpus. This is limited in that it does not provide any evidence for disambiguation, and cannot produce annotated sentences that contain multiple entities (unless all are known). It also loses the applicability of features related to long-distance dependencies that some have found advantageous for NER (what Chieu and Ng (2003) call *global features*), but unlike the use of lists alone may help identify sentence-internal patterns for NE recognition while simply discarding more difficult sentences. The initial corpus produced by An et al. (2003) was much larger than available Korean annotated corpora, and produced marginally improved results in an NER task.

An alternative presented by Richman and Schone (2008) takes advantage of structure of Wikipedia to produce NE-annotated corpora in languages other than English. Their method involves classifying Wikipedia articles and annotating linked texts with the entity class of their targets, and is discussed in detail in section 2.3.2. When training a NER system with thousands of automatically-annotated Wikipedia articles, their NER performance was comparable to being trained on 15–40,000 words of manually-annotated corpora.

Such approaches to automatic corpus acquisition present free data sources from which NER knowledge can be learnt. Automatic annotation processes produce a lower quality of data than manual annotation, but may be more flexible in producing corpora suitable for domain or language-specific tasks without expert annotators, and can produce much larger training texts than are otherwise available.

2.3 Wikipedia and named entities

Wikipedia is a multilingual online encyclopedia written by many thousands of its users. While Wikipedia's dubious verity is pertinent for some information extraction tasks, the sheer size and open-domain coverage of such a partially-structured natural language corpus makes it attractive for many computational linguistics tasks that often are not concerned with factual accuracy. Wikipedia's freely downloadable data[9] is now the primary data source in the INEX Workshop for XML information retrieval (Denoyer and Gallinari, 2006), for instance. The term *Wikipedia mining* refers to approaches for obtaining useful knowledge from Wikipedia data.

The Wikipedia corpus is especially useful for NE-related tasks. Most famous entities are allocated an article of their own, and conversely a large proportion of Wikipedia articles detail named entities. When an entity A is mentioned in an article B, it will usually hyperlink to the article about A, thus disambiguating the term in its context (Appendix A outlines Wikipedia features). Articles, and thus entities, are usually marked with one or more categories to which they belong, and articles of common-class entities (such as actors or countries) often include *infobox* templates listing critical details of the entity. When different entities share a name, the name often designates a *disambiguation page*, which lists the alternative meanings of a term. The literature describes many mining tasks using Wikipedia's semi-structured collection of information about entities, such as entity name classification and disambiguation or ontology construction.

[9] From `http://download.wikimedia.org/`

Ian Fleming (18): 1908 births, 1964 deaths, Bibliophiles, Black Watch officers, British spies, Deaths by myocardial infarction, English children's writers, English novelists, English short story writers, English spy fiction writers, English thriller writers, Old Etonians, Old Sunningdalians, People from Mayfair, Royal Navy officers, Royal Navy personnel of World War II, Sandhurst graduates, World War II spies.
James Bond (14): Characters in written fiction, Fictional English people, Fictional Old Etonians, Fictional assassins, Fictional bodyguards, Fictional detectives, Fictional gamblers, Fictional golfers, Fictional martial artists, Fictional secret agents and spies, Fictional socialites, James Bond, Media franchises, The League of Extraordinary Gentlemen characters.
Sydney (10): Australian capital cities, Cities in New South Wales, Coastal cities in Australia, Host cities of the Commonwealth Games, Host cities of the Summer Olympic Games, Metropolitan areas of Australia, Port cities in Australia, Settlements established in 1788, Spoken articles, Sydney.
University of Sydney (6): Educational institutions established in 1850, Gothic Revival architecture in Sydney, New South Wales Government statutory bodies, Universities in Sydney, University of Sydney, Worldwide Universities Network.

FIGURE 2.2: The diverse Wikipedia categories for a small selection of articles

2.3.1 Article classification

Although Wikipedia is structured around articles referring to entities, they can be enhanced as a knowledge source when classified into ontological groupings, broadly corresponding to standard NE classes. A core collection of recent work therefore involves classifying Wikipedia's articles.

Wikipedia does include its own category hierarchy. It allows its authors to tag articles (and categories) as belonging to one or more categories, but these categories do not neatly correlate with ontological relationships. A number of publications have proposed using the distance measures between Wikipedia categories as markers of semantic similarity (Chernov et al., 2006; Thom et al., 2007; Zesch et al., 2007), but these tasks do not require *taxonomic* integrity. Others have identified Wikipedia's category graph as a *folksonomy*, comparable to other collaborative online document tagging (Strube and Ponzetto, 2006). They point out that the semantic relationship between an article and its categories is heterogeneous, as evident from the categorisations of four example articles given in Figure 2.2. Some articles are members of many more categories than others. The relationships described by category labels include the following:

is a: Sydney is a coastal city in Australia.

has a: Ian Fleming had a 1908 birth.

is part of: University of Sydney is part of Worldwide Universities Network.

is in topic: James Bond relates to the topic James Bond.

administrative metadata: Sydney's article is a spoken article.

Suchanek et al. (2007) hence divide the Wikipedia categories into conceptual, relational, thematic and administrative. They further suggest that the sets of administrative and relational (e.g. 1908 births) categories are closed classes and can be manually ignored for ontology construction. Conceptual categories, they suggest, almost always contain a plural head noun, such as novelists, characters, cities or institutions in the examples of Figure 2.2, which often indicate an entity's ontological groupings. As well as the diversity and uneven distribution of the Wikipedia category graph, ontological subtlety is also prone to cause error. For example, the James Bond article states that it is actually about the James Bond phenomenon or concept, rather than about the character which is allotted its own article (James Bond (character)). Therefore its many categories beginning with fictional are inaccurate, leaving Media franchises as its only accurate conceptual category. Although editorial projects exist to improve and systematise Wikipedia's category graph,[10] it currently consists of popular labelling that does not correspond to the ontological classification required for NLP applications.

Those seeking the classification of Wikipedia articles therefore resort to less direct methods of identifying an article's class. A few approaches avoid machine learning: both Toral and Muñoz (2006) and Suchanek et al. (2007) do so by utilising WordNet. Toral and Muñoz (2006) suggest that words in an article's opening sentence will often come from an entity class-related synset, and so are able to identify a large portion of articles about locations and people, but organisations to a much lesser extent. Suchanek et al. (2007) instead rely on Wikipedia's category system, using the head nouns of conceptual categories as keys to associate articles with WordNet synsets. For example, Ian Fleming's categorisations as a spy, novelist and bibliophile would all ultimately lead to the WordNet concept for a human being. Although Suchanek et al. (2007) and our own research both suggest that the category *hierarchy* is highly unreliable apart from leaf nodes (i.e. articles' direct memberships in categories), Richman and Schone (2008) use a limited list of key phrases to identify categories in each class, and if none is matched by a leaf category, their system checks parent categories recursively until a sufficient number of categories agreed in support of a particular classification. They also consider the listing of a Wikipedia title as a common noun in Wiktionary[11] as evidence that the title refers to a non-entity. Some initial attempts by Bhole et al. (2007) also showed limited success in using a heuristic approach alone. They first attempted the low-recall method of extracting entities from Wikipedia's *list pages* (e.g. List of biologists), and eventually used rules determined by the presence of certain infobox templates, of dates of birth, or of geographical coordinates. This again gave very poor results for the ORG category (10%), and only 49%

[10] See `http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Categories`, `http://en.wikipedia.org/wiki/Wikipedia:Categorization`

[11] A collaborative dictionary project, `http://en.wiktionary.org`.

Feature space	SVM	NB
Whole article	**89.7**	86.9
Outgoing links	84.0	82.5
Structured data	88.1	84.8
First paragraph	86.0	**88.4**

TABLE 2.4: Reported average F measures when classifying (using SVM and Naïve Bayes) Wikipedia articles with bag-of-words features from different portions of the article text. (Dakka and Cucerzan, 2008)

for LOC, although at almost-perfect precision for all classes. For all of these approaches, if insufficient information was available, an article would simply be classified as belonging to no class.

Just as in heuristic-based approaches, machine learning methods utilise various features of Wikipedia articles. Bhole et al. (2007) eventually used a bag-of-words classification method to moderate success.[12] More of Wikipedia's structural features were utilised by Watanabe et al. (2007), who used Conditional Random Fields (CRF) to explore list structure, with the understanding that entities together in a list are commonly of the same category. They train both this CRF-based system and another SVM with features such as article texts, headings, categories and the text of incoming links, and show that list structure can give a marginal improvement over a text-only approach.

Dakka and Cucerzan (2008) achieve a 90% F measure (90% P, 89% R) when classifying of a set of articles into one of PER, MISC, ORG, LOC and common (non-entity) using SVM with a bag-of-words approach over whole article texts. This exceeds results obtained utilising a more selective feature set, such as the words appearing in structured data (section titles, infoboxes, etc.), in an article's outgoing links, or in a context window surrounding incoming links (see Table 2.4). Using Naïve Bayes over the words in only the first paragraph of each article they achieve a 88% F measure. Non-entity articles are the most difficult to classify correctly (F 88% for full-text SVM; LOC and PER reached 95%), due to their heterogeneity. The training and test data for their classifiers were initially compiled by human-judgement of 800 random articles, and then expanded in the same class distribution by bootstrapping over the co-presence of entities in lists (as in Watanabe et al. (2007)). They do not report the size of their expanded data set, although it presumably has some bias towards entities that are popular enough to be mentioned in lists. These results represent the state-of-the-art in classifying Wikipedia articles into named entity categories, although we present a competitive alternative method in Chapter 5.

[12]The reported results are difficult to interpret summarily.

2.3.2 NER with Wikipedia data

Wikipedia's content has been mined for NER in a few ways: (a) to provide name gazetteers; (b) to provide knowledge for a NER feature; and (c) to derive training data. On the basis of article classification methods discussed in the previous subsection, the creation of categorised lists of entity names is trivial. Toral and Muñoz (2006) suggest the use of such gazetteers in NER, and although this could increase recall of famous named entities when compared with traditional entity extraction methods (e.g. Etzioni et al. (2005)), this approach includes the failings usually associated with gazetteer use in NER (see section 2.2.1).

Wikipedia's knowledge may be incorporated directly as features in NER learning rather than requiring the intermediary of classification which is prone to error. Kazama and Torisawa (2007) choose a single Wikipedia-based feature, and use it to augment a complete NER system. They note, as Toral and Muñoz do, that most articles begin with a definitional statement such as Thomas Cruise Mapother ... is an American actor and film producer. Within the first sentence of each article they search for a copula (is, are, etc.) and select the first head noun that follows it, actor in the case of Tom Cruise. Where available, this one-word *category label*, they suggest, will be sufficient to predict the entity's membership of a particular class (much like the conceptual category head nouns used by Suchanek et al. (2007)). They therefore add this label as a feature to a CRF-based named entity recogniser: each appearance of an unambiguous Wikipedia title in the training or test data is marked with its category label. In training, the machine learner is able to learn an association between gold-standard entity class annotations and Wikipedia-based labels, although its effectiveness is constrained by the training corpora available: any category labels which do not appear in training are unusable when tagging unseen text. This method improves their baseline performance on the English CONLL-2003 test data by 1.6% F score. The inability to label ambiguous names seems to be a cause for finding that their Wikipedia feature and standard gazetteers had significantly different coverage. Using gazetteer and Wikipedia labels together, they achieved 3.0% F score above their baseline (88.0 from 85.0), placing their results higher than all but two of the CONLL-2003 entrants. Because they do not directly classify Wikipedia articles, Kazama and Torisawa's (2007) methodology is equally applicable to finer-grained named entity hierarchies. They also consider as future work an entity-disambiguation method for correctly labelling ambiguous entities found in text, which would certainly increase the applicability of their system, but overall they provide an interesting and simple approach to improving NER using Wikipedia features.

A third approach to improving NER with Wikipedia involves producing annotated training data and is the goal of my thesis work. Richman and Schone (2008) transfer English Wikipedia article classification knowledge to non-English articles and categories by way of Wikipedia's inter-language links (see section A.2). Selecting over 50,000 articles from each target language's Wikipedia, they form a NE-annotated corpus where each link is transformed into an entity annotation on the basis of the link target's class. Additional entities are marked up if they: (a) are mentioned elsewhere in the article; (b) are mentioned in corresponding English Wikipedia articles; (c) correspond to any Wikipedia titles in the target language; (d) match abbreviation forms found parenthesised after organisation names in the article; (e) have the form of an initialised personal name. Like MUC evaluations, Richman and Schone (2008) also mark up dates and numerical information by way of regular expressions specific to the target language. Comparing the performance of a Wikipedia-trained system to human-annotated training data in Spanish, French and Ukrainian they report that their Wikipedia corpora corresponded in performance to 15–40,000 words of gold-standard annotated data depending on language.[13]

2.3.3 Entity disambiguation

Instead of predicting the entity class of names in text as in NER, a system may attempt to identify each name's particular referent, represented by its corresponding Wikipedia article.

Wikipedia provides a mechanism for determining the series of candidate entities with a given name. Alternative names (synonyms) for many entities are covered by *redirect* pages that forward a viewer to the appropriate article (see appendix section A.1), while ambiguous entity names (homonyms, including some acronyms and abbreviations) often are associated with a disambiguation page. For instance, the disambiguation article on Bush includes links to common noun meanings like flora, fauna (bush babies), undeveloped land, survival skills, mechanical fixings; many people with Bush as a surname; a few place names; some organisations (including two bands); and an aircraft carrier named after a former US president. This is a substantial listing of entities that may be referred to by that title, though necessarily incomplete and giving little designation of which are its most common uses. But given candidate entities corresponding to a particular term, systems have been built that attempt to identify the particular referent, or to determine that none can be found in Wikipedia.

To limit this disambiguation to articles about proper-noun entities only, Bunescu and Paşca (2006) used capitalisation-based heuristics to select NE-related articles. Then, given a candidate named entity found

[13] They do not report the number of words in Wikipedia-derived corpora resulting from their method.

in a text, a context window of 55 words was compared using a bag-of-words SVM classifier to candidate article texts. This method suffered due to short articles and the use of related but different words in the two texts being compared. In response they expanded the set of words used for comparison to include those common to other articles in the same categories as the candidate entity; the target sentence John Williams *lost* a Taipei death *match* would readily match a wrestler, while John Williams *conducted* a summer Star Wars *concert* has keywords in common with other composers.

Mihalcea and Csomai (2007) also used a simple word-overlap method between target contexts and candidate articles (including non-entity keywords), although they augmented this with a naive Bayes classifier trained on the contexts of incoming links to those articles. For example, if the John Williams (composer) article was linked to elsewhere in Wikipedia in a sentence that mentioned conducting, this correlation would improve the chances of correct disambiguation when John Williams and conducted are found together in a target text.

The approach described in Cucerzan (2007) does not use the entire article text, and instead splits identifying features into three groups: surface forms, categories and context. Surface forms are essentially a collection of the strings by which a given entity is referred to in Wikipedia, collected from the article title, redirect titles, disambiguation pages that point to that article, and the various anchor texts used for incoming links (with frequency ≥ 2). Context extends this with all the articles that are linked from or link to the candidate entity, while category labels include both the Wikipedia categorisation, as well as the titles of lists in which the entity appears, with significant filtering applied. Rather than merely checking for string matches to surface forms to find entity strings in a target text, Cucerzan uses a truecaser, a named entity recogniser, some web-based heuristics, and in-document coreference resolution to identify the boundaries of entity mentions in text. With context and category labels used as binary features, a vector-based comparison between target context and candidate entity vectors determines the best match. If multiple matches are still found, the target context is reduced from document to paragraph to sentence levels for more refined matching.

Each of these systems was able to evaluate its performance by taking a sample of Wikipedia documents, removing the links, and trying to redetermine them correctly. Nonetheless, it is difficult to compare the results of each disambiguation approach, partially because they conflate their results for unambiguous cases with those for ambiguous terms, in addition to each system being tested on different documents and domains, and not having a standard definition of a correct match. This method of evaluation also does not judge the system's performance for *unknown* entities, such as a non-notable individual by the

name John Williams, although Bunescu and Paşca (2006) did evaluate this aspect by treating individual instances as if they were not in Wikipedia and scoring correct results if their system correctly identified this fact. Moreover, both Cucerzan (2007) and Mihalcea and Csomai (2007) found instances where their system could more accurately disambiguate a link than that given by Wikipedia's authors, highlighting Wikipedia's failings as "correct" data for evaluation.

2.3.4 Other Wikipedia entity mining

As discussed above, Wikipedia is an obvious resource for NLP tasks that deal with entities, which further recent IE and IR work has attempted to capitalise upon.

Wikipedia has readily been associated with notions of automatically constructing a *Semantic Web* (Berners-Lee et al., 2001). This concept revolves around lists of machine-readable facts that describe attributes and relationships of entities. DBpedia (Auer et al., 2007) and Freebase (Bollacker et al., 2008) are both online resources with interfaces for application development consisting of basic factual data primarily extracted from Wikipedia's infoboxes and some category material (Auer and Lehmann, 2007), and additional relations that connect the Wikipedia data to other structured data-sources on the web, such as IMDb[14]. Wu et al. (2008) provide a method for extending knowledge bases derived from infoboxes to articles without them.

A similar use of Wikipedia entity data is in the creation of ontologies. One approach involves mapping Wikipedia to an existing hierarchical ontology, be it WordNet (Suchanek et al., 2007; Toral et al., 2008) or Cyc (Medelyan and Legg, 2008). The alternative is to create ontologies directly from Wikipedia data (Ponzetto and Strube, 2007; Nastase and Strube, 2008; Cui et al., 2008). Most of these approaches utilise the Wikipedia category graph, noting useful patterns in the naming of category titles (Ponzetto and Strube, 2007; Suchanek et al., 2007; Nastase and Strube, 2008). Cui et al.'s (2008) work creates domain-specific ontologies and compares results to the Library of Congress Classification scheme.

Other projects used a more text-based approach to extract facts and relations between entities. Ruiz-Casado et al. (2005), for instance, used seeded pattern learning, while Bhole et al. (2007) scoured entity articles for descriptions of other entities and events, in order to construct timeline data.

Wikipedia has also been used in Information Retrieval, where the task of entity ranking (finding the most relevant entities for a given query) has been prominent. Current attempts have explored the usefulness

[14] `http://www.imdb.com/`

of Wikipedia's link graph for this purpose (Zaragoza et al., 2007b; Vercoustre et al., 2007; Ollivier and Senellar, 2007), using standard techniques such as PageRank (Brin and Page, 1998) or HITS (Kleinberg, 1999), although Vercoustre et al. (2007) also utilised Wikipedia category associations. Ollivier and Senellar (2007) frame their work assessing the use of Green measures for this task as an "illustration with Wikipedia", an example of Wikipedia being used as an experimental corpus for work elsewhere. Clearly the dense linking between entity-related articles on Wikipedia make it ideal for such network-based approaches.

With continued exploration of named entities in Wikipedia, techniques are becoming more refined for accessing Wikipedia's content and structural features as resources for entity classification, disambiguation, relation and ranking.

2.4 Further NLP applications of Wikipedia

To complete our survey of Wikipedia mining literature, we describe another few areas where Wikipedia's knowledge has been utilised. This includes its used as a corpus of world knowledge to aid question answering (Ahn et al., 2004; Lita et al., 2004); or using its interconnections to collect lexical semantic data (Strube and Ponzetto, 2006; Zesch et al., 2007; Milne, 2007). A converse approach considers methods of improving Wikipedia's structured knowledge by taking advantage of less explicit structure. For example, Sorg and Cimiano (2008) attempt to find pairs of corresponding articles in Wikipedia so as to predict new inter-language links; Wu et al. (2008) extracts data with which to expand incomplete or non-existent infoboxes. In subsections below we discuss some other general applications of Wikipedia's knowledge as training data or as a knowledge source for describing arbitrary documents.

Perhaps the largest area of Wikipedia research not yet covered here is that of using Wikipedia as an aid in classifying or labelling arbitrary documents. Gabrilovich and Markovitch (2006) finds Wikipedia features especially useful for the classification of very short documents which traditional classifiers are particularly challenged by. Like Kazama and Torisawa's (2007) use of Wikipedia-derived category labels as NER features, they use an auxiliary classifier to identify those articles in Wikipedia that are most relevant to a document or a paragraph, and then uses those article titles as features for document classification. A number of others use Wikipedia to determine document topics in a similar manner, labelling documents or collections of documents with appropriate Wikipedia articles or categories (Schönhofen,

2006; Medelyan et al., 2008; Syed et al., 2008), and some follow Gabrilovich and Markovitch (2006) in using this labelling as a feature for further classification (Weale, 2006; Banerjee, 2007).

Wikipedia is also used as training knowledge for NLP tasks in addition to Richman and Schone's (2008) NER work described above (section 2.3.2). The fact that in languages such as Hebrew clitics are commonly unlinked when attached to link terms has promoted Wikipedia's use in word segmentation (Gabay et al., 2008). By correlating facts given in infoboxes to occurrences of the same information within article text, Wu et al. (2008) are able to learn patterns for information extraction. For the task of summarising multiple sources into biographies, Biadsy et al. (2008) build a system which learns from biographical Wikipedia articles to appropriately order summary sentences. Nelken and Yamangil (2008) propose the novel exploitation of Wikipedia articles' revision histories, which store every edit that is made to an article. They exhibit its use to find certain types of spelling errors or to train a system for sentence compression. They further propose potential uses of revision data in text summarisation or anaphora resolution.

Note also that research regarding Wikipedia extends beyond language applications. Wilkinson and Huberman (2007) explores the rate at which articles are edited, finding that edits cause other edits, and that the number of edits to an article tends to correlate with its quality; Thomas and Sheth's (2007) research suggests that Wikipedia articles converge to stability; other researchers have performed graph-theoretic analyses of Wikipedia's category and link networks (Muchnik et al., 2007; Zesch et al., 2007).

2.5 Summary

The task of Named Entity Recognition has received much attention since its advent in the 1990s, being a central component of many NLP systems, particularly those related to information extraction. The task involves challenging semantic ambiguities, and can be benefited by the provision of both linguistic and world knowledge sources to adaptable machine learning systems. Due to its size, structure, open licencing and continuous updating, Wikipedia has recently become a popular source of world knowledge for NLP tasks, especially those involving named entities. The application of Wikipedia to improving NER has been introduced in the literature (Kazama and Torisawa, 2007; Richman and Schone, 2008), and we present a new method of extracting sentences with which to train statistical NER systems in Chapter 4.

CHAPTER 3

Evaluating Training Corpora

Most of the literature on statistical NER considers the impact of features on performance given a fixed training corpus and corresponding test data. Our experiments with new corpora involve varying training data, while keeping the modelling system fixed. This unusual experimental variable introduces many complicating factors such as:

- the choice of other corpora as baselines;
- formatting training corpora for fair comparison;
- analysis of internal and inter-corpus inconsistency; and
- the effect of mismatched data sources on inter-corpus results.

In this chapter, we introduce our method for evaluating and comparing gold-standard NE-annotated data, which will be reapplied to analyse Wikipedia-derived corpora in Chapters 6 and 7.

3.1 Gold-standard corpora

We evaluate our generated corpora against three sets of manually-annotated data from (a) the MUC-7 Named Entity Task (MUC, 2001a); (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). Basic details of each corpus are given in Table 3.1 with more statistics in Table C.1.

Sources. The text of each corpus is extracted from a different news source text, with standardised divisions into training (TRAIN), development test (DEV) and final test (TEST) data available for MUC and CoNLL corpora. TRAIN data are used to train the NER system, DEV is used while altering experimental features and for error analysis, and TEST is used for final evaluation. The DEV-TEST separation avoids a system fitting a particular evaluation corpus too closely. MUC consists of New York Times News

Corpus	Source	# ent. tags	Number of tokens		
			TRAIN	DEV	TEST
MUC-7	New York Times	3	83,601	18,655	60,436
CONLL-03	Reuters	4	203,621	51,362	46,435
BBN	Wall Street Journal	54	901,894	142,218	129,654

TABLE 3.1: Summary of the corpora used for comparison and evaluation, including the number of relevant entity types marked up and corpus sizes

BBN class prefix	CONLL class	BBN class prefix	CONLL class
EVENT:	MISC	LANGUAGE	MISC
FAC:AIRPORT	ORG	LAW	MISC
FAC:ATTRACTION	ORG	LOCATION:	LOC
FAC:BRIDGE	LOC	NORP	MISC
FAC:BUILDING	LOC	ORGANIZATION:	ORG
FAC:HIGHWAY_STREET	LOC	PERSON	PER
FAC:HOTEL	ORG	PRODUCT:	MISC
FAC:OTHER	LOC	WORK_OF_ART:	MISC
GPE:	LOC		

TABLE 3.2: Our mapping from BBN to CONLL named entity classes. Hierarchy is mapped by matching label prefixes. All entities labels not matched are removed.

Service articles which were topically selected to related to aeroplane crashes for TRAIN and DEV data, and launch events for TEST. CONLL was taken from the Reuters Corpus (Rose et al., 2002) with TRAIN and DEV data covering ten days in August 1996 and TEST data from December 1996. The BBN corpus is the Penn Treebank (Marcus et al., 1995), 1.1 million words of 1989 Wall Street Journal material, split into sections. We use 03–21 for TRAIN, 00–02 for DEV and 22-24 for TEST.

Tags. Each corpus uses a different set of entity labels. MUC marks locations, organisations and personal names as well as numerical and time data (Chinchor and Robinson, 1997). CONLL (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) marks person (PER), organisation (ORG), location (LOC) and a broad miscellaneous class (MISC; e.g. events, artworks and nationalities). BBN use a hierarchy of 105 fine-grained tags (Brunstein, 2002): 54 corresponding to CONLL entities; 21 for numerical and time data; and 30 others (animals, chemicals, diseases, etc.).

In order to compare training corpora we consider the four-class CONLL tag set, and map BBN classes down to these coarse groupings as shown in Table 3.2; non-entity tags in the BBN and MUC data are removed. Since no MISC entities are marked in MUC, we perform all evaluations twice: one model

includes MISC tags and is evaluated on CONLL and BBN data; the other leaves MISC entities unlabelled, and is evaluated across all three gold-standard corpora.

Tokenisation. Unlike the CONLL and BBN corpora, MUC data is marked up using SGML and so is not ordinarily distributed as tokenised data. Official tokenisation rules are provided to define positions where it is valid to annotate entity boundaries in addition to whitespace (MUC, 2001b). We initially used a MUC corpus tokenised precisely according to these rules, but found that discrepancies with the other corpora caused very poor inter-corpus performance. In particular, all trailing full-stops were separated from abbreviations, while in BBN and CONLL they are usually stripped only at the ends of sentences. Rejoining all full-stops to abbreviations mid-sentence caused greatly improved performance of BBN-trained models on MUC (up to 12% EXACT F-score). We will generally report results from this corpus with modified tokenisation, although in this chapter we refer to the MUC corpus with unadjusted tokenisation as MUC*. Note that there are other tokenisation discrepancies which are not as easily fixed (c.f. section 3.4), such as the splitting of hyphenated terms, where London-based is marked up as [LOC London] - based in MUC and [LOC London-based] in BBN.

Quotes. The CONLL data does not distinguish between left and right quotation marks, so this distinction is removed from the other corpora.

3.2 Training the C&C tagger

We normalise the format of each candidate training corpus in order to train the C&C NER system (Curran and Clark, 2003b). The system by default uses a pipe-delimited format to annotate each token with a part of speech (POS) tag and a NE tag given in IOB format. For example, [PER Murdoch]'s [ORG News Corp] would be formatted as:

Murdoch|NNP|I-PER 's|POS|O News|NNP|I-ORG Corp|NNP|I-ORG

where NNP is a POS tag indicating a singular proper noun, and POS indicates a possessive.

We POS tag training corpora using the C&C tools (Curran et al., 2007) tagger trained on the Penn Treebank (Marcus et al., 1995), ignoring the manual POS annotations available with the BBN corpus (also from the Penn Treebank). While we do this in order to fairly compare the corpora, POS tagger accuracy is likely to be higher on the BBN corpus, having been trained on the same text. This situation

is not ideal and the impact of this POS tagging bias may be evaluated or avoided in future work with a POS tagger trained on another corpus.

IOB is a representation of non-overlapping, non-nested sequences in the annotation of individual tokens, suitable for machine learning (Ramshaw and Marcus, 1995). The original definition of IOB (later designated IOB1) labelled each token with with one of:

O: for tokens not in any phrase;
I-X: for tokens in a phrase of type X; or
B-X: for the first token in a phrase of type X immediately following another phrase of type X.

IOB2 uses B-X to mark the beginning of all phrases rather than just neighbouring phrases of the same type (Tjong Kim Sang, 2002). Although CoNLL-2002 and BBN use IOB2, we follow CoNLL-2003 in annotating entities with IOB1 for the present evaluation.[1]

With training corpora in a uniform format, they are each used to train the C&C maximum entropy tagger for NER (Curran and Clark, 2003b), with its default use of local, global and external knowledge features as detailed in Appendix B. The tagging model produced is then used to predict tags on DEV or TEST data, which are then compared to gold-standard annotations.

3.3 Comparing corpora

Typically NER performance is measured by evaluating a model on a designated TEST corpus, comparing tag predictions to those annotated in the gold-standard data, and then applying some metrics (discussed in section 2.1.1). In considering the training corpus as a variable in NER, this is again the most obvious approach to comparison, although it is also possible to study the systematic errors produced by the predictive model in terms of differences between the TRAIN and TEST corpora. In analysing a corpus, it may be considered either with regard to itself or in comparison to other corpora; also, we may analyse the training annotations, or the model built on those annotations and its predictive performance. We have developed three methods to explore intra- and inter-corpus consistency which we apply to the gold-standard corpora in section 3.4.

[1] While the choice between IOB1 and IOB2 impacts performance, for the present work it is only necessary to fix one method as an experimental constant.

3.3.1 N-gram tag variation

Dickinson and Meurers (2003) present a method for finding inconsistencies within POS-annotated corpora, which we apply to NER corpora. Their method finds all n-grams in a corpus for which a subsequence (the *nucleus*) has tags which vary between occurrences of that n-gram. Multiple tag sequences for the same text within a corpus are often valid, as the correct tag depends on context. To remove such valid ambiguity, they suggest using (a) a minimum n-gram length; (b) a minimum margin of invariant terms around the nucleus, except at sentence boundaries. For our analysis, we require at least one invariant token (or boundary) on either side of the nucleus. To simplify results, they suggest including each originating context only once in results, subsuming shorter n-grams in longer ones.

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. Such variations are likely to be annotation inconsistencies or errors.

We do not currently have a metric that summarises and compares results of this analysis, although an information-theoretic measure may be of use to measure the level of variability for n-grams. Because it requires sufficient repetition to be useful, the analysis is only applicable to large corpora. The increase in inconsistencies when combining multiple corpora for analysis may indicate inter-corpus disagreement, but we have not yet developed this approach.

3.3.2 Entity type frequency

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

POS sequences: Entity types may often be distinguished by their POS tags, e.g. nationalities are often JJ (adjectives) or NNPS (plural proper nouns).

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. We distinguish single from multiple characters by duplication, e.g. USS Nimitz (CVN-68) has wordtype AA Aaa (AA-00).

Wordtype with function words: We map content words only to wordtypes; function words are retained,[2] e.g. Bank of New England Corp. maps to Aaa of Aaa Aaa Aaa..

Alone, no grouping provides sufficient discrimination: wordtype patterns are able to distinguish among terms with 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 with this orthographic form. Instead, CONLL has many names of form A. Aaa, e.g. S. Waugh, while BBN and MUC have none. We can therefore predict possible incompatibilities between systems trained on BBN and evaluated on CONLL or vice-versa.

3.3.3 Tag sequence confusion

The previous two analyses consider the training corpus. However, it may be more meaningful to consider the performance of the predictive model. In classification tasks, it is common to analyse errors by constructing a confusion matrix between predicted and correct classes. This does not apply to phrasal sequence tagging, where class confusion is only applicable either to entities whose boundaries are identified precisely by the tagger, or on a per-token basis, losing any sense of phrasal matches.[3]

For phrasal tagging tasks such as NER, it is more helpful to consider a matrix which illustrates both boundary and class confusion. We therefore compile two matrices to analyse confusion in tagger output:

CORR-PRED**:** comparing correct entity classes against predicted tag sequences; and
PRED-CORR**:** comparing predicted classes against correct tag sequences.

If oversized boundaries are predicted by the tagger, matrix 3.3.3 considers them correct matches, and tabulates cases of undersized boundaries. For example, if [ORG Johnson and Johnson] was tagged [PER Johnson] and [PER Johnson], it is marked in the matrix cell corresponding to correct tag ORG and predicted PER O PER.[4] PRED-CORR does the opposite: if gold-standard Mr. [PER Ross] is tagged PER, it is counted as confusion between predicted tag PER and correct tag sequence O PER. See Tables C.8 and

[2] We use a list of 129 prepositions, conjunctions, determiners and pronouns from `http://www.marlodge.supanet.com/museum/funcword.html`.

[3] These are not useless measures either, but an indication of phrasal confusion is more meaningful.

[4] Here we use O to refer to a string of tokens without any entity annotations.

C.9 for an example. To further distinguish types of error, entity groupings from section 3.3.2 may also be used.

This method is useful for both analysing the performance of the tagger, and for cross-corpus evaluation, e.g. BBN versus CONLL on a BBN test set. As with entity type frequency analysis, this cross-corpus analysis involves finding confusion matrix entries where BBN and CONLL's trained models differ significantly in prediction. Unlike comparing training corpora directly, this method identifies the impact of inter-corpus variation on the tagging model and its performance.

3.4 Evaluation of gold-standard corpora

We have compared the NER performance of gold-standard training corpora. In Table 3.3 we present the DEV[5] results for training and testing on each manually-annotated corpus, including two variant tokenisations of the MUC data, according to a number of the evaluation metrics given in section 2.1.1. For each testing corpus considered, the performance of training data from the same corpus is higher than other training data, by up to 32% EXACT F-score between evaluating MUC and CONLL models on CONLL DEV. The smallest cross-corpus variance is 2% EXACT F-score between BBN and MUC on MUC DEV. Note that this corresponds with the fact that the MUC (and MUC*) corpus performs relatively poorly on its associated DEV data, presumably because the training corpus is simply too small, whereas BBN is the largest of the three human-annotated corpora. The table also indicates that MUC and BBN data have much greater compatibility with each other than with CONLL. The large impact that the choice of training corpus has on NER performance was also reported by Ciaramita and Altun (2005) who manually annotated section 00 of the Penn TreeBank's Wall Street Journal text with CONLL-style labels and reported a 26.5% decrease in F-score when testing a CONLL-trained tagger on the WSJ text as compared to CONLL data.

To analyse this corpus incompatibility, we apply the methods described in the previous section.

Table 3.4 lists some n-gram tag variations for BBN and CONLL (TRAIN + DEV). These include cases of schematic ambiguities (e.g. the full-stop 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 provides only a few

[5]The choice of DEV rather than TEST corpora here is somewhat arbitrary, and TEST data shows similar trends, but DEV data will initially be considered in our experiments with Wikipedia-derived corpora below in Chapters 6 and 7. Test performance for gold-standard corpora is indicated in Table C.5.

TRAIN	DEV	With MISC	TEXT F	TYPE F	MUCEVAL P	MUCEVAL R	MUCEVAL F	EXACT P	EXACT R	EXACT F
MUC*	MUC*		91.3	87.5	88.9	89.9	89.4	83.0	83.9	83.4
MUC	MUC*		85.5	82.7	82.2	86.1	84.1	74.1	77.6	75.8
CONLL	MUC*		80.3	74.7	75.8	79.4	77.5	63.0	66.0	64.5
BBN	MUC*		85.7	82.0	83.6	84.0	83.8	75.0	75.3	75.1
MUC*	MUC		90.4	86.1	88.2	88.3	88.2	82.2	82.3	82.2
MUC	MUC		91.7	86.3	87.7	90.4	89.0	81.0	83.6	82.3
CONLL	MUC		86.2	76.6	81.2	81.6	81.4	69.7	70.1	69.9
BBN	MUC		89.9	85.3	87.5	87.7	87.6	80.1	80.2	80.2
MUC*	CONLL		73.4	59.1	74.1	59.9	66.2	61.3	49.5	54.8
MUC	CONLL		76.8	59.6	71.0	65.5	68.2	57.2	52.8	54.9
CONLL	CONLL		92.0	89.7	89.0	92.8	90.8	85.2	88.8	86.9
BBN	CONLL		74.9	68.6	71.7	71.8	71.8	59.0	59.0	59.0
MUC*	BBN		76.6	68.3	73.7	71.3	72.5	60.7	58.7	59.7
MUC	BBN		82.8	75.6	77.5	81.0	79.2	67.8	70.9	69.3
CONLL	BBN		75.2	70.0	69.4	76.1	72.6	57.6	63.1	60.2
BBN	BBN		92.1	90.9	89.4	93.7	91.5	86.0	90.1	88.0
CONLL	CONLL	✓	92.5	89.2	90.3	91.5	90.9	85.3	86.5	85.9
BBN	CONLL	✓	77.0	68.4	75.8	69.8	72.7	62.0	57.1	59.4
CONLL	BBN	✓	80.1	70.0	73.1	77.2	75.1	60.2	63.6	61.9
BBN	BBN	✓	92.6	89.6	90.3	92.0	91.1	85.7	87.3	86.5

TABLE 3.3: Gold-standard cross-corpus performance (DEV)

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

TABLE 3.4: Examples of n-gram tag variations in BBN (top) and CONLL (bottom). Nuclei are in bold.

examples, echoing BBN in the ambiguities of trailing full-stops and leading determiners. Whether to include leading modifiers in entity tags is often ambiguous.

Wordtype distributions were also used to compare the three gold-standards. We investigated all word-types which occur with at least twice the relative frequency for each entity class in one corpus as in

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

TABLE 3.5: Examples of tag sequence confusion on BBN DEV when training on gold-standard corpora (no MISC).

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 headline capitalisation.
- when abbreviations appear at the end of a sentence, the '.' token is separated from the text. This full-stop is commonly included within tags in BBN and MUC, but less often in CONLL.
- BBN tags text like Munich-based as LOC; CONLL tags it as MISC; MUC separates the hyphen as a token.
- CONLL is sports biased and has many event names in the form of 1990 World Cup.
- CONLL marks currencies such as A$ while MUC and BBN mark them only in the context of amounts.
- BBN separates organisation names from their products as in [ORG Commodore] [MISC 64].
- BBN makes many more mentions of full company names.
- BBN commonly uses R. and D. as abbreviations for Republican and Democrat.
- Since BBN also annotates common nouns, some have been mistakenly labelled as proper-noun entities.
- 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. Table 3.5 shows some examples where the resulting 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. A reasonable automatically-annotated corpus may likewise have relatively poor performance on gold-standard data.

CHAPTER 4

Method

4.1 Approach

The size and structure of Wikipedia suggests its use as a source of data for knowledge-based tasks. We use the fact that many links between Wikipedia articles correspond to standard entity annotations to add NE-annotations to Wikipedia text. Wikipedia links introduce new topics or entities within an article and provide a reference for encyclopaedic definition of the topic. We therefore make the following assumption: *If an article A is about an entity of class K then all links to article A correspond to entity annotations of class K.*

On the basis of this assumption, we derive NE-annotated text from Wikipedia by the following steps:

(1) Classify all articles into entity classes (Chapter 5).
(2) Clean Wikipedia articles and split them into sentences with links (section 4.2).
(3) Annotate sentences with NEs according to link targets (section 4.3).
(4) Select sentences for inclusion in a corpus (section 4.3).

Figure 4.1 illustrates the application of our approach to a single example sentence, Holden is an Australian automaker based in Port Melbourne, Victoria. The proper-noun terms Holden, Australian and Port Melbourne, Victoria each link to appropriate articles which, when classified as describing an organisation or a location, allow us to label the links in the original sentence as named entities of these classes. In order for the resulting annotations conform to the schema of standard NE-annotated corpora, the resulting annotations may need to be adjusted (see section 6.3.2).

Figure 4.2 provides more detail into each of the processes used in transforming Wikipedia into a NE-annotated corpus. The only manually-built data source required is a collection of articles with entity class labels assigned, used to seed the classification process (see section 5.3).

FIGURE 4.1: Tracing an example through our system, from Wikipedia page to NE-annotated sentence

4.1.1 Key ideas: links as annotations

Key to our approach of using Wikipedia links as NE annotations are the following two notions:

Link targets tell us more than context. In statistical NER, a system is usually only able to decide the boundaries and class of an entity on the basis of prior knowledge and context. Since Wikipedia links usually target structured articles whose topic is the link text, the target article provides much more information on the entity class than a computerised system can easily infer from the context of a reference.

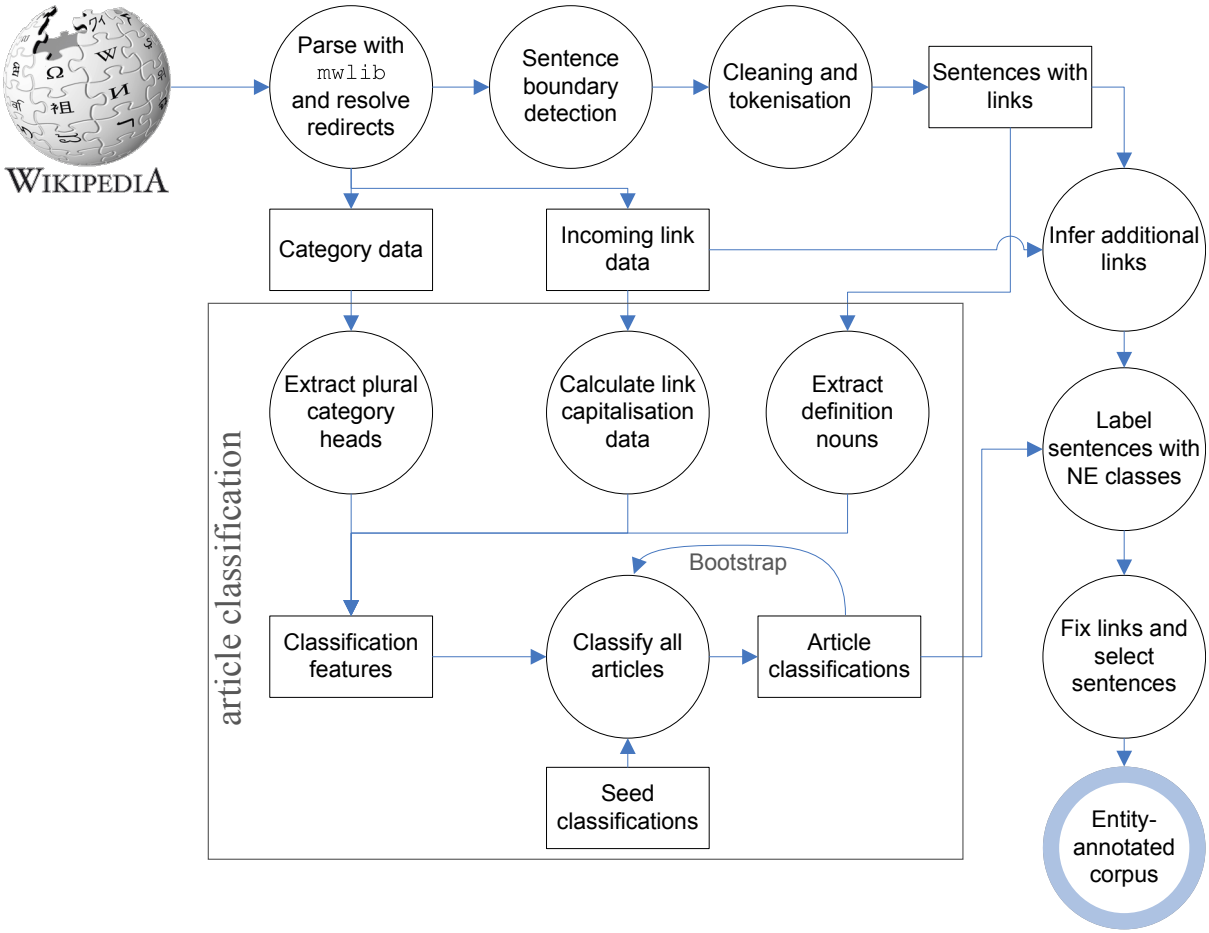


FIGURE 4.2: Detailed data-flow for deriving NE-annotated data from Wikipedia

Links disambiguate their referents. An important challenge of NER is the ambiguity of many names. Wikipedia encourages editors to choose the most appropriate target article for a particular link, and facilitates this task with disambiguation pages. For example, one may link the text David Jones to David Jones Limited or to David Jones (poet). If such linked text is correctly transformed to an entity annotation, a machine learner may learn to appropriately differentiate mentions of David Jones as a person from the company.

4.1.2 Caveats to our assumption

Although we rely on the assumption given above, in practice it does not always hold. Though concern regarding Wikipedia's *factual* quality is not entirely pertinent to our purpose, the online encyclopaedia includes errors in linking which break our assumption, and great stylistic variability when compared to

Referent	Frequency	Example
Planet	91	Shaneeswaran is the Hindu lord of Saturn.
Gaming console	2	. . . given the rise of Playstation and Saturn in that year. . .
Mythological figure	1	Equuleus is also linked to the story of Philyra and Saturn.

TABLE 4.1: Link disambiguation errors for links to the article Saturn—whose topic is the planet—appearing with link text Saturn. The table gives the number of links for each referent.

precise NE corpus annotation schemes (e.g. Chinchor et al. (1999)). This complicates direct translation from Wikipedia's linked sentences to an annotated corpus.

A few examples break the following corollary of our above assumption: *The logical referent of linked text in Wikipedia corresponds to the link's target article (or is of the same entity class).* The most obvious deviation from this corollary is error on the part of Wikipedia contributors. One class of such errors is *link disambiguation errors*: Wikipedia handles the ambiguity in article titles in one of two ways: the default page for a given term is either (a) a disambiguation page, which serves to list all possible referent articles (e.g. David Jones); or (b) the article for what Wikipedia editors have considered its most common referent (e.g. for Saturn, a planet), with an editorial note preceding the article that indicates alternative referents or points to the relevant disambiguation page. When composing articles, Wikipedia editors may lazily assume that the referent they intend by a particular term is the topic of the article with that name. Thus they might link to the article entitled David Jones when discussing Australian department stores, or Saturn in the context of Greek mythology. Both would be erroneous links, but errors of the former kind can be much more easily identified than the latter. Link disambiguation errors like the Saturn example may break our assumption (if a term has referents of different classes) and introduce error into a NE-annotated corpus derived from Wikipedia. As shown in Table 4.1, we manually identified 3% of links to Saturn with link text Saturn as disambiguation errors.

Another form of this target-referent mismatch may occur without error on the editor's part. An entity name may link to a related entity or topic which is perhaps more noteworthy or more descriptive for the given context. A link labelled Palestinians may therefore have target Peace process in the Israeli-Palestinian conflict. Redirects may also be formed from one article title to another despite non-identity of the title referents: Arik Vardi redirects to Mirabilis (company) which he founded; Eliza Dolittle redirects to Pygmalion (play) in which she is the main character.[1] In such instances, the linked text's context would

[1]Often the redirect instead points to an appropriate section heading of the target page, which is less problematic for our assumption. For example, the name of the Harry Potter character Katie Bell redirects to a character-specific subsection of the article Dumbledore's Army.

be more useful for entity classification than the target article. Presently we ignore such discrepancies, erroneously labelling the link with the class of the target article.

In other cases, the above corollary may hold while our main assumption is violated. Wikipedia editors may *refer* to an entity without using its name, and link the referring text to the appropriate entity article though it would not usually be marked up in a NE-annotated corpus. For example the fragment:

> In the Ukraine, anarchists fought in the civil war ...

includes a link from the text civil war to an article on the Russian Civil War.

More systematic violations of our assumption also occur, such as the linking of nationalities to articles about countries as in:

> Thomas Cruise ... is an American actor and film producer.

where American links to United States. While these arguably share the same referent, common entity annotation schema do not mark nationalities with the same entity label as geo-political entities. The systematic nature of such violations means they may be detected and altered to conform to an annotation scheme, as we implement for nationalities in section 6.3.2.

It is clear from each of these examples that our assumption is not upheld in practice in Wikipedia articles. Although it is naive, our results (Chapter 7) suggest that the assumption is maintained sufficiently to create viable annotated corpora. Recall also that all linguistic corpora, even when labelled as *gold standards*, have some level of error in their annotation, as described by measures of human annotator agreement where available. A generated corpus may have a greater but estimable level of error. We are able to successfully produce a working training corpus with errors such as these included, with some common violations (e.g. nationalities) corrected by way of manual rules.

4.2 Processing Wikipedia

Wikipedia is designed primarily for human viewers, and not for use as a textual corpus. Publicly available Wikipedia dumps need to be processed to extract useful information and remove noise; they need to be split into sentences, tokenized and cleaned, as exemplified in Figure 4.3.

Wikipedia markup	Tokenised sentences with links
[[Image:Sydney locator-MJC.png|thumb|Location of Sydney]] '''Sydney''' ({{pronEng|'sɪdni:}}) is the most populous city in [[Australia]], with a [[metropolitan area]] population of approximately 4.28 million (2006 estimate).<ref name="2006population">. . . </ref> Sydney	Sydney is the most populous city in Australia|Australia , with a metropolitan_area|Metropolitan_area population of approximately 4.28 million (2006 estimate) . Sydney is the state capital of New_South_Wales| New_South_Wales , and is the site of the first . . .

FIGURE 4.3: An example of processing Wikipedia markup to produce sentences with links. Link text and target are separated by a pipe in our sentence format.

4.2.1 Wikipedia snapshot dumps

The article source texts of all public wikis run by the Wikimedia Foundation (of which English Wikipedia is the largest) are made available at `http://download.wikimedia.org/`. They provide snapshot dumps in three formats:

- HTML dumps of the formatted output as served to web browsers;
- XML dumps with metadata and article sources in Mediawiki's native markup format[2] (with or without revision history, etc.);
- SQL dumps of structured databases, such as the category or link networks.

The static HTML dumps are much larger than equivalent data in Mediawiki format (approximately 14GB and 4GB respectively), and although they can be processed more easily with commonly available tools for HTML, they lack the semantic and syntactic simplicity of purpose-made Mediawiki markup. Use of the XML dumps is therefore preferable for accessing Wikipedia article text. We considered using the SQL dumps as a source of structured data, but found that their snapshot times did not correspond to the XML dumps, making the two sources difficult to integrate.

4.2.2 Parsing Mediawiki markup

We initially considered a number of publicly-available NLP-directed processors for accessing structured content from Wikipedia dumps. As shown in Table 4.2, a variety of systems are available, either providing the processing system or its output freely online. Considering only systems whose source code was available and which minimally produced article text with links and extracted useful structural features (link and category graphs), we experimented with the WikiXML system (ILPS, 2007), but found that its dependence on a modified version of the official PHP Mediawiki engine made it excessively slow.

[2] See `http://www.mediawiki.org/wiki/Help:Formatting`.

Processing system	Source code	Output format	Article text	Structured features							Notes
				Red.	Tpl.	In	Out	Cat.	Lang.	Dab.	
Auer et al. (2007)	PHP	RDF		X	X	X		X		X	Category graph cleaned; incorporates YAGO (Suchanek et al., 2007) and WordNet data.
Denoyer and Gallinari (2006)	-	XML	✓		N						
Gabrilovich (2005)	PERL	XML	✓	R	R	X	AX	AX			Collects *related links* into external resource; stub articles marked.
ILPS (2007)	PHP	XML, DB	✓	X	ANX	X	NX	AX	AX		Adds id to every tag.
Metaweb Technologies (2008)	-	XML, DB	✓	RX	NX		N	X			Sentences split; infers synthetic links (where linked term is repeated unlinked).
Schenkel and Kasneci (2007)	-	XML	✓	R	N	N	N	A			Labels articles and links with WordNet synsets.
Toral (API)	C++	XML, DB		X				X	X		
Zaragoza et al. (2007a)	-	Multitag	✓	R	R		N	A	A		Based on ILPS (2007); sentences split; tokenized; tagged with POS, NEs, anaphora.
Zesch et al. (2008)	-	DB, Java API		X		X	X	X		A	Can query based on article properties.

TABLE 4.2: Comparison of a number of NLP-targeted Wikipedia processing systems. *Source code* indicates a programming language if the processing system source code is available. *Output format* indicates the formats produced by the system, where DB may be any database format. *Article text* indicates whether the full article text is provided, rather than only structured data. *Structured features* indicates the representation or handling of some of Wikipedia's structured features (see Appendix A for definitions): redirect pages, templates, in-links, out-links, categories, language links, disambiguation pages. For each, the following codes may be indicated: A– data is appended to the article content; N– some form of annotation is made in-place; R– redirects are resolved or templates expanded; X– data is extracted as an external resource.

We instead opt for a parser which produces a structured representation of Wikipedia markup which can then be processed for our needs, selecting `mwlib`[3] for this purpose.[4] Available as Python open-source code, `mwlib` produces generic parse tree structures from Mediawiki article source, which can then be transformed into other formats. Outermost Article nodes are subdivided into Sections which are further split into Paragraph nodes. Paragraphs contain plain-text nodes interspersed with nodes for links, styling, tables, lists, etc. A parse tree facilitates flexible application of Wikipedia data, and `mwlib`'s active development and popular use elsewhere[5] ensures robust coverage of Mediawiki formatting.

Using `mwlib` parse trees of Wikipedia pages, we extract the following data (see Appendix A for definitions of Wikipedia features):

Redirects: A table of titles and their redirect targets. Redirect articles are removed from later processing, and used to resolve final link targets.

Template categories: In each template page, templates are expanded recursively in order to extract any category links associated with particular templates.

Category data: Category links are extracted for all articles and category pages, incorporating categories of included templates.

Outgoing links: Outgoing links from each article are extracted and redirects resolved.

Incoming links: Outgoing link data was transposed to list incoming links to each article.

Templates: The titles of templates used in each article are extracted.

Tokenised article text: Non-text of portions of Wikipedia articles are removed and the remainder split into sentences and tokens with links as described in the following sections.

4.2.3 Sentence boundary detection

We split Wikipedia's text into sentences for two purposes:

- NER systems are commonly trained on sentence units of text; and
- sentences are largely self-contained units which may be retained or discarded on the basis of selection criteria for our final corpus (see section 4.3).

[3] `http://code.pediapress.com`

[4] A list of Mediawiki markup parsers apart from the official engine is maintained at `http://www.mediawiki.org/wiki/Alternative_parsers`.

[5] `mwlib` is primarily used for producing PDF documents from collections of Mediawiki articles.

Parameter	Qty	Examples
Abbreviations	13,508	sept.; a.k.p.; gb.d.; etc.; hon.sec.; m.comm.; u.s.
Frequent sentence-starters	7,650	according; needless; contraindications; doan; augustus; nietzsche
Collocations	2,642	inc. announced; dept. of; st. albans; law. article; mr. potato
Orthographic data	1,174,479	ethonyms: LC; smerwick: UC; archerfish: LC,UC

TABLE 4.3: Size and examples of parameter collections for Punkt (Kiss and Strunk, 2006) learnt from 1.6 million Wikipedia articles

Detecting sentence boundaries in English requires distinguishing between two uses of the full-stop (.): marking sentence boundaries and marking abbreviations. Abbreviations and ellipses may further double as sentence boundaries. Punkt (Kiss and Strunk, 2006) is an unsupervised sentence boundary detection algorithm which learns the following parameters from text:

Abbreviations: Words frequently collocated with trailing full-stops, with additional heuristics for word length and number of internal full-stops;

Frequent sentence-starters: Words frequently collocated with preceding sentence breaks;

Collocations: Frequently collocated word pairs where the first word ends in a full-stop;

Orthographic data: Whether each word appears capitalised or lowercase sentence-internally;

These parameters together form heuristics for resolving sentence boundary ambiguity.

We use the Punkt implementation included in NLTK[6] (Loper and Bird, 2002), with parameters learnt from 1.6 million Wikipedia articles (see Table 4.3). We apply sentence boundary detection to the structured parse tree by assigning children of paragraph nodes to sentence nodes.

All un-styled text fragments in paragraph nodes of the article parse tree are considered as candidates for sentence boundaries, assuming that styled text (italics, linked, etc.) rarely include sentence breaks. Block elements (tables, lists, etc.) are also considered sentence boundaries. If closing punctuation (e.g. right parenthesis) is split from the end of a sentence, the sentence boundary is realigned to include it.

While the Punkt algorithm generally produced reasonable sentence breaks, we found it had labelled a number of surnames as frequent sentence starters (i.e. they commonly appear after a full-stop). If found after an abbreviation, frequent sentence starters are presumed to begin new sentences. This results

[6]We rewrote portions of the implementation, improving speed and enabling incremental training, and submitted our modifications to the project.

in false-positive sentence breaks in titles like Mr. Jones. We therefore prohibit sentence breaks after common personal title abbreviations (Mr., Ms., etc.) and initials.

Punkt's boundary decision is completely determined by a pair of tokens where the first ends in a full-stop, exacerbating any errors that may affect our corpus. False negatives (sentence boundaries which Punkt does not detect) are not likely to be very detrimental: compound "sentences" thus formed are less likely to meet sentence selection criteria which is judged on a per-token basis (see section 4.3) than shorter actual sentences. False positives are both more likely to be selected and to be detrimental to training, by losing important training instances. For instance, if Punkt fails to consider Mr. an abbreviation, it will produce a sentence break at every occurrence, removing useful NER training patterns relating Mr. to a following PER entity. This suggests that fine-tuning the sentence boundary detection algorithm may improve the overall performance of Wikipedia-derived entity corpora.

4.2.4 Tokenisation

Tokenisation, which largely involves separating punctuation from words, is also performed on the parse tree. The text nodes within each sentence are joined into a single string, then tokenised, and tokens redistributed to appropriate nodes. We define tokens by a hand-built regular-expression, using common conventions for the tokenisation of affixes ('ve, n't, etc.), separating full-stops from abbreviations only at the end of sentences, treating hyphenated expressions as singe tokens, etc.

In addition to manual inspection, we attempted to find systematic errors in tokenisation (and data cleaning as below) by tokenising 2000 articles and counting the appearances of different orthographic patterns (combinations of uppercase, lowercase, digits and punctuation). Unexpected frequent patterns were investigated and the tokeniser modified if appropriate.

4.2.5 Extracting sentences with links

In a final stage of pre-processing, we effectively remove non-sentential data and all markup other than links from Wikipedia articles. Each node in the parse tree is handled according to its type:

Handling the *potential noise* category introduces a dummy token. If this token appears within a bracketed expression, this expression is removed. Otherwise, the entire sentence is discarded. We also discard

Node class	Description	Examples	Action
Safely discarded	Block elements unlikely to interfere with sentential data	Tables, lists, section headings	Discard node
Potential noise	Inline elements not easily translated to sentential data	HTML tags, template inclusions	Replace node with dummy token
Redundant styling	Markup not affecting its textual content	Bold text, Wikipedia-external links	Retain text but ignore structure
Useful markup	Structural elements useful in further processing	Sentences, paragraphs, inter-article links	Retain structure

TABLE 4.4: Handling of different types of parse tree nodes to producing sentences with links

text that results from parsing errors[7] and poorly edited articles, such as loose tabular data or incorrectly-formatted lists (using - for bullets rather than *), as well as URLs. The resulting textual data is stored with each sentence on a new line, and linked text notated as link_text|Target_article (see Figure 4.3).

4.3 Labelling and selecting sentences

Given entity classifications for all articles in Wikipedia, links can be replaced with entity tags, the text formatted identically to gold-standard corpora (IOB NE annotation; POS-tagged; see section 3.2) and a NER system can be trained with the resulting corpus, treating each Wikipedia article as a new training document. Although all sentences could be included in a training corpus, many would introduce error; in addition, corpus sizes are limited by memory constraints in NER training. We therefore need to select sentences on the basis of *confidence* and *utility*.

Intuitively, we would like to include all sentences for which we are confident that all entities and only entities are correctly labelled. Since standard classes of named entities (LOC, ORG, PER, etc.) in English are almost always capitalised, we use the following basic criterion for sentence confidence:

SELECTION CRITERION (Confidence). *Include a sentence if all capitalised words are labelled with entity tags.*

This criterion (detailed further in section 4.3.1) is too restrictive: many words which English conventionally capitalises should not be labelled as entities, but under this criterion all sentences containing such capitalised words are discarded. We therefore describe loosening its application in section 4.3.2.

[7] Some of these occur because we do not expand templates, which may contain part of a structural element which the parser cannot otherwise identify.

1500s in music; 1939 in Wales; Abdominal wall; Administrative proceeding; Arcade emulator; Augustan poetry; Bartonellaceae; Canal; C-sis internal ribosome entry site (IRES); Data logger; Daytime wetting; Flag of Guadalajara; Heterobranchia; Honor system virus; Horses in art; Inclination; Insulin lispro; List of films directed by Mani Ratnam featuring A R Rahman; List of Galaxy Express 999 episodes; List of places named after Josip Broz Tito; Madman theory; Monetary reform; Rana Pierniamarilla; Trains in the Netherlands; Vegetarianism; Washer (hardware); Wheelchair basketball.

FIGURE 4.4: A selection of articles that would usually be considered non-entity topics

We use another simple criterion to include a higher density of useful training data:[8]

SELECTION CRITERION (Utility). *Include a sentence only if it includes at least one tagged entity.*

We have not found the inclusion of sentences without entity annotations to improve NER performance.

4.3.1 Non-entity targets and anomalous capitalisation

Not all links can be easily labelled as entity annotations. Some link targets:

- describe non-entity topics;
- are disambiguation or administrative pages, not about particular topics;
- refer to particular sections of articles;
- do not exist; or
- cannot be classified.

All but the first can be considered *indeterminate link targets*. We currently have no way to transform such links into entity annotations, and hence discard sentences containing them. We allow an exception for all-lowercase links, which we assume are non-entities and do not label.

In practice, we treat non-entity links similarly, but the motivation differs. An initial approach to non-entity links is simply to leave the linked text unannotated. This is complicated by the fact that non-entity articles cover a very broad range of topics, as suggested by the examples in Figure 4.4. The titles of many non-entity articles include the names of particular entities (e.g. Politics of Indonesia), and if such a title is linked within a sentence, standard entity annotation would require the entity reference to be annotated. Therefore simply leaving all non-entity links unlabelled introduces error into our corpus.

[8] More extensive criteria may consider quality or novelty of the data, although eventually we aim to include all sentences.

Conversely, we need to generally exclude all-lowercase links to entity articles. Entities are not always referred to by name as is the case of our previous example:

> In the Ukraine, anarchists fought in the civil war ...

where the text civil war links to an article on the Russian Civil War. As an exception to this rule, the names of some entities are conventionally written in lowercase, as is the case with gzip and Canadian poet bill bissett. Since Wikipedia generally assumes that all titles begin with a capital, it allows for such exceptions to be marked by including the template Lowercase. This allows us to include some correctly-lowercase entities in our corpus, while removing others.

These notions lead to the following expansion of the Confidence Criterion:

SELECTION CRITERION (Confidence: expanded). *Include a sentence only if it does not contain:*

- *All-lowercase links to entity articles, unless the target is marked Lowercase;*
- *Capitalised text, other than sentence-initial capitalisation, which is:*
  - *unlinked;*
  - *linked to a non-entity article; or*
  - *linked to an indeterminate target.*

4.3.2 Conventional capitalisation

While we assume that proper name entities are capitalised in English and use this as the basis of the Confidence Criterion, sentence selection must be relaxed to allow for other instances of conventionalised capitalisation in English which do not require entity annotations. These include:

- first words of sentences;
- pronouns (only first person I in English);
- dates (e.g. Monday, February, 70 AD);
- acronyms;
- personal titles (e.g. Mr., President);
- adjectival forms of names (e.g. nationalities).

These should form exceptions to our general sentence selection criterion. In producing an initial Wikipedia-derived training corpus, we include only a basic exception for capitalised sentence-initial words. Specifically, we allow inclusion of sentences whose first word is in an edited list of 285 frequent sentence starters extracted in the process of training Punkt (Kiss and Strunk, 2006) on a few thousand Wikipedia articles (see section 4.2.3). Similar exceptions to our selection criterion are described below in section 6.3.1.

4.4 Summary

We have presented a method for transforming Wikipedia into very large named entity-annotated corpora, by assuming that linked terms correspond to entity annotations whose classes are given by the target article topic. Snapshots of Wikipedia data are readily available online and can be processed into sentences with links. Given the entity classification of each Wikipedia article, these links are replaced with entity annotations, and sentences with sufficient *confidence* (where capitalised words correspond to entity links) and *utility* (at least one entity tagged) are selected for a corpus. A procedure for classifying all of Wikipedia's articles into entity classes is described in the following chapter, which is followed by an assessment of this baseline approach (Chapter 6). The viability of this method for producing high-quality corpora is ultimately realised (see Chapter 7) only after extensions are applied to loosen the capitalisation-based sentence selection criterion, increase sentence coverage, and adjust entity annotations to better conform to standard annotation schemes.

CHAPTER 5

Article Classification

We classify Wikipedia's articles into a predefined set of named entity classes which can be used to label the text of incoming links. The key motivation for this approach is that links' target articles are much more indicative of the linked text's entity classes than the sentential context of the link.

As reviewed in section 2.3.1, related classification tasks have been attempted in recent work. Rather than relying on manually-crafted rules (e.g. Richman and Schone (2008)) or bag-of-words machine learning methods (e.g. Dakka and Cucerzan (2008)), we classify the articles through a bootstrapping approach (see Figure 5.1):

(1) Infer a number of heuristics from a hand-labelled sample of articles;
(2) Attempt to use these heuristics to classify all Wikipedia articles;
(3) Infer further heuristics from the resulting classifications;
(4) Repeat from (2) until convergence.

This approach reflects our intuitions that (a) it is difficult to manually design rules with high coverage for a data set as diverse as Wikipedia; (b) bag-of-words approaches ignore many structural and linguistic

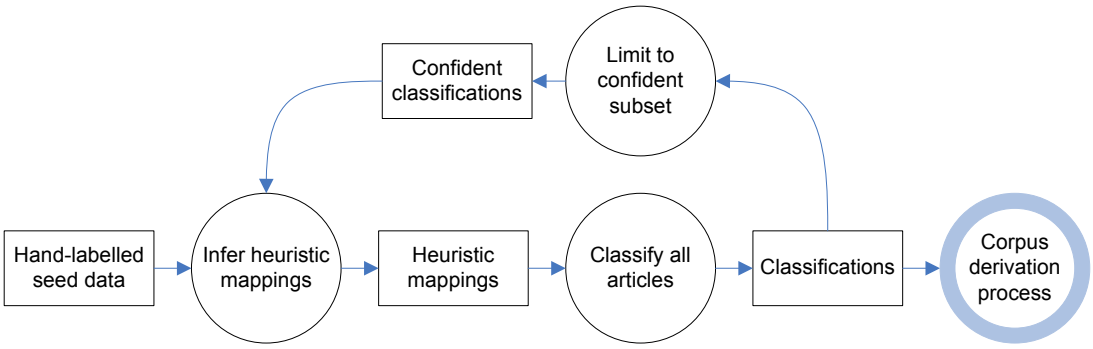


FIGURE 5.1: A bootstrapping approach to article classification

cues available in Wikipedia; (c) while the training data may provide insufficient heuristics by itself, a semi-supervised approach is able to learn heuristics from unlabelled data in addition to the training set.[1]

In producing a NE-annotated corpus, high accuracy of entity labels is much more important than high sentence coverage, so we use a classification methodology which may mark some article classifications as unknown. Sentences linking to articles of unknown class are then excluded from our training corpus. Each article is classified as one of: a named entity class (e.g. LOC, PER); a disambiguation page (DAB); a non-entity page (NON); or unknown (UNK).

Our approach to classification is detailed in this chapter, beginning with the heuristics used for classification (section 5.1) and the application of those heuristics through the bootstrapping process (5.2). We put this approach into practice using a set of manually-labelled seeds (5.3), used in evaluation as described in section 5.4.

5.1 Classification heuristics

Our bootstrapping classifier uses a number of heuristics to label articles with appropriate classes. Non-entity articles are often identified by a few static heuristics as discussed in the following subsection, while more generic heuristics are used to label articles of all classes. Some of the Wikipedia features used—as illustrated in Figure 5.2—include: the definitional noun phrase found in the first sentence of many articles (The *University of Sydney* ... is the oldest *university* ...); some of the article's Wikipedia categories (e.g. *Educational institutions* established in 1850); and a heuristic based on the capitalisation of incoming links. Table 5.1 gives examples of the combination of features used to predict article classes.

5.1.1 Non-entity classification

The class of non-entities is broad and therefore difficult to capture with a limited set of semantic heuristics and a small sample of hand-labelled training examples. It incorporates such broad subjects as Politics of Indonesia, List of AFL teams or Philosophy. By using additional heuristics to identify non-entity articles, we also avoid false-positive assignments to entity classes. Hence we use a capitalisation heuristic

[1] Co-training (Blum and Mitchell, 1998) is another semi-supervised approach that provides similar advantages, but we have found the present method to be sufficiently effective.

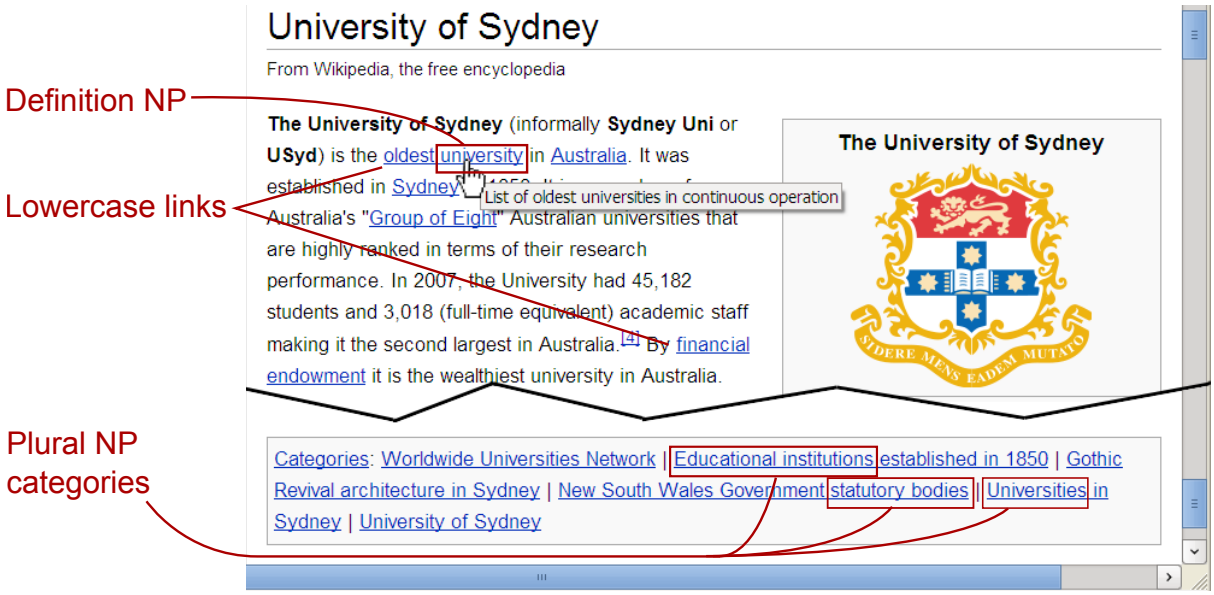


FIGURE 5.2: A portion of the Wikipedia article on the University of Sydney with some useful features for classification marked

in addition to rules for specific types of non-entity articles. Common articles not identified by these rules may additionally be picked up by the generic heuristics discussed in the following section.

Capitalisation heuristic. For the most common generic named entity classification taxonomies (see section 2.1), all named entities are proper nouns, and hence are ordinarily capitalised in English. Therefore, like Bunescu and Paşca (2006), we use capitalisation as an indication of whether an article likely refers to an entity, prior to other classification strategies. Their approach counts mentions of the title within the article itself, but there are often very few such verbatim references. We instead rely on the text of incoming links to a given article. When linked within a sentence, the names of non-entities are likely to not be capitalised. We therefore mark an article as NON if:

- it has at least 5 incoming links whose link text is a case-insensitive substring of the target;
- 30% of these links begin with a lowercase letter;
- it does not include the template Lowercase; and
- its title does not begin with a determiner (A, An or The).

These parameters were selected through inspection of link capitalisation statistics, comparing entity and non-entity articles. While one may expect a higher threshold than 30%, a number of contexts conventionally require capitalisation of typically lowercase terms (e.g. at the start of a sentence or list

(a) Mapped article features

Article	Mapped plural category NPs	Definition NP
'Allo 'Allo! (series 5)		
4th Medical Battalion		
Analogue (literature)	concepts, motifs	
Ben van Berkel	architects, architecture firms, births, people (×2)	
It's My Time	albums (×3), album stubs	studio album
Jeffrey Sonnenfeld		Professor
List of NAIA Conferences		list
Tickencote	villages	small village
United Federation of Teachers	professional associations, trade unions	labor union
University of Sydney	educational institutions, universities (×2)	

(b) Heuristic values and predicted classes

Article	DAB	List	Cap.	Plural category mappings					Def'n	Class
				LOC	ORG	PER	MISC	NON		
'Allo 'Allo! (series 5)	-	-	-	0	0	0	0	0	-	UNK
4th Medical Battalion	+	-	-	0	0	0	0	0	-	DAB
Analogue (literature)	-	-	+	0	0	0	0	2	-	NON
Ben van Berkel	-	-	-	0	1	4	0	0	-	PER
It's My Time	-	-	-	0	0	0	4	0	MISC	MISC
Jeffrey Sonnenfeld	-	-	-	0	0	0	0	0	PER	PER
List of NAIA Conferences	-	+	-	0	0	0	0	0	NON	NON
Tickencote	-	-	-	1	0	0	0	0	LOC	LOC
United Federation of Teachers	-	-	-	0	1	0	0	1	ORG	ORG
University of Sydney	-	-	-	0	3	0	0	0	-	ORG

TABLE 5.1: The classification heuristics for a sample of articles. Table (a) shows mapped article category and definition nouns. Table (b) gives the predicted classes and values of all heuristics used to classify the articles (disambiguation, list, capitalisation heuristics; number of category NPs in each class; class of definition NP).

item), while the capitalisation of proper nouns is more consistent. With less than five samples, the number of lowercase links is insignificant.

The last two conditions provide exceptions. The Lowercase template is used in articles which require that the first letter of their title should not be capitalised, in opposition to the Wikipedia norm.[2] Such words, such as iPod, al-Qaeda, or Canadian poet bill bissett, will generally be found in text with a lowercase initial, but we must not exclude them from being considered as entities.

If a proper name begins with a determiner, a lack of consistent stylistic convention may lead to arbitrary capitalisation of the determiner. For instance, the musical group A Ragmuffin Band has 17 incoming

[2]By default, all characters in the title of a Wikipedia article are case-sensitive except for the first. The first letter is always shown capitalised unless overridden by special templates such as Lowercase.

links, 7 (41%) of which begin with a lowercase character. Without a rule pertaining to determiner-initial titles, this article would be falsely marked as a non-entity.

Of our hand-labelled articles, this capitalisation heuristic identifies 25% of non-entity articles (excluding disambiguation pages), without mis-classifying any entities as non-entities.

For a named entity classification scheme which incorporates non-proper name entities such as colours or animals (e.g. Brunstein (2002); Sekine et al. (2002)), capitalisation would not specify the entity class directly, but may limit the possible categories for further hierarchical classification. Such a heuristic may also only be applied to languages with capitalisation conventions comparable to English.

List heuristic. If the title of an article begins with List of , we mark it as NON, identifying a further 7% of the hand-labelled non-entity articles. Lists are a common type of non-entity page which enumerate or tabulate related data, as in List of English poets or List of geological phenomena. Occasionally these articles are linked to within text, as in Figure 5.2 where the text oldest university links to the page List of oldest universities in continuous operation, in which case capitalisation heuristics may apply. Often they are linked only within the context of *See also* sections of articles where links are usually capitalised. Hence it is useful to identify this group of articles with a separate heuristic.

Disambiguation heuristic. Disambiguation pages, which distinguish the different meanings of a particular title text, need to be differentiated from articles about *particular* entities. We also distinguish them from other non-entity articles for specialised processing (see section 6.3.3). We mark an article as a disambiguation page if any of the following conditions are met:

- its title contains [dis]ambiguation);
- it is in the Disambiguation category;
- it is in the Surnames category;
- it is in a category ending disambiguation; or
- it is in a category containing [a]mbiguous.

These rules find 92% of the articles marked as disambiguation pages in our hand-labelled data, and mark no false positives.

5.1.2 Category nouns

Wikipedia's own category hierarchy is of limited use for article classification, as discussed in section 2.3.1. By following Suchanek et al.'s (2007) suggestion of identifying useful ontological categories as those whose first noun phrase is plural, we are able to group articles of the same class. We therefore map plural head nouns of an article's categories to named entity classes. For example, institutions from the category Educational institutions established in 1850 might map to ORG. Since numerous categories are often assigned to a single article, we may also infer further mappings from the co-occurrence of category head nouns. This method effectively clusters a subset of Wikipedia's categories on the basis of their head nouns and maps these clusters to NE classes. While Suchanek et al. (2007) derived this mapping from WordNet, we do so through a bootstrapping process.

In order to classify a particular article, the head noun of each of its categories is identified and, if plural, mapped case-insensitively to an entity class, if such a mapping is known. Thus a score for each class can be counted (see Table 5.1 for examples). The article is then labelled as:

- the most frequent class if one exists;
- UNK if multiple classes tie as most frequent;
- UNK if no categories with class mappings were present.

The following subsections describe a number of caveats and practical issues involved in the determination of useful head nouns from category titles.

5.1.2.1 Identifying head nouns

To this end, we POS-tagged and chunked the titles of all Wikipedia categories using the C&C tools (Curran et al., 2007) to identify the head of the first noun phrase (NP) of each and whether its head was plural (tagged NNS for common nouns). For instance, the category Educational institutions established in 1850 was chunked as:[3]

```
Educational|JJ|I-NP institutions|NNS|I-NP established|VBN|I-VP in|IN|I-PP 1850|CD|I-NP
```

Although the entire category is a single noun phrase, the shallow parsing performed by the chunker does not identify the relationship between Educational institutions and its post-modifier established in 1850,

[3]The output of the chunking process is a list of tokens each followed by a pipe, then the part-of-speech tag, another pipe and the chunk tag given in IOB format.

such that the first noun phrase group marked by the chunker ends with the phrase's head noun, in this case institutions. Similarly we attempt to find the head noun of each category title.

Errors in POS tagging and chunking cause this method to fail for some categories. These errors are compounded because the taggers are trained on sentential data rather than the noun phrases that we are chunking presently.[4] The following are examples of some common errors:

(1) Within|IN|I-PP Temptation|NNP|I-NP songs|NNS|I-NP
(2) Woodworking|VBG|I-VP hand|NN|I-NP tools|NNS|I-NP
(3) Scouting|VBG|I-VP in|IN|I-PP the|DT|I-NP Bahamas|NNPS|I-NP
(4) Islands|NNP|I-NP of|IN|I-PP Wisconsin|NNP|I-NP

In the first three examples the first token should be identified as part of the noun phrase (chunk tag I-NP) but is not. The head of the first noun phrase is still the appropriate plural head, except in example (3), since Bahamas—possibly considered a plural head (see below)—is identified as its first head noun. We therefore relabel all gerunds (VBG) appearing before the first noun phrase as part of a noun phrase, such that the first NP in example (3) is Scouting. The final example illustrates a case where Islands is incorrectly identified as a singular proper noun. After briefly reviewing some cases where this occurs (i.e. the first token is incorrectly POS-tagged NNP), we have constructed a rough rule to modify the POS tagger output, replacing the first token's tag with NNS if the word ends with -s (but not -ss) and if a preposition is found elsewhere in the category title. This catches many important cases, while missing some and over-generating in others, but facilitates the identification of many entities otherwise lost to our heuristic (categories containing lakes, for instance, are consistently subject to this POS tagging error). Although it is possible to modify tagger output to handle some cases, we see that tagging errors are able to induce some amount of error into this process.

5.1.2.2 Inappropriate plural nouns

Our assumption that a plural head noun is indicative of an ontological category is false for words like mathematics or Bahamas which are conventionally plural.

It is not clear whether proper noun plurals (POS-tagged NNPS) should be considered for our classification heuristic. Head nouns such as Americans or Scientologists are indicative of PER entities, but others, such as United States or Sydney Swans, are not similarly useful. Ideally, since the latter would likely

[4] Training a POS and chunk model on only NPs is left to future work.

appear in articles of various class, the bootstrapping process may effectively ignore them. When evaluating this variable on our held-out test data, we nonetheless found a slight improvement (1% micro F) if category titles with proper noun heads were removed. This performance increase is logical since useful NNPS heads generally refer to PER entities which are almost always identifiable from other features, whereas the noise created by less useful NNPS heads is likely to be detrimental. For this reason plural proper noun heads were ignored in our final experiments.

In other cases, plural head nouns are not indicative of ontological categories, as with nouns that are conventionally pluralised, such as physics or media; or which categorise the article itself rather than the entity it describes, such as stubs or articles (as in Articles needing attention); or describe a property of the entity, as in 1980 births or 1867 establishments, which are not ontological and the term establishments can refer to geo-political entities such as countries in addition to organisations. We experimented with compiling lists of such plural exceptions, but they consistently caused losses in both precision and recall on our held-out and cross-validation data, up to 2% micro F.

5.1.2.3 N-gram category heads

In many cases a unigram (single-word) category head does not sufficiently discriminate the mapping between noun phrases and entity classes. For instance, train stations are facilities, unlike radio stations; media players are products, while baseball players are people. For this reason, we produce a list of likely bigram collocations and allow both unigrams and bigrams to be mapped to entity classes in the inference stage as detailed below. Thus if a class mapping is known for railway stations, an article in the category Railway stations in Sydney will be labelled with that class and any mapping for the unigram stations will be ignored. To identify collocations, we used Pearson's chi-square test (Manning and Schütze, 1999) for all noun phrases from category titles, producing 7,925 bigrams ($\chi^2 > 3.5$; we experimented with smaller lists but found they decreased performance).

5.1.2.4 Still a folksonomy

Despite some success in using the Wikipedia category graph through plural head nouns, inconsistent categorisation in Wikipedia still generates noise. For example, the article Company of Guinea, a Portuguese colonial organisation, is in the category Former Portugese colonies. The topic is thematically related to former Portuguese colonies, but is not one itself. If such errors were to appear in the training data, without competing evidence, an erroneous mapping may be made from colonies to ORG, causing

the mis-classification of other former colonies such as Australia. Hence without sufficient training data and guards against conflicting information in the bootstrapping process, our classification methodology can be very sensitive to popular thematic categorisation used in Wikipedia.

5.1.3 Definition noun

We utilise definitional noun phrases from the first sentence of many articles in a similar manner to our use of plural category nouns. In some cases, articles do not have any ontological categories (1.2% have no categories at all), or any which we are able to map to entity classes. For these cases, and when the category nouns result in a conflict, it is useful to have a back-off heuristic for determining the entity class. Specifically, if an article is marked UNK by previous heuristics, a definition noun may relabel it as an entity; if only one category mapping is available for an article, and the definition noun mapping contradicts it, we relabel the article as UNK.

Kazama and Torisawa (2007) used a classification feature based on the first noun phrase following a copula (is, was, are, etc.) in the first sentence of each article, if present. For example, university would be extracted from the opening sentence of Figure 5.2, The University of Sydney . . . is the oldest university in Australia. We use a similar *definitional noun phrase* feature as a heuristic in our bootstrapping process.

The definitional noun phrase is extracted from each article as follows:

(1) Extract the first sentence (see section 4.2)
(2) Delete any parenthesised expressions that may complicate tagging;
(3) POS-tag and chunk the sentence with the C&C tools as above;
(4) Find the first copula ("to be" word) in the sentence; if none, give up;
(5) Find the noun phrase immediately following the copula; if none, give up;
(6) If the head noun is one of {form, kind, one, sort, type, variety} and is followed by of, find the next noun phrase;
(7) If the noun phrase is followed by the possessive 's, find the next noun phrase.

Finding the appropriate definition noun therefore relies on correctly identifying the first sentence, its containing a copula, and the correct chunking of the sentence. While an article may include multiple definition nouns in its opening sentence, as is the case with Tom Cruise who is an American actor and film producer, we presently only extract and use the first definition noun. The last two conditions here

(inspired by Kazama and Torisawa (2007)) allow for cases where the first noun phrase chunk found directly after the copula is not the best to define the article entity, as in . . . is one of Australia's leading universities where one is not as appropriate as universities; or children's book in which the chunker marks 's as a new phrase. The list of words in (6) was compiled with reference to WordNet and some inspection of the data, although the step may be more effective with some alteration, as type of company should be distinguished from company, while universities extracted from the previous example would be better lemmatised to university. Fine tuning this is left to future work. This method extracts definition nouns from 81.3% of Wikipedia articles.

As with category nouns, some of the tagging output needs to be corrected. In particular, a peculiar case:

An|DT|I-NP abacus|NN|I-NP is|VBZ|I-VP . . . a|DT|I-NP calculating|VBG|I-VP tool|NN|I-NP

The taggers fail to mark calculating as pre-modifier of tool. In such cases, where a single-word verb phrase (VP) is found between a determiner and a noun phrase, we relabel the VP as NP.

Many opening sentences of articles are not accessible to our approach. While the opening sentence of University of Sydney in Figure 5.2 is a definition, the opening sentence had been The University of Sydney was established in Sydney in 1850 and is the oldest university in Australia. when we acquired the Wikipedia data for processing.[5] This sentence has two copulas, the first providing no definition noun. In other cases, the first sentence is not formulated as a definition.

As with category nouns, we allow unigrams or bigrams from definition NPs to be mapped to entity classes. We produce a list of bigrams as collocations from all first sentences of Wikipedia articles, excluding those with a determiner. This gives us 66,856 bigrams ($\chi^2 > 3.5$).

5.2 Bootstrapping: inference and confidence

The bootstrapping process allows us to make use of both labelled and unlabelled data by inferring heuristic mappings from confident classifications (see Figure 5.1).

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

[5] The 22 May 2008 English Wikipedia snapshot

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

TABLE 5.2: Varying threshold (t) values used for mapping inference

m articles contradict it, then we accept the mapping if

$$(n \geq t) \wedge \left(\frac{m}{n+m} < p \right)$$

for some constant thresholds t and p. We have used $p = 0.25$, while t varies with values given in Table 5.2, to account for the fact that definition nouns vary much more than category nouns, and that we are more confident about our manual classifications than our system's predictions. We performed a few experiments in varying p and t values. Reducing p to 0.1, for instance, increased precision by 1% and decreased recall by 3% on our held-out data.

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). We considered removing this second requirement, but found that doing so reduced precision by 2.2% and increased recall by only 0.4%.

We also performed early experiments in which we inferred new category noun mappings through Wikipedia's category hierarchy. For example, since Educational institutions established in 1850 appears in the category Organizations established in 1850, one may infer that educational institutions and organisations have the same class. Whether inferring up or down the category hierarchy, we found this to be quite detrimental, again suggesting the unreliability of Wikipedia's category network, except for leaf node membership where we use it successfully.

5.3 Data

In order to seed the bootstrapping process, we manually classified a set of 1300 articles. Originally, a set of 1100 was sampled randomly from among all Wikipedia articles. We found that this data set was highly populated by entities that are frequent in Wikipedia, such as authors or albums. It very poorly represented entities which, although popular in usage, were few in number, such as countries, and as a result, bootstrapping failed to determine a mapping from countries as a category noun to LOC. We therefore augmented the data set with 200 articles randomly sampled from among the articles with at

Category label	Qty	Category label	Qty	Category label	Qty
?	7	MISC→event→other	9	NON→substance	28
DAB	61	MISC→event→sports	15	NON	189
FAC→airport	3	MISC→event→war	5	NUM→date	16
FAC→building	17	MISC→language	5	ORG→?	1
FAC→other	15	MISC→law	7	ORG→army	4
FAC→road	12	MISC→NORP→nationality	10	ORG→band	20
LOC→GPE→country	12	MISC→NORP→religion	1	ORG→corporation	42
LOC→GPE→other	57	MISC→product→album	33	ORG→educational	27
LOC→GPE→state	6	MISC→product→other	10	ORG→government	13
LOC→GPE→town	81	MISC→product→software	14	ORG→hospital	1
LOC→non-GPE→?	2	MISC→product→vehicle	5	ORG→hotel	3
LOC→non-GPE→other	19	MISC→product→weapon	2	ORG→museum	2
LOC→non-GPE→region	8	MISC→vessel	10	ORG→other	34
LOC→non-GPE→river	7	MISC→work→?	1	ORG→political	5
LOC→non-GPE→space	3	MISC→work→book	10	ORG→religious	2
LOC→non-GPE→water	6	MISC→work→film	17	ORG→sport	23
MISC→?	13	MISC→work→other	46	PER→animal	2
MISC→courtcase	1	MISC→work→play	1	PER→fictional	18
MISC→event→?	1	MISC→work→song	14	PER→person	283
MISC→event→hurricane	1	NON→life	40		

TABLE 5.3: The 58 class labels manually assigned to 1300 articles

least 700 incoming links, which are more likely to appear in sentences of our NE-annotated corpora. Although we retain the initial 1100 articles they are less likely to be refined than popular articles and hence may produce more noise in our classification.

A single annotator manually labelled the articles using a entity class hierarchy based on used in tthe BBN corpus (Brunstein, 2002) as shown in Table 5.3. We added some new categories, e.g. DAB for dis-ambiguation pages; and MISC→product→album, which may be better as MISC→work→album. Some BBN types had no instances in our article sample. We also designed the named entity hierarchy such that it could be easily downscaled to coarser standard NER annotations. When classification at some level of hierarchy was left undecided it was marked '?'. Seven articles were left with undecided coarse classification as shown in Table 5.4.

The 1293 articles with decided broad classifications were reduced to CONLL-compatible classes {LOC, MISC, ORG, PER } plus NON and DAB for evaluation and use in our system. In particular, while facil-ities are often ambiguously organisations or locations, FAC was mapped to LOC with the exception of FAC→airport which was mapped to ORG.

Article	Description	Ambiguity
Age of Enlightenment	a historical era	MISC or NON
Bukharan tenga	a currency	MISC or NON
Cyrus and John	multiple persons	possibly PER→person
Fahrenheit	unit of measure	MISC or NON
Liberian Premier League	sports league	MISC→event or ORG
Milly-Molly-Mandy	series of novels	PER→fictional or MISC→work→other
Singh sabha movement	a reformation movement	ORG→religious or perhaps MISC

TABLE 5.4: The seven articles left unclassified in manual labelling

5.4 Evaluation

In order to evaluate our performance, we held out 15% (195 articles) of the manually-labelled data as a development evaluation set. For final testing we report results of a ten-fold cross-validation.

We use standard measures for evaluating single-class classification: precision (P), recall (R) and $F = \frac{2PR}{P+R}$ calculated for each class. Per-class values may be combined to an overall metric with either a micro-average or macro-average approach. In the former, each test instance is weighted equally, while the latter equates all classes. The unbalanced size of entity classes suggests that a micro-average approach is more appropriate for our classification task, but the predominance of PER entities which are often easy to classify because of their relatively systematic Wikipedia categorisation[6] means that high micro-average results can be easily obtained, while the macro-average is more indicative of performance in minority classes.

5.5 Results

We present the results of our classification, looking at progressive performance during bootstrapping, the effect of our heuristics, and final cross-validation results.

Table 5.5 shows the convergence of classification test results during bootstrapping, with proportions of heuristics for each class shown in Figure 5.3. For all our experiments on the held-out data we found that results were unchanging after the third feedback stage. In general, all other results in this section follow three iterations of feedback. The high micro-averaged results from the seed classification is due to consistently high-precision as well as high recall (above 90%) for LOC and PER, which together account

[6] As per `http://en.wikipedia.org/wiki/Wikipedia:Categorization_of_people`, the categorisation of biographical articles is quite well-specified; most PER articles belong to categories giving the year of birth and/or death.

Iter.	# of mappings		%	Class F						Macro			Micro		
	Cat.	Def'n	UNK	NON	DAB	LOC	MISC	ORG	PER	P	R	F	P	R	F
Seed	1050	132	22	77	100	90	71	76	94	94	78	85	93	73	82
1	8256	24277	14	81	100	94	76	88	92	94	85	88	93	79	85
2	8763	26454	13	84	100	94	76	88	92	94	85	89	93	80	86
3	8890	26976	11	84	100	99	78	88	92	96	86	90	95	84	89
4	8951	27048	11	84	100	99	78	88	92	96	86	90	95	84	89
5	8992	27127	11	84	100	99	78	88	92	96	86	90	95	84	89

TABLE 5.5: Performance statistics between stages of bootstrapping on the held-out test set

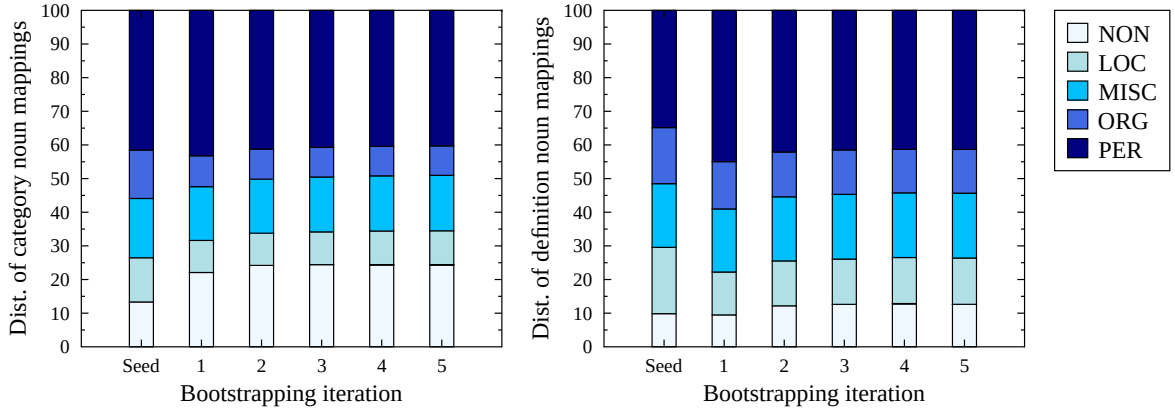


FIGURE 5.3: Distributions of heuristic mappings during bootstrapping

Classification heuristics	P	R	F
All heuristics	95	84	89
No definition nouns	98	80	87
No capitalisation	92	80	85
No category nouns	74	42	54

TABLE 5.6: Effect of removing some heuristics for classification on the held-out test set (micro-average results)

for more than a third of the held-out test data. For most classes, recall and precision are increased by bootstrapping, but PER precision decreases 6% with a recall increase of 2%.

In Table 5.6 we show the effect of removing some classification heuristics. These results are somewhat artificial: in order to show the impact of each heuristic alone, we infer category and definition noun mappings in bootstrapping in all cases. We nonetheless see that category nouns are by far the most powerful feature, and that definition nouns increase recall while losing some precision. The loss of only

(a) Our results

Class	Dist. %	P	R	F
NON	21	89 ± 7	69 ± 10	78 ± 7
DAB	5	100 ± 0	93 ± 10	96 ± 6
LOC	19	95 ± 5	94 ± 4	95 ± 4
MISC	18	93 ± 5	72 ± 13	80 ± 9
ORG	14	92 ± 6	80 ± 9	85 ± 7
PER	23	96 ± 4	98 ± 2	97 ± 2
Macro-average		94 ± 2	84 ± 3	88 ± 3
Micro-average		94 ± 2	84 ± 3	89 ± 2
% labelled UNK: 10 ± 2				

(b) Results from Dakka and Cucerzan (2008)

Class	Dist. %	P	R	F
NON	11	89	87	88
LOC	12	96	95	95
MISC	25	93	92	92
ORG	11	94	93	93
PER	41	94	96	95
Micro-average		90	89	90

TABLE 5.7: Comparative results of overall classification. Table (a) gives our cross-validation classification results (average and standard deviation). Table (b) shows results from Dakka and Cucerzan's (2008) SVM-based bag-of-words classification.

Class	Bootstrapping feedback iteration									In-links
	Seed	1	2	3	4	5	6	7	8	
UNK	17.0	10.6	9.8	9.5	9.4	9.4	9.4	9.4	9.4	12.4
NON	12.9	14.5	14.9	15.0	15.0	15.0	15.0	15.0	15.0	28.7
DAB	5.1	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	1.6
LOC	16.8	18.3	18.5	18.6	18.8	18.9	18.9	18.9	18.9	19.8
MISC	12.9	14.4	14.6	14.7	14.7	14.7	14.7	14.7	14.7	10.2
ORG	9.6	10.4	10.4	10.4	10.3	10.3	10.3	10.3	10.3	10.8
PER	25.8	26.7	26.8	26.8	26.8	26.8	26.8	26.8	26.8	16.6

TABLE 5.8: Final class distribution of Wikipedia articles at each stage of bootstrapping; also the distribution of incoming links

4% F score when removing the capitalisation heuristic also suggests our approach's applicability to classification schemes and languages where capitalisation cannot be relied upon.

The results of ten-fold cross-validation are shown in Table 5.6(a), and are compared with the state-of-the-art performance reported by Dakka and Cucerzan (2008) in Table 5.6(b). The two methods perform similarly, although the distribution of classes is not closely comparable. Our classification precision is slightly higher, with a lower recall particularly in the broad ORG, MISC and NON classes. Our approach implies a lower recall since it relies on features not present in all articles. If we consider only entity classes, we achieve a micro-averaged recall of 87%, and 92% without the poorly-defined MISC category.

The final article classifications used in the rest of our work were seeded by the entire set of hand-labelled articles. In Table 5.8 we see that while the class distribution is fairly stable after three bootstrap feedback

iterations, a marginal portion of unknown articles is able to be classified with more feedback iterations. The proportion of articles with unknown class stabilises at 9.4% in the fourth feedback iteration. We use the predicted classifications which follow the fifth feedback iteration in the rest of our present work.

We mapped the targets of all links in Wikipedia to these classes to determine their distibution over the set of classes, as shown in the rightmost column of Table 5.8.[7] For some classes, the distribution of incoming links for each entity class differs drastically from the number of articles. For example, links to disambiguation articles rarely appear intentionally in text; many PER entities are covered by Wikipedia, but relatively few are popular.

5.6 Conclusion

We have presented an article classification approach which uses structural-semantic features of Wikipedia articles in a semi-supervised method, achieving higher overall precision (94%) than reported elsewhere for similar tasks, and high recall on entity classes (87%). Our use of semantic features reduces classification errors that might occur in a bag-of-words approach. Bootstrapping takes account of the broad terminology inherent in lexical-semantic tasks by spreading semantic knowledge to unseen features and thus incorporating unlabelled data in subsequent classifications. Unlike rule-based classification found in the literature, e.g. Richman and Schone (2008), our classification method may easily be reapplied to fine-grained classification schemes, or those considering a specific entity domain. We use article classifications produced by this method to transform links in Wikipedia articles into entity annotations, and discuss the results of this process in the following chapters.

[7] This data consists of 90.9% of all links; the other 9.1% link to non-existant articles, sections within articles or other unclassified targets.

CHAPTER 6

Initial Results and Extensions

We evaluate our method for deriving NE-annotated corpora from Wikipedia as it has been described heretofore and find it performs poorly. After analysing some of the problems in our corpus, we detail methods to improve NER corpora derived from Wikipedia.

6.1 Data

Our classification data and NE-annotated corpora are derived from the English Wikipedia snapshot dated 22 May 2008. It includes 4,894,828 pages in the main namespace (see section A.1), 2,593,996 of which were identified as redirects, leaving 2,300,832 articles. Using the processing described in section 4.2, we extracted sentences with links comprised of 758 million tokens. Statistics on the result of this process are shown in Table 6.1.

	Articles	Sentences	Tokens	Links
Total	2,161,722	32,253,055	758,352,596	37,859,007
Per article		14.9	350.8	17.5
Per sentence			23.5	1.2

TABLE 6.1: Size and statistics of the sentential data extracted from the 22 May 2008 Wikipedia snapshot

As with most of the Wikipedia-derived corpora used in our experiments, we use a baseline corpus of approximately 3.5 million tokens. Sentences are selected sequentially (as specified in section 4.3) from an arbitrarily-ordered collection of Wikipedia articles. When the corpus reaches 3.5 million tokens or more, the remaining sentences are discarded. This ensures that corpora contain complete sentences while being of approximately equal size. Precise statistics for the Wikipedia-derived baseline corpus (WPB) are given in Table C.1.

(a) MUCEVAL scoring

TRAIN	With MISC		No MISC		
	CONLL	BBN	MUC	CONLL	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
WPB	51.0	63.0	65.0	52.8	65.2

(b) EXACT scoring

TRAIN	With MISC		No MISC		
	CONLL	BBN	MUC	CONLL	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
WPB	37.1	45.7	52.7	39.6	51.4

TABLE 6.2: DEV F score performance for gold-standard corpora and the WPB baseline Wikipedia corpus

6.2 Results and analysis

Performance of the baseline Wikipedia-derived training corpus is compared to gold-standard corpora in Table 6.2. The Wikipedia corpus consistently produces poorer performance than manually-annotated training corpora. We analyse it using the methods described in section 3.3.

N-gram tag variation. Examples of the most frequent nucleus variations are shown in Table 6.3.[1] Some are similar ambiguities to those found in gold-standard corpora (Table 3.4), such as the labelling of premodifiers (the, United States ..., etc.) and nouns which entities modify (metropolitan area, language). Compounds such as Serbia and Montenegro are similarly ambiguous. Others common inconsistencies include the incorporation of trailing punctuation or possessive 's in a link, and the fact that an editor may link the text Burlington County , New Jersey, or may link Burlington and New Jersey separately. Variation in French results from links to two Wikipedia articles: French people and France.

Entity type frequency. We compared the frequency of Wikipedia-derived entity patterns with BBN and CONLL corpora in Tables C.6 and C.7, and found a number of results. WPB has:

- very few all-caps entities, especially compared to CONLL which uses uppercase headings;
- many more entities of form X of Y than either gold standard;
- no abbreviated locations (e.g. Calif.) although 3.5% of BBN LOCs are of this form;
- 2.2% of WPB LOCs are of the form X , Y while there are none in BBN or CONLL;
- many adjectives (e.g. Australian) classified as LOC which gold standards would mark MISC;
- many more entities that are proper nouns followed by common nouns than in other corpora;

[1] The example n-grams may not commonly vary, but they are instances of nuclei which vary often.

N-gram	Tag	Freq	Tag	Freq
County **,** New	O	25	LOC	168
, **the** Netherlands ,	O	40	LOC	12
Shakespeare **'s** play	O	12	PER	2
Academy Award **for** Best Actress	O	2	MISC	43
Trinity College **,** Cambridge	O	2	ORG	15
During **the** First World War	O	11	MISC	2
in England **and** Wales	O	12	LOC	18
The New York Times	O	11	ORG	3
the **French** ,	LOC	10	MISC	2
of **the** Rolling Stones	O	4	ORG	3
Earth **'s** surface	O	14	LOC	1
of **the** Pope	O	3	PER	1
the State **of** Colorado	O	1	LOC	6
the English **language** ,	O	1	MISC	19
by the **United States** Army	LOC	1	ORG	12
Romulus **and** Remus	O	1	PER	10

TABLE 6.3: Examples of frequent nucleus variations in our baseline corpus. Nuclei are in bold.

- many few entities with pattern Aaa-aa (e.g. [LOC London-based]) which is common as LOC in BBN and as MISC in CONLL. Our Wikipedia tokenisation splits hyphens from links;
- like CONLL and MUC, our data has no instances of MISC with pattern A., while BBN uses R. and D. to abbreviate Republican and Democrat.

Tag sequence confusion. The confusion matrices in Appendix C highlight numerous annotations where WPB differs both from gold-standard data and from the predictions of standard-trained models, indicating their relative impact on performance. WPB's prominence in the PRED-CORR matrices (Tables C.8 and C.10) suggests that it tends to predict much larger entities than in the evaluation data. In particular, it regularly marks many non-entities as entities, and has many instances of tagging LOC O LOC (e.g. "[LOC London] , [LOC Ontario]") or LOC O as a single LOC entity. In CONLL, where sports score data often follows names, these are falsely included in entity tags. While BBN would split a product name like Toyota Corolla into two entities (ORG and MISC), WP2 tags it as a single entity. BBN's many ORGs with possessive suffixes are also not tagged correctly.

On the CORR-PRED matrices (C.8 and C.10), most PER entities are shown as correctly tagged, but PRED-CORR shows that around half of these had oversized boundaries, being O PER in the gold-standard data (e.g. President Nixon, Mr. Jones). We also see that organisations ending with abbreviations (e.g.

Inc.) are often tagged without their full-stops. Very few BBN MISC entities are identified, most being misclassified as LOC (e.g American), and many ORGs are tagged LOC, while others are left untagged at the beginning of sentences.

This presents a wide variety of errors and inconsistencies in our data, many of which may be valid NE annotations, but do not conform to gold standards.

6.3 Method extensions

Motivated by this error analysis, we present a number of modifications to our corpus derivation process.

6.3.1 Handling conventionalised capitalisation

We provide additional methods to handle non-entity text that is capitalised in English (see section 4.3.2).

First words. Having found that too many sentence-initial words were marked as entities, we extend the list of words allowed unlabelled at the start of a sentence with a collection of words which are found lowercase sentence-internally more than t_1 times and also appear at the start of sentences more than t_2 times in our Wikipedia data. With $t_1 = t_2 = 50$ we extracted 1500 such words from 1.5 million sentences, including a much higher proportion of content nouns than in the 285 frequent sentence starters previously used. Together, 1520 words are admitted unlinked when sentence-initial.

We also loosened our definition of *first word* for this purpose to include words appearing after some punctuation marks (left parenthesis, quote, colon).

Dates. Date expressions are identified using regular expressions for inclusion in the corpus. Additionally, if a date is linked to a Wikipedia article, the link is removed. For single-word links, we require that the link target is identical to the link text, admitting for instance a link labelled Wednesday to the article on fictional character Wednesday Addams.

Acronyms. Although sentences with capitalised links to NON articles are generally discarded, we make an exception for all-caps links, assuming they are acronyms.

Pronouns. We allow the inclusion of personal pronoun I in the corpus.

Personal titles. Personal titles (e.g. Brig. Gen., *Her Majesty*, Prime Minister-elect, Playmate of the Year) are capitalised in English, although not considered entities by gold-standard corpora.BBN *marks personal titles by* PER_DESC*, but it also applies this tag to numerous other descriptors of people.* Titles are sometimes linked in Wikipedia, but the target articles, e.g. US President, are found in Wikipedia categories like Presidents of the United States, causing their incorrect classification as PER.

Noting examples of the form [PER President] [PER Barack Obama]), we initially implemented a trivial solution which marks links immediately preceding PER entities as titles, leaving them unannotated in the final corpus. This approach increased performance on CONLL DEV by 7% F in one experiment, but since it only includes titles adjacent to PER entities and incorrectly marks other entities as titles, performance was further improved by a more elaborate approach.

We have extracted the text of links which frequently precede PER entities, forming 774 entries which were edited to remove non-titles and to strip titles to their basic form (e.g. US Vice President becomes President). We thus acquire a list of 365 titles, to which we append 19 abbreviated titles found by searching the BBN corpus for tokens labelled PER_DESC and ending in full-stops. We also list 11 title prefixes (e.g. then-, Deputy) and 3 suffixes (e.g. Emeritus, -Elect). This data together is used to form a regular expression which marks titles in text (and as link text) allowing their un-annotated inclusion in a corpus.

6.3.2 Emending the corpus

In order to conform to the annotations of gold-standard evaluation corpora, a number of changes need to be made to our default method of transforming links into entity annotations. Rarely, it is useful also to modify the text of the Wikipedia-derived corpus.

Adjusting entity 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: always for LOC, and for ORG and PER if the link does not end in a full stop;
- possessive markers ('s or ' when labelled as possessive by the POS tagger);

- other punctuation (comma, semicolon, hyphen, etc.)

Personal title links. The method for finding personal titles above does not directly apply to links whose text includes a title following an organisation or location name (e.g. Microsoft CEO, Australian Prime Minister), or preceding a personal name (President Barack Obama). We thus use the personal title regular expression described above adjust boundaries for links of form:

(1) TITLE PER

(2) LOC/ORG TITLE

(3) TITLE of/for LOC/ORG

For (1) we search for PER entities whose prefixes match. For (2) and (3) we allow any class of entity, but require that the mentioned entity can be identified as ORG or LOC.[2] If not, the sentence is discarded as likely noise, although this also discards entities such as Burger King.

Adjectival forms. Adjectival forms of entity names, such as Australian or Islamic, are capitalised in English. While these are not exactly entities, both the CONLL and BBN corpora tag them (BBN entitles this class NORP). Since the text Australian would usually link to Australia in Wikipedia, which is classified LOC, our baseline system simply annotated adjectival names like their nominal forms.[3] An initial solution POS tagged the corpus and marked all adjectival (tag JJ) entity annotations as MISC, improving performance but missing instances where nationalities are used as nouns, e.g. five Italians.

Like with personal titles, we use this initial solution to create a list-based alternative with greater coverage. We extracted common 339 adjectival LOC and ORG entities from the corpus. Entity annotations whose text have a match on this list, optionally with a plural -s suffix, are emended to MISC, irrespective of POS tag. This unfiltered list includes some errors, e.g. First, Emmy, and others where MISC is only occasionally the appropriate tag, e.g. the Democrats, an ORG, compared to Democrat leader, a MISC.

State abbreviations. A gold standard may use stylistic forms which are rare in Wikipedia, notably abbreviations. For instance, the Wall Street Journal (BBN) regularly uses US state abbreviations, while Wikipedia nearly always refers to states in full. We boosted our corpus performance by randomly selecting 1/3 of appearances of US state names in Wikipedia to substitute with their standard abbreviated forms.

[2] We use the alternative title trie described in the next section or check the title article's classification directly.

[3] Note that this too is a reasonable annotation scheme, but it is not what is used by our gold-standard evaluation corpora.

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 exclude personal names including of which are much more frequent in Wikipedia than in newswire corpora.

Fixing tokenisation. 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 system had split hyphenated portions from entities. Fixing this to match the BBN approach improved performance significantly. Note that the MUC corpus we use splits such hyphenated expressions into three tokens.

6.3.3 Inferring additional links

Wikipedia editors commonly only the link first appearance of an entity named within an article. Relying on links to produce entity annotations therefore introduces a number of sentence selection biases to our corpus:

Short sentences: Longer sentences mention more entities, and every capitalised entity mentioned has to satisfy the Confidence Criterion for sentence selection (see section 4.3);

Low repetition: The same entity is unlikely to appear in multiple sentences in an article, reducing the significance of in-document features of NER learning systems (see section 2.1.2);

Early sentences: Sentences appearing later in Wikipedia articles tend to have much sparser links;

Full names: First mentions of entities are rarely abbreviated forms (e.g. surname, acronym, etc.).

Relying on links also implies that only a short amount of text from each Wikipedia article will be selected (WPB has an average of 71 tokens per article, many times fewer than gold-standard corpora; see Table C.1), such that NER machine learner features related to in-document coreference are less likely to be activated than in real-world texts.

We increase coverage of sentences and entity name forms by inferring additional links in articles, as others processing Wikipedia have done (Metaweb Technologies, 2008; Richman and Schone, 2008). To ensure the accuracy of the links inferred, we only produce additional links to targets already linked from

Type	Alternative titles
1	007; Agent 007; Bames jond; Bond; Bond , James; Bond , James Bond; Bond 007; Bond Extreme; Bond film; Bond movie; Bond. James Bond .; Commander Bond; Double Oh Seven; James Blonde; James Bond; James Bond 007; James Bond formula; James bond; James bonf; Mr bond; OO7; Sir James Bond
2	Bonded; Bonding; Bonds; JB; Jb; Tax bond
3	James
4	007 / Bond car; Another role; Bond 's; Bond Villain; Bond films; Bond movies; Bond-esque; Double naught spy; Ian Fleming 's James Bond 007; J. Bond; James Bond 's; James Bond (007); James Bond (Video game); James Bond movies; James Bond novels and films; James Bond-style; James Bondian; James Bonds; Jimmy; Jimmy Bond; Mr. Bond; O-O-Seven; Rosh Hashanah; Ross Hashanah; Sir James Bond 007

TABLE 6.4: Alternative titles for the article James Bond

an article. Our method for link inference involves first compiling a list of *alternative titles* for each Wikipedia article. Then, given an article A^*, we infer new links by:

(1) determining the targets of all outgoing links from A^*;

(2) compiling a prefix trie of alternative titles of all articles linked from A^*;

(3) matching strings in A^* to trie entries, and labelling them with the corresponding link target.

6.3.3.1 Collecting alternative titles

Alternative titles for an article A include:

Type 1: The title of A and those of redirects to A;
Type 2: The titles of disambiguation pages linking to A;
Type 3: The first or last word of A's title if A is of class PER;
Type 4: The text of all links whose target is A.

Examples of each type of alternative title are given in Table 6.4. The ordering of the types of alternative articles is intended such that more reliable titles are of a lower-numbered type; in the case of duplicates, the lower type is preferred. Article titles and redirects often include disambiguation information, in parentheses (e.g. Saturn (mythology)) or following a comma (e.g. London, Ontario), and forms with and without this additional information are considered as alternative titles in each case.

Type 2 uses disambiguation pages (see section A.1) as a source of alternative titles, and is particularly important for obtaining acronyms and abbreviations (the disambiguation page AMP links to AMP Limited

and Ampere), and given or family names (Howard links to Howard Dean and John Howard). Since Wikipedia links usually apply to the first mention of an entity, the text is usually a disambiguated form, hence inferring links for abbreviated forms is essential to improving sentence coverage. It is nonetheless difficult to accurately determine appropriate disambiguation page titles, since disambiguation pages are not consistent in their format, and since redirects to disambiguation pages need to be considered in addition to their own titles, which adds, for instance, Bonded and Tax bond to the alternative titles for James Bond. 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. Valid targets the do not appear at the beginning of list items are not discovered by this approach; and linked terms at the beginning of list items on disambiguation pages may be included though not relating to the disambiguated term. The impact of errors acquired through processing disambiguation pages is dampened by the matching process below.

Initially, alternative title Type 3 was intended to cover given and family names, and improved performance when Type 2 was not in use. With disambiguation page data available, this type of title is largely redundant. The text used to link to an article is often a name for that article's referent. Link texts are often subsumed by lower title types, but some additional useful forms might be obtainable. We see that in the case of James Bond, Type 4 titles are largely noise.[4]

Because of the loss or gain that each type of alternative title may introduce to link inference, we report experiments incorporating each additional level in the following chapter.

6.3.3.2 Matching titles

Each string beginning with a word containing at least one capitalised letter is considered a candidate for link inference. The longest sequence of words which matches an appropriate article title is inferred as a new link. When inferring links in an article A^*, a trie maps each alternative title of A^* and A^*'s outgoing links to its associated article with the following constraints:

[4]The strangest example, perhaps, is Rosh Hashanah. The Wikipedia article on The Sack (Robot Chicken episode) describes a parody of the James Bond film *Goldfinger*, in which Ross Hashanah is the name of a character modelled after James Bond. In the article, Ross Hashanah is linked to James Bond, as is Rosh Hashanah, a typographical error.

Sentence type	% of sentences
Contain no entities	21.0
Linked entities alone	6.0
+ Type 1	9.3
+ Types 1-2	2.2
+ Types 1-3	3.2
+ Types 1-4	2.4
Fail confidence criterion	55.8

TABLE 6.5: The increased sentence coverage due to link inference

(1) If T is a title of both articles A_1 and A_2:
   (a) if A_1 or A_2 is A^* then the other is ignored;
   (b) if $Type(T, A_1) < Type(T, A_2)$ then T is mapped to A_1;
   (c) if $Type(T, A_1) = Type(T, A_2)$ then T is marked as being conflicted.

(2) A title is not added to the trie if it matches one of the boundary adjustment rules above.

(3) A title T is not added to the trie if the following conditions are all met:
   - T and T' are titles of A;
   - $Type(T', A) \leq Type(T, A)$;
   - T is the concatenation of T' and some R; and
   - R is an all-lowercase string.

Constraint (1) handles conflicts when the same title refers to two candidate articles, prioritising the target article and titles of lower Type. The second constraint avoids matching strings like London , Ontario when London is a better match. Constraint (3) attempts to avoid added non-entity words, as in James Bond novels and films when James Bond would be a better match.

We performed some experiments with not requiring the first word of inferred links to include a capitalised letter, but found this introduced an excessive amount of error.

Table 6.5 shows that using link inference, we are able to increase our coverage of Wikipedia sentences selected for corpora from 6% to over 23%. It also increases the average number of tokens per article in the derived corpus from 82 to 506 (see Table C.1).

CHAPTER 7

Results and Discussion

We present results from experiments to determine:

(1) the performance increase due to changes described in Chapter 6;
(2) the effect of types of alternative titles for link inference (see section 6.3.3);
(3) TEST results for our best-performing corpus;
(4) the use of Wikipedia in addition to gold-standard training corpora;
(5) the impact of corpus size on performance;
(6) the variability of Wikipedia corpora of constant size; and
(7) whether Wikipedia-derived annotations are predictable.

We conclude by analysing the remaining errors produced by NER models trained our Wikipedia corpus.

7.1 Performance

Table 7.1 shows the performance of six Wikipedia-derived corpora, each of approximately 3.5 million words. Using a corpus derived from only Wikipedia's explicit links (WP0), we are able to achieve an EXACT F-score performance 17-26% above that of the baseline corpus (WPB). Corpora WP1-4 cumulatively incorporate additional types of inferred link alternative titles (see section 6.3.3). The performance impact of Types 1 (titles and redirects) and 2 (disambiguation titles) is consistently positive. Type 3 (first and last PER names) changes results only marginally, while Type 4 (link texts) clearly reduces performance. Note though, that since varying link inference modifies which sentences are selected for the training corpus, the performance difference between corpora is not necessarily a direct result of the inference of additional links. Link inference adds 4-8% to WP0's EXACT F-score, resulting in a 24-32% improvement on the baseline. Although WP2 and WP3 performance is very similar, we select WP2 as our best performing corpus and use it primarily for future experiments.

(a) MUCEVAL scoring

TRAIN	With MISC		No MISC		
	CONLL	BBN	MUC	CONLL	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
WPB	51.0	63.0	65.0	52.8	65.2
WP0: no inf	71.0	79.3	76.3	71.1	78.7
WP1	74.9	82.3	81.3	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

(b) EXACT scoring

TRAIN	With MISC		No MISC		
	CONLL	BBN	MUC	CONLL	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
WPB	37.1	45.7	52.7	39.6	51.4
WP0	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	66.2	72.3	75.6	67.3	73.3

TABLE 7.1: DEV F-score performance for gold-standard corpora and Wikipedia-derived corpora with varying levels of link inference. Pale blue cells mark corresponding corpora; pink cells mark the best performing Wikipedia corpus for each evaluation.

(a) MUCEVAL scoring

TRAIN	With MISC		No MISC		
	CONLL	BBN	MUC	CONLL	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

(b) EXACT scoring

TRAIN	With MISC		No 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
WP2	60.9	69.3	76.8	61.5	69.9

TABLE 7.2: TEST F-score performance for gold standard corpora and WP2

Our Wikipedia data also performs better than the best cross-corpus result with gold standard training data by up to 12% EXACT F-score (on BBN with MISC). Using the BBN model on MUC is the only cross-corpus pair that is not outperformed by our Wikipedia corpora. This is a key result of our work: while CONLL and BBN data perform poorly on each other, a Wikipedia-derived corpus achieve better performance on either CONLL or BBN. In TEST evaluation (Table 7.2), this result is retained, although by a smaller margin: we outperform cross-corpus evaluation by up to 7% EXACT F-score, 5% less than on DEV. This suggests, perhaps, that our corpus has been somewhat overfitted in design to DEV data. The result of outperforming other cross-corpus evaluation pairs nonetheless shows Wikipedia's viability as a source of automatically-annotated NER training data.

A class-by-class breakdown of WP2's performance on the TEST data is given in Table 7.3. Performance is particularly low for ORG on CONLL, most likely because it often refers to sports teams, labelled ORG, by their place of origin (e.g. Milan coach Oscar Tabarez), as suggested by the correspondingly low

(a) MUCEVAL scoring

Class	With MISC		No MISC		
	CONLL	BBN	MUC	CONLL	BBN
LOC	69.2	79.4	80.5	69.0	80.8
MISC	58.0	56.9	—	—	—
ORG	43.0	74.4	74.8	40.0	68.7
PER	82.0	79.3	90.8	80.6	82.0

(b) EXACT scoring

Class	With MISC		No MISC		
	CONLL	BBN	MUC	CONLL	BBN
LOC	65.8	75.9	77.9	66.4	77.1
MISC	52.7	53.3	—	—	—
ORG	36.8	65.1	70.3	33.9	60.2
PER	78.2	77.5	87.0	77.0	80.9

TABLE 7.3: TEST F-score performance for WP2 by class. See Table C.5 for comparison.

(a) TYPE scoring

Corpora		With MISC		No MISC		
Training	Eval.	CONLL	BBN	MUC	CONLL	BBN
TRAIN	DEV	90.9	91.1	89.0	90.8	91.5
TRAIN + WP2	DEV	91.7	91.2	90.6	91.5	91.9
TRAIN	TEST	87.8	91.1	81.0	87.9	91.9
TRAIN + WP2	TEST	87.6	91.2	83.6	88.1	91.2

(b) EXACT scoring

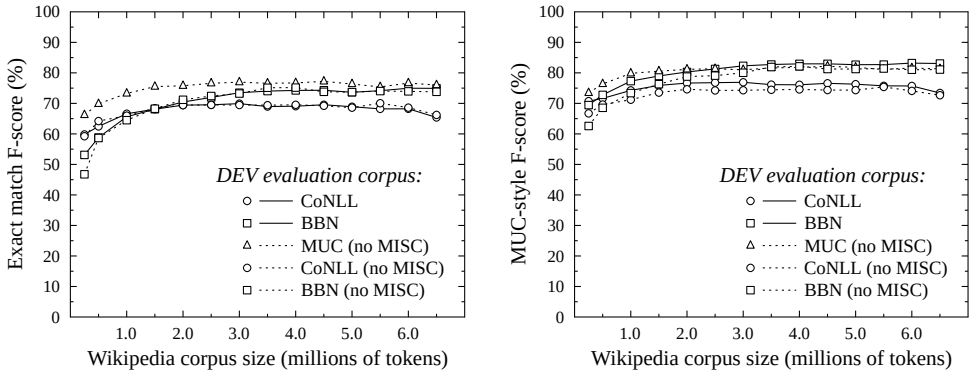
With MISC		No MISC		
CONLL	BBN	MUC	CONLL	BBN
85.9	86.5	82.3	86.9	88.0
87.6	86.9	87.4	88.4	88.7
81.2	86.7	73.5	82.1	88.4
81.5	87.2	79.5	83.4	87.9

TABLE 7.4: F-score performance for Wikipedia as additional training data: rows marked TRAIN give scores for evaluating e.g. CONLL TRAIN on CONLL DEV; each cell immediately below shows the impact of training with Wikipedia data in addition to the appropriate TRAIN corpus

performance for LOC due to many false positives. For both BBN and CONLL data, MISC performance is low due to the breadth of the category and the different distribution of entities represented in Wikipedia as compared to newswire corpora.

We have also considered using a Wikipedia-derived corpus as *additional* training data, and evaluate the performance when WP2 is added to training for traditional TRAIN-DEV pairs. The Wikipedia data dominates the manually-annotated corpus to which it is added. Table 7.4 shows that although performance improves in most cases, this improvement is marginal except on MUC which is handicapped by a very small training corpus. The variability of Wikipedia-derived corpora (see the next section) suggests that marginal improvements are meaningless, and hence that our corpora are not able to overcome the advantage of training and evaluation data having the same source.

Complete tables of overall performance statistics are given in Appendix C.



(a) EXACT scoring (b) MUCEVAL scoring

FIGURE 7.1: Effect of changing Wikipedia corpus size on performance

(a) σ with outlier

TRAIN	MUCEVAL			EXACT		
size	MUC	CONLL	BBN	MUC	CONLL	BBN
3.5M toks	1.6	2.0	1.4	2.1	2.7	2.1
300K ents	1.4	1.5	1.6	1.7	2.1	2.4
150K sents	1.8	1.7	1.6	2.0	2.2	2.5

(b) σ without outlier

MUCEVAL			EXACT		
MUC	CONLL	BBN	MUC	CONLL	BBN
1.7	2.0	0.8	2.2	2.6	1.2
1.5	1.4	0.9	1.5	1.5	1.3
1.9	1.8	0.9	2.1	2.1	1.2

TABLE 7.5: DEV F-score standard deviations for ten Wikipedia training corpora of the same size. All evaluations are without MISC.

7.2 Wikipedia variability

We have also examined performance of Wikipedia-derived data independent of gold-standard sources.

As shown in Figure 7.1 we have constructed Wikipedia corpora from 0.25 to 6.5 million tokens to assess the effect of size on performance (full statistics in Table C.4).[1] While it was plausible that increased Wikipedia knowledge would increase performance, we find it plateaus—and for CONLL evaluation even degrades—for more than 3-4 million tokens of training data. This is likely due to unresolvable inconsistencies between Wikipedia and gold standard data, although it may also represent constraints of the machine learning process and over-training.

In another experiment, we produced ten Wikipedia-derived corpora of approximately the same size, to determine variability due to different selections of Wikipedia text. Fixed size could be defined by holding the number of tokens, named entities, sentences, or even articles constant: which gives the lowest variance? Our results in Table 7.4(a) suggest surprisingly that for MUC and CONLL, keeping entities or

[1]The smaller corpora are strict subsets of the larger corpora; the 3.5M-token corpus is WP2.

(a) MUCEVAL scoring

TRAIN	With MISC		No MISC	
	$\bar{x}$	σ	$\bar{x}$	σ
MUC	—	—	67.7	1.2
CONLL	71.7	1.7	70.8	1.5
BBN	73.2	1.1	72.5	1.3
WP2	89.5	0.7	89.1	0.8

(b) EXACT scoring

TRAIN	With MISC		No MISC	
	$\bar{x}$	σ	$\bar{x}$	σ
MUC	—	—	54.2	1.9
CONLL	55.6	2.8	59.6	2.3
BBN	57.7	1.8	61.4	1.8
WP2	83.6	1.2	86.1	1.0

TABLE 7.6: F-score performance with ten folds of WP2-like test data

sentences constant produced less variability in performance than tokens. In each case, we found that one of the ten corpora gave outlying low performance on BBN DEV. Without this outlier, EXACT F-score standard deviation ranged from 1.2 on BBN to 2.6 on CONLL as shown in Table 7.4(b). This variance is small enough to consider our overall performance against gold standard corpora significant.

Finally, we investigate how predictable WP2's annotations are. If tested on same-similar data, can we achieve high results similar to e.g. BBN on BBN? And is cross-corpus performance similarly low with Wikipedia test data? We produced ten WP2-like non-overlapping 150,000-token test corpora. Table 7.6 seems to affirm both our queries, although WP2 performance on Wikipedia-derived test data is 2.3 to 2.9 EXACT F-score lower than CONLL on CONLL or BBN on BBN with MISC included. Wikipedia data is therefore only a little less predictable than human annotation, especially for the broad MISC category.

7.3 Error analysis

Corpus analyses all clearly show the effect of changes to our baseline system as reducing the difference between Wikipedia and gold standard corpora, but also introducing some new errors.

Top tag variations in WP2 are shown in Table 7.7. Like WPB, some result from variations in linking and the choice of target article in context, such as Rome, England and EU. Link inference has also introduced some errors such as the labelling of To as a location; allowing non-entity acronyms has admitted US without annotation. Variations Jackson and Batman do not indicate corpus inconsistency, being valid reference ambiguities. Absences when compared to Table 6.3 are also notable.

The entity type frequency data summarised in tables C.6 and C.7 shows WP2 at times retaining data distributions similar to WPB, and at others showing progress towards the distributions of entities in gold standard data, such as increases in all-caps entities and relative decreases of full forms of names due to

N-gram	Tag	Freq	Tag	Freq
, **the** Netherlands ,	O	20	LOC	7
In **Europe** ,	LOC	17	ORG	3
in England **and** Wales	O	5	LOC	13
The Canterbury Tales	O	2	MISC	1
Award **for** Best	O	5	MISC	26
in **Rome** .	LOC	28	MISC	4
former **England** captain	LOC	1	ORG	1
of **the** Book of	O	38	MISC	1
To this end	O	20	LOC	1
the **EU** and	LOC	1	ORG	16
Jackson was	LOC	2	PER	7
Batman 's	MISC	1	PER	11
the **US** dollar	O	3	LOC	6

TABLE 7.7: Examples of frequent nucleus variations in WP2. Nuclei are in bold.

Token	Corr.	Pred.	Count	Why?
.	ORG	O	90	Inconsistencies in BBN
House	ORG	LOC	56	Article White House is classified LOC in bootstrapping
Wall	O	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	O	26	A difficult NER ambiguity in e.g. Robertson , Stephens & Co.
's	ORG	O	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	O	MISC	20	Rare in Wikipedia, and inconsistently labelled in BBN
D.	MISC	PER	14	BBN uses D. to abbreviate Democrat

TABLE 7.8: Tokens in BBN DEV that the WP2 model frequently mislabels

link inference, and a non-negligible number of LOC entities with abbreviated forms, although not yet approaching BBN. From POS tag types we see an increase in miscellaneous adjectives and plural proper nouns in accordance with our handling of nationalities, bringing proportions closer to BBN and CONLL.

Tag sequence confusion PRED-CORR matrices (C.9 and C.11) highlight a great reduction in oversized entity boundaries from WPB to WP2, and fewer non-entities labelled as entities. CORR-PRED tables (C.8 and C.10) show that this co-occurs with frequently leaving gold standard entities untagged. They also highlight a new excess of undersized entity annotations, such as leaving common nouns at the beginning of ORG names untagged, or splitting them into two entities at punctuation or function words. Errors retained include: including . after an ORG where BBN does not; and marking ORG-product compounds as single ORG or MISC entities where BBN tags ORG MISC.

Finally, Table 7.8 summarises analysis of the BBN DEV tokens most often mislabelled by WP2 data.

Extensions to our corpus-derivation process still retain and introduce tagging errors and inconsistencies with NER tagging schemes. As we have shown in section 3.4, gold-standard corpora are also divided by schema inconsistencies which dramatically reduce cross-corpus performance. Our experiments and analysis show that our method for automatically annotating Wikipedia for NER produces fairly stable corpora that are competitive with gold-standard data.

CHAPTER 8

Conclusion

8.1 Future work

Our results, while indicating Wikipedia's viability as NER training data, suggest many areas for extension and exploration. Individual components of our method could be improved, including article classification, selection criterion exceptions, and link boundary emendations. A more thorough and meaningful evaluation scheme could be developed to ensure the value of new features, such as finding the average gain over multiple folds of Wikipedia training data. The models could also be evaluated on genre-appropriate test data, such as manually-annotated Wikipedia articles or web content. We could also attempt to measure corpus accuracy, either through manual assessment, or by developing a statistical measure of inconsistency.

8.1.1 Different domains

The breadth of topics covered in Wikipedia suggests its applicability to much broader named entity types than those considered presently (ORG, LOC, PER). It may be possible to apply our method to fine-grained entity hierarchies (e.g. Sekine et al. (2002); Brunstein (2002)), possibly also incorporating their many non-proper noun classes (e.g. animals). These hierarchies also include many types of entities that are much more common in Wikipedia than in the news text commonly used as gold-standard corpora, such as works of art. Many such entities, e.g. When Harry Met Sally..., also have a more complicated linguistic structure than personal or location names, making them difficult to identify with traditional NER systems. Recognising a breadth of entity types is particularly useful for question answering.

Our approach may also be applicable to producing training data for more specific NER domains, e.g. biomedical literature. Feasibility is dependent on sufficient coverage of a domain in Wikipedia. We have begun some experiments into using our method for astronomy texts (see Murphy et al. (2006)).

Finally, Wikipedia is currently available in 132 languages, twenty-three of which have over 100,000 articles. As Richman and Schone (2008) have suggested, Wikipedia-derived corpora could produce NER models in languages which are otherwise poor in NLP resources.

Note that with any of these domain variations, capitalisation may no longer indicate entities, and article classification and sentence selection approaches would need to be adjusted accordingly.

8.1.2 Using more of Wikipedia

Currently we are only able to utilise a small fraction of our Wikipedia-derived training data, due to resource constraints for machine learning. Portions currently not used in training may include useful entity examples for real-world application. This could be solved by:

- parallelising the machine learner;
- building models from many small Wikipedia corpora and combining them through voting or some other mechanism;
- selecting only useful sentences, e.g. by removing repeated or formulaic sentences, or only using text from articles that meet quality criteria.

8.1.3 Wikipedia as additional data

We reported results of an experiment where we naively added Wikipedia training data to gold-standard corpora to only marginal effect. Many other approaches have been considered for combining training sources of different genres or domains and adapting them for a new task (e.g. Daumé III (2007)). Using such domain adaptation techniques may overcome inconsistencies between Wikipedia and manually-annotated corpora and hence better illustrate the advantage of using Wikipedia text.

An alternative approach may augment gold-standard corpora with only the Wikipedia sentences that would be most beneficial. For example, the result of running a CoNLL-trained NER system on Wikipedia could be compared to automatically-derived annotations. Sentences which the system labels incorrectly could then be selected for automatic addition to the training data in a bootstrapping process.

8.2 Conclusion

We have presented a method for transforming Wikipedia into annotated corpora from which a system may learn patterns useful in the identification and classification of names. Such corpora are able to significantly outperform human-annotated training data when tested on an evaluation corpus from a different source, emphasising the feasibility of learning NER from Wikipedia data.

Our work has three main contributions:

The first concerns the role of training corpora in NER performance, where we have identified significant performance losses when NER training and evaluation data are from different sources. This led us to develop methods for analysing corpus inconsistencies, comparing corpora and finding common errors caused by the training data, which are complicated by the phrasal nature of the NER task.

Secondly, we have presented an approach to classifying Wikipedia articles using their structural features and have used its sentential content as training data for Named Entity Recognition. Our results suggest that despite controversy surrounding error in Wikipedia, its content may be harnessed productively. To do so we had to systematically emend some data to conform to existing standards, illustrating the flexibility of data automatically acquired from semi-structured sources. Hence our work promotes further use of Wikipedia and other online resources in acquiring natural language knowledge.

Finally, we are able to provide a new training corpus for NER, having experimentally shown the viability of Wikipedia-derived corpora. While a BBN-trained model tags our introductory example as

> [ORG Paris Hilton] visited the [ORG Paris Hilton].

a NER model learnt from Wikipedia's contemporary knowledge instead produces:

> [PER Paris Hilton] visited the [LOC Paris Hilton].

Moreover, the corpus is huge (over 100 million tokens), free, broad in its domain coverage, and can be automatically regenerated as Wikipedia grows. These features make it a useful resource for a wide range of NLP applications.

Bibliography

David Ahn, Valentin Jijkoun, Gilad Mishne, Karen Müller, Maarten de Rijke, and Stefan Schlobach. 2004. Using Wikipedia at the TREC QA Track. In *Proceedings of the Text REtrieval Conference*.

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.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, volume 4825 of *LNCS*, pages 722–735.

Sören Auer and Jens Lehmann. 2007. What have Innsbruck and Leipzig in common? extracting semantics from wiki content. In *Proceedings of the 4th European Semantic Web Conference*, volume 4519 of *LNCS*, pages 503–517.

Somnath Banerjee. 2007. Boosting inductive transfer for text classification using Wikipedia. In *Proceedings of the 6th International Conference on Machine Learning and Applications*, pages 148–153. IEEE Computer Society, Washington, DC, USA.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Assocation for Computational Linguistics*, pages 26–33.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American*, 284(5):34–43.

Abhijit Bhole, Blaž Fortuna, Marko Grobelnik, and Dunja Mladenić. 2007. Extracting named entities and relating them over time based on Wikipedia. *Informatica*, 31:463–468.

Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An unsupervised approach to biography production using Wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 807–815. Association for Computational Linguistics, Columbus, Ohio.

Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100. Madison, WI.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM*

SIGMOD International Conference on Management of Data, pages 1247–1250.

Andrew Borthwick. 1999. *A Maximum Entropy approach to Named Entity Recognition*. Ph.D. thesis, New York University.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. NYU: Description of the MENE named entity system as used in muc-7. In *Proceedings of the 7th Message Understanding Conference*.

Sergey Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International Conference on World Wide Web*, pages 107—-117. Brisbane, Australia.

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.

Sergey Chernov, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou. 2006. Extracting semantic relationships between Wikipedia categories. In *Proceedings of the First Workshop on Semantic Wikis*, pages 161–171.

Hai Leong Chieu and Hwee Tou Ng. 2003. Named entity recognition with a maximum entropy approach. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 160–163.

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

Nancy Chinchor, Erica Brown, Lisa Ferro, and Patty Robinson. 1999. 1999 named entity recognition task definition. MITRE and SAIC.

Nancy Chinchor and Patty Robinson. 1997. MUC-7 named entity task definition. MITRE and SAIC. `http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html`.

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.

Alessandro Cucchiarelli, Danilo Luzi, and Paola Velardi. 1998. Automatic semantic tagging of unknown proper names. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 286–292.

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.

Silviu Cucerzan and David Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Gaoying Cui, Qin Lu, Wenjie Li, and Yirong Chen. 2008. Corpus exploitation from Wikipedia for ontology construction. In European Language Resources Association, editor, *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco.

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.

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.

James R. Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and boxer. In *Proceedings of the ACL 2007 Demonstrations Session*, pages 33–36. Prague, Czech Republic.

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.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Ludovic Denoyer and Patrick Gallinari. 2006. The wikipedia xml corpus. *SIGIR Forum*, 40(1):64–69.

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.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ace) program — tasks, data, and evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 837–840.

Encyclopædia Britannica, Inc. 2006. Fatally flawed: Refuting the recent study on encyclopedic accuracy by the journal *nature*. `http://corporate.britannica.com/britannica_nature_response.pdf`.

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.

Richard Evans. 2003. A framework for named entity recognition in the open domain. In *Proceedings of Recent Advances in Natural Language Processing*.

Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 168–171.

David Gabay, Ziv Ben-Eliahu, and Michael Elhadad. 2008. Using Wikipedia links to construct word segmentation corpora. In *Proceedings of the AAAI '08 Workshop on Wikipedia and Artifical Intelligence*.

Evgeniy Gabrilovich. 2005. Wikipedia preprocessor (wikiprep). `http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep/`.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1301–1306.

Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature*, 438:900–901.

ILPS. 2007. WikiXML collection. http://ilps.science.uva.nl/WikiXML/.

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.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–415. Association for Computational Linguistics, Columbus, Ohio.

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.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 180–183.

Jon M. Kleinberg. 1999. Authoritative sources in hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

George R. Krupka and Kevin Hausman. 1998. IsoQuest, Inc.: Description of the NetOwl™ extractor system as used for MUC-7. In *Proceedings of the 7th Message Understanding Conference*.

(Linguistic Data Consortium) LDC. 2008. ACE (Automatic Content Extraction) English annotation guidelines for entities, version 6.5.

Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249.

Lucian Vlad Lita, Warren A. Hunt, and Eric Nyberg. 2004. Resource analysis for question answering. In *Proceedings of the ACL-2004 Interactive Posters/Demonstrations Session*, pages 18–21.

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

Christopher D. Manning. 2006. Doing named entity recognition? don't optimize for f_1. NLPers Blog, 25 August. `http://nlpers.blogspot.com/2006/08/`

doing-named-entity-recognition-dont.html.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*, chapter 5: Collocations, pages 151–189. MIT Press.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1995. Treebank-2. LDC95T7, Linguistic Data Consortium, Philadelphia.

Elaine Marsh and Dennis Perzanowski. 1998. MUC-7 evaluation of IE technology: Overview of results. In *Proceedings of the 7th Message Understanding Conference*.

David D. McDonald. 1996. Internal and external evidence in the identification and semantic categorization of proper names. In Bran Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, chapter 2, pages 21–39. MIT Press, Cambridge, MA.

Olena Medelyan and Catherine Legg. 2008. Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense. In *Proceedings of the AAAI '08 Workshop on Wikipedia and Artifical Intelligence*.

Olena Medelyan, Ian H. Witten, and David Milne. 2008. Topic indexing with Wikipedia. In *Proceedings of the AAAI '08 Workshop on Wikipedia and Artifical Intelligence*.

Metaweb Technologies. 2008. Freebase wikipedia extraction (WEX). http://download.freebase.com/wex/.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 233–242.

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.

David Milne. 2007. Computing semantic relatedness using Wikipedia link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference*.

MUC. 2001a. *Message Understanding Conference (MUC) 7*. LDC2001T02, Linguistic Data Consortium, Philadelphia.

MUC. 2001b. MUC-7 tokenization rules 2.0. LDC2001T02, Linguistic Data Consortium, Philadelphia.

Lev Muchnik, Royi Itzhack, Sorin Solomon, and Yoram Louzoun. 2007. Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 76(016106):12 pages.

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

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.

Vivi Nastase and Michael Strube. 2008. Decoding Wikipedia categories for knowledge acquisition. In *Proceedings of the 23rd Conference on Artificial Intelligence*.

Rani Nelken and Elif Yamangil. 2008. Mining Wikipedia's article revision history for training computational linguistics algorithms. In *Proceedings of the AAAI '08 Workshop on Wikipedia and Artifical Intelligence*.

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

Jorge Nocedal and Stephen J. Wright. 1999. *Numerical Optimization*, chapter 8.1: The BFGS Method, pages 194–201. Springer.

Yann Ollivier and Pierre Senellar. 2007. Finding related pages using green measures: An illustration with Wikipedia. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*.

Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 1440–1445.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the 3rd ACL Workshop on Very Large Corpora*, pages 82–94.

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, pages 133–142. Philadelphia, PA.

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.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Nonference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference*, pages 474–479.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus volume 1 — from yesterday's news to tomorrow's language resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 827–832. Las Palmas, Canary Islands.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *Proceedings of Advances in Web Intelligence: Third International Atlantic Web Intelligence Conference*, pages 380–386. Springer, Lodz, Poland.

Fabian Suchanek Ralf Schenkel and Gjergji Kasneci. 2007. YAWN: A semantically annotated Wikipedia XML corpus. In *Proceedings of GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW2007)*, pages 277–291.

Petér Schönhofen. 2006. Identifying document topics using the Wikipedia category network. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 456–462. IEEE Computer Society, Washington, DC, USA.

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.

Philipp Sorg and Philipp Cimiano. 2008. Enriching the crosslingual link structure of Wikipedia – a classification-based approach. In *Proceedings of the AAAI '08 Workshop on Wikipedia and Artifical Intelligence*.

Mark Stevenson and Robert Gaizauskas. 2000. Using corpus-derived name lists for named entity recognition. In *Proceedings of the 6th Applied Natural Language Processing Conference*, pages 290–295.

Michael Strube and Simone Paolo Ponzetto. 2006. Wikirelate! computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1419–1424.

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.

Zareen Syed, Tim Finin, , and Anupam Joshi. 2008. Wikipedia as an ontology for describing documents. In *Proceedings of the 2nd International Conference on Weblogs and Social Media*.

Partha P. Talukdar, Thorsten Brants, Mark Liberman, and Fernando Pereira. 2006. A context pattern induction method for named entity extraction. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 141–148.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pages 214–221.

James A. Thom, Jovan Pehcevski, and Anne-Marie Vercoustre. 2007. Use of Wikipedia categories in entity ranking. In *Proceedings of the 12th Australasian Document Computing Symposium*. Melbourne, Australia.

Christopher Thomas and Amit P. Sheth. 2007. Semantic convergence of Wikipedia articles. In *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 600–606.

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.

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.

Antonio Toral. API. wiki_db_access (C++ Wikipedia API). online resource.

Antonio Toral and Rafael Muñoz. 2006. A proposal to automatically build and maintain gazetteers for named entity recognition using Wikipedia. In *Proceedings of the Workshop on NEW TEXT, 11th Conference of the European Chapter of the Association for Computational Linguistics*.

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.

Anne-Marie Vercoustre, Jovan Pehcevski, and James A. Thom. 2007. Using Wikipedia categories and links in entity ranking. In *Pre-proceedings of the sixth International Workshop of the Initiative for the Evaluation of XML Retrieval*.

Yotaro Watanabe, Masayuki Asahara, and Yuji Matsumoto. 2007. A graph-based approach to named entity categorization in Wikipedia using conditional random fields. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 649–657.

Timothy Weale. 2006. Utilizing Wikipedia categories for document classification. Unpublished.

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

Casey Whitelaw and Jon Patrick. 2003. Named entity recognition using a character-based probabilistic approach. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*, pages 196–199.

Dennis Wilkinson and Bernardo Huberman. 2007. Cooperation and quality in Wikipedia. In *Proceedings of the 2007 International Symposium on Wikis*.

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.

Hugo Zaragoza, Jordi Atserias, Massimiliano Ciaramita, and Giuseppe Attardi. 2007a. Semantically annotated snapshot of the English Wikipedia v.1 (SW1). `http://www.yr-bcn.es/semanticWikipedia`.

Hugo Zaragoza, Henning Rode, Peter Mika, Jordi Atserias, Massimiliano Ciaramita, and Giuseppe Attardi. 2007b. Ranking very many typed entities on Wikipedia. In *Proceedings of the 16th ACM International Conference on Information and Knowledge Management*, pages 1015–1018.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007. Analyzing and accessing Wikipedia as a lexical semantic resource. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer, editors, *Data Structures for Linguistic Resources and Applications*, pages 197–205. Gunter Narr, Tübingen, Germany.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In European Language Resources Association (ELRA), editor, *Proceedings of the 6th International Language Resources and Evaluation*. Marrakech, Morocco.

APPENDIX A

Wikipedia Features

Wikipedia runs the Mediawiki[1] software. In Mediawiki, pages are accessed by uniquely identifying titles, the first character of which is capitalised in English Wikipedia (except when the article contains the Lowercase template). Authors edit pages in a markup format specific to Mediawiki, and their changes are stored in a versioned article history.

A.1 Types of pages

Page namespaces. Each Mediawiki page belongs to a namespace, determined by its prefix. For instance, English Wikipedia uses prefixes: User:, Wikipedia:, Image:, Mediawiki:, Template:, Help:, Category: and Portal:. Each namespace has a corresponding discussion namespace, e.g. User_talk:; Talk: is for main namespace discussion. All page titles with none of these prefixes is part of the *main namespace*, such as all content articles, which we generally consider below.

Redirects. A page's sole purpose may be to redirect the browser to another title, e.g. from Sydney University to University of Sydney as in Figure A.1. Such redirect pages consist of a markup like: #REDIRECT[[University of Sydney]].

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

FIGURE A.1: Example of an article redirect from Sydney University to University of Sydney

(a) Disambiguation page

(b) In-article disambiguation

FIGURE A.2: Disambiguation for titles David Jones (Fig. (a)) and Saturn (Fig. (b))

Disambiguation. Wikipedia designates pages used to list alternative meanings of a title as disambiguation pages, which may be the default page for a title (e.g. David Jones, Figure A.2(a)). Otherwise a title may designate an article about its primary referent, with links provided for alternative meanings or to a disambiguation page (Saturn linking to Saturn (disambiguation), Figure A.2(b)). Page titles disambiguate their referents in parentheses (e.g. Saturn (mythology)), after a comma for locations (Lonon, Ontario), or with initials for personal names (David E. H. Jones).

Lists. Wikipedia has many pages which simply list items with something in common. Their titles often begin List of, e.g. List of places named after Josip Broz Tito.

Stubs. Very short articles are labelled as stubs.

FIGURE A.3: Useful structural features of the Australia Wikipedia article

A.2 Markup features

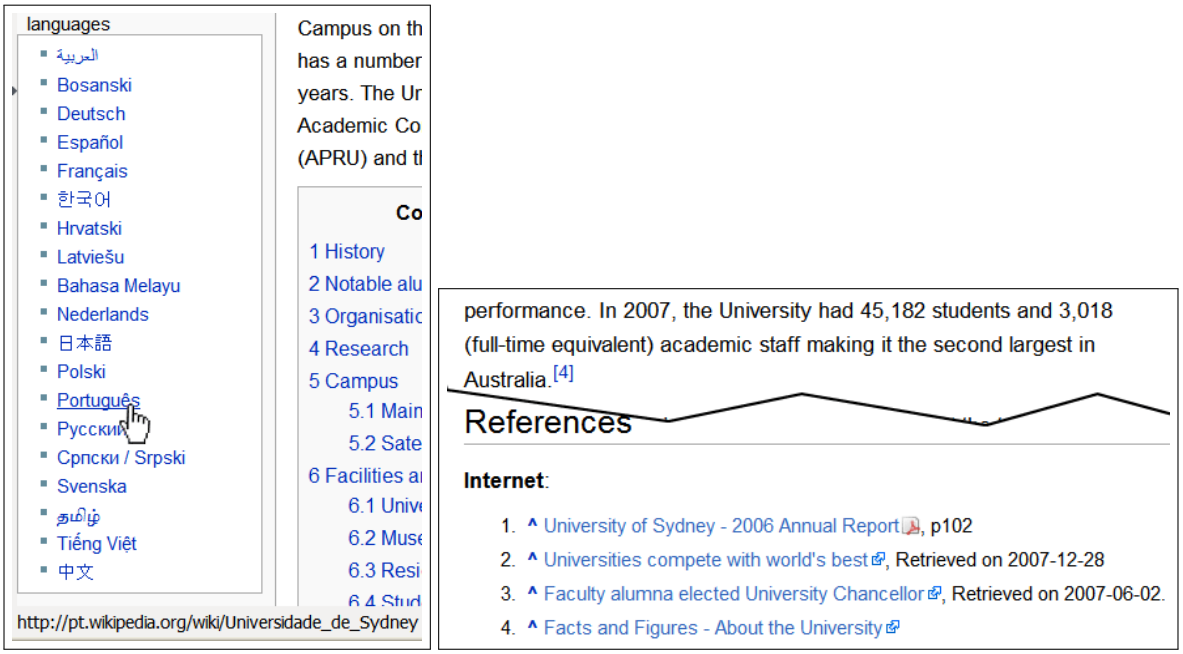
Templates. Mediawiki allows markup reuse through the transclusion of templates, for which the Template: namespace is reserved. Templates can be transcluded recursively, may take arguments, and can perform some basic logic.

Sections. Articles may be split under headings into hierarchical sections. Multiple headings are summarised in a table of contents at the top of an article.

Links. Articles may link to other titles. The markup [[Saturn]] indicates a link to the article on Saturn; [[Saturn|6th planet]] links to Saturn, with link text 6th planet. Links to web pages use a different markup.

Categories. An article's membership in a category is indicated by including a link of the form [[Category:*TITLE*]]. An article may be a member of many categories, which appear at the bottom of an article, as illustrated in Figure A.3.

Language links. An article may indicate an equivalent article in another language by including a link of the form [[*LC*:*TITLE*]] where *LC* is the language code (e.g. de for German). Language links appear in a panel beside the article (see Figure A.4(a)).



(a) Language links (b) References

FIGURE A.4: Examples of language links and references for University of Sydney

Other elements. Mediawiki defines markup for the display of elements such as tables, lists, block-quotes, etc., and for emphasising text and other font formatting. It also admits some HTML.

Extensions. Extensions to the basic Mediawiki markup enable the display of TEX-based mathematical formulae, timelines, image galleries and references to cited works as shown in Figure A.4(b).

A.3 Semantic features

Typical article structure. Wikipedia articles have a structure typical of encyclopedias, beginning with definitions, detail split into sections, and a list of related topics and references at the end. When main articles become too large and detailed, they are split into articles on sub-topics which are summarised briefly in the main article. Thus Australia points users to History of Australia, Government of Australia, Culture of Australia among other sub-articles.

Infoboxes. For many articles of common classes, basic details of the article topic are listed in templated *infoboxes*, shown on the right side of the article, as in Figure A.3.

Article status. Some Wikipedia articles are marked by editors as *featured articles*; some have recordings of the article being spoken; some are protected from being edited by most users. These status facts are indicated with icons in the top-right corner of articles.

APPENDIX B

The C&C NER System

Curran and Clark (2003b) use a maximum entropy (MaxEnt) tagger based on Ratnaparkhi's (1996) solution for part-of-speech tagging, as described in Curran and Clark (2003a). As part of the C&C tools (Curran et al., 2007), its open source is freely available at `http://svn.ask.it.usyd.edu.au/trac/candc`. In all of our experiments, we have used the BFGS method (Nocedal and Wright, 1999) for parameter estimation.

The tagger incorporates a number of features as listed in Table B.1. These include lexemes and POS tags in the neighbourhood of each token; the predicted NE tags for the immediately-preceding unigram and bigram; many orthographic features; some global tag features; and personal name gazetteers. For *word type* features, words are simplified to representations of orthographic patterns (c.f. Collins (2002)) similar to those used in section 3.3.2. The *memory* NE *tag* represents the tag previously assigned to the current token in the document being tagged, while the *unigram tag* is the token's most frequent tag in training. Sizes of the gazetteers used—including data on the capitalisation frequency of words in an external corpus—are given in Table B.2.

Feature type	Condition	Contextual predicates
Lexemes	$\forall w_i$	$w_i = X$ $w_{i-1} = X$, $w_{i-2} = Y$ $w_{i+1} = X$, $w_{i+2} = Y$
Part of speech tags	$\forall w_i$	POS $_i = X$ POS $_{i-1} = X$, POS $_{i-2} = Y$ POS $_{i+1} = X$, POS $_{i+2} = Y$
Predicted NE tags	$\forall w_i$	NE $_{i-1} = X$ NE $_{i-2}$NE $_{i-1} = XY$
Orthography	$freq(w_i) < 5$	X is prefix of w_i, $\|X\| \leq 4$ X is suffix of w_i, $\|X\| \leq 4$ w_i contains a digit w_i is only digits w_i is a number w_i contains a hyphen w_i contains a period w_i contains punctuation w_i contains uppercase character w_i is upper,lower,title,mixed case w_i is alphanumeric w_i has only Roman numerals w_i is an initial (X.) w_i is an acronym (ABC, A.B.C.) length of w_i
Orthographic patterns	$\forall w_i$	unigram of word type bigram of word types trigram of word types
Orthographic knowledge	$\forall w_i$	w_i not lowercase and $f_{LC} > f_{UC}$
Global tag history	$\forall w_i$	memory NE tag for w_i unigram tag of w_{i+1}
Gazetteer lookup	$\forall w_i$	w_i in a gazetteer w_{i-1} in a gazetteer w_{i+1} in a gazetteer

TABLE B.1: Features used by C&C for NER (Curran and Clark, 2003b)

Gazetteer	Entries
First name	6,673
Last name	89,836
$freq_L C > freq_U C$ list	778,791

TABLE B.2: Size of gazetteers used by Curran and Clark (2003b)

APPENDIX C

Detailed Corpus Comparisons

This appendix contains tables of statistics comparing training corpora and their performance:

C.1: gives sizes and distributions of tokens and entities in our primary training corpora;

C.2 and C.3: show performance of each training corpus on DEV data (respectively without and with MISC), including where WP2 is used to augment gold-standard corpora;

C.4: reports DEV performance as the size of the training corpus is changed;

C.5: shows TEST performance of gold-standard and WP2 corpora, including class break-down;

C.6 and C.7: exemplify results of entity type frequency analysis;

C.8 and C.9: compare tag sequence confusion (CORR-PRED and PRED-CORR respectively) for four training corpora evaluated on BBN DEV;

C.10 and C.11: compare tag sequence confusion similarly on CONLL DEV.

TRAIN	Number of tokens				% distribution of entities				Average entity length (toks)			
corpus	Total	/entity	/sentence	/article	LOC	MISC	ORG	PER	LOC	MISC	ORG	PER
BBN	901894	18.0	23.8	508.1	22	9	46	21	1.3	1.5	2.0	1.6
CONLL	203621	8.7	14.5	215.0	30	14	26	28	1.2	1.3	1.6	1.7
MUC*	84051	19.5	24.1	824.0	33	-	40	26	1.5	-	1.7	1.6
MUC	83601	19.4	24.0	819.6	33	-	40	26	1.4	-	1.7	1.5
WPB	3500008	12.0	21.4	70.6	37	20	14	27	1.8	2.4	2.4	2.1
WP0	3500011	12.5	21.9	81.6	36	23	14	25	1.7	2.2	2.4	2.0
WP1	3500022	12.3	23.7	360.1	34	22	13	28	1.4	1.8	2.0	1.6
WP2	3500032	12.1	23.9	417.2	31	22	14	31	1.4	1.8	1.9	1.5
WP3	3500039	12.0	24.0	452.3	29	21	13	35	1.4	1.8	1.9	1.4
WP4	3500000	11.8	24.2	506.4	29	21	14	34	1.4	1.8	1.8	1.4

TABLE C.1: Basic training corpus statistics: total number of tokens, and number of tokens per number of entities, sentences and articles; distribution of entity annotations; average length of entity annotations of each class

TRAIN	DEV	With MISC	TEXT F	TYPE F	MUCEVAL P	MUCEVAL R	MUCEVAL F	EXACT P	EXACT R	EXACT F
MUC	MUC		91.7	86.3	87.7	90.4	89.0	81.0	83.6	82.3
CONLL	MUC		86.2	76.6	81.2	81.6	81.4	69.7	70.1	69.9
BBN	MUC		89.9	85.3	87.5	87.7	87.6	80.1	80.2	80.2
WPB	MUC		66.0	63.9	66.1	63.9	65.0	54.4	52.7	53.5
WP0	MUC		77.8	74.8	85.9	68.6	76.3	78.5	62.7	69.7
WP1	MUC		84.2	78.5	90.6	73.8	81.3	83.8	68.3	75.3
WP2	MUC		83.5	79.7	91.9	73.4	81.6	86.2	68.9	76.6
WP3	MUC		83.3	80.5	91.7	74.0	81.9	86.4	69.8	77.2
WP4	MUC		82.8	79.0	89.7	73.7	80.9	83.8	68.9	75.6
WP2+MUC	MUC		91.4	89.8	95.1	86.5	90.6	91.8	83.5	87.4
WP2+CONLL	MUC		85.2	83.3	91.5	78.1	84.3	85.7	73.2	79.0
WP2+BBN	MUC		86.2	83.1	91.7	78.6	84.6	86.0	73.7	79.4
MUC	CONLL		76.8	59.6	71.0	65.5	68.2	57.2	52.8	54.9
CONLL	CONLL		92.0	89.7	89.0	92.8	90.8	85.2	88.8	86.9
BBN	CONLL		74.9	68.6	71.7	71.8	71.8	59.0	59.0	59.0
WPB	CONLL		52.7	52.9	57.3	49.0	52.8	46.4	39.6	42.8
WP0	CONLL		72.2	70.0	80.4	63.7	71.1	73.1	58.0	64.7
WP1	CONLL		75.7	70.5	82.8	65.4	73.1	76.7	60.6	67.7
WP2	CONLL		76.2	72.7	84.5	66.5	74.5	78.8	62.0	69.4
WP3	CONLL		77.0	72.3	84.7	66.8	74.7	78.8	62.2	69.5
WP4	CONLL		75.9	70.3	84.0	64.7	73.1	77.3	59.5	67.3
WP2+MUC	CONLL		76.3	74.5	84.1	68.4	75.4	77.5	63.0	69.5
WP2+CONLL	CONLL		92.5	90.4	93.1	89.9	91.5	90.0	86.9	88.4
WP2+BBN	CONLL		77.9	75.1	85.6	69.1	76.5	80.0	64.6	71.5
MUC	BBN		82.8	75.6	77.5	81.0	79.2	67.8	70.9	69.3
CONLL	BBN		75.2	70.0	69.4	76.1	72.6	57.6	63.1	60.2
BBN	BBN		92.1	90.9	89.4	93.7	91.5	86.0	90.1	88.0
WPB	BBN		62.6	67.8	64.8	65.7	65.2	50.7	51.4	51.0
WP0	BBN		81.1	76.2	83.0	74.8	78.7	73.9	66.6	70.0
WP1	BBN		83.7	78.3	84.5	77.8	81.0	76.8	70.7	73.6
WP2	BBN		84.7	79.1	85.8	78.4	81.9	78.7	71.9	75.1
WP3	BBN		83.8	77.7	85.0	76.9	80.7	77.5	70.2	73.7
WP4	BBN		83.3	78.1	83.9	77.8	80.7	76.2	70.6	73.3
WP2+MUC	BBN		85.5	79.9	86.2	79.5	82.7	79.4	73.2	76.2
WP2+CONLL	BBN		86.6	81.4	85.8	82.3	84.0	78.7	75.5	77.1
WP2+BBN	BBN		92.6	91.3	92.5	91.4	91.9	89.3	88.2	88.7

TABLE C.2: DEV performance without MISC

TRAIN	DEV	With MISC	TEXT F	TYPE F	MUCEVAL P	MUCEVAL R	MUCEVAL F	EXACT P	EXACT R	EXACT F
CONLL	CONLL	✓	92.5	89.2	90.3	91.5	90.9	85.3	86.5	85.9
BBN	CONLL	✓	77.0	68.4	75.8	69.8	72.7	62.0	57.1	59.4
WPB	CONLL	✓	54.4	47.6	51.5	50.6	51.0	37.4	36.8	37.1
WP0	CONLL	✓	74.4	67.6	78.4	64.8	71.0	69.4	57.4	62.8
WP1	CONLL	✓	79.6	70.2	80.3	70.2	74.9	72.0	62.9	67.2
WP2	CONLL	✓	80.5	71.8	81.2	71.6	76.1	73.6	64.9	69.0
WP3	CONLL	✓	80.5	72.2	81.8	71.5	76.3	73.9	64.6	68.9
WP4	CONLL	✓	78.9	69.6	80.1	69.2	74.3	71.4	61.7	66.2
WP2+CONLL	CONLL	✓	93.3	90.0	92.9	90.5	91.7	88.7	86.5	87.6
WP2+BBN	CONLL	✓	82.5	75.4	84.4	74.1	78.9	77.0	67.6	72.0
CONLL	BBN	✓	80.1	70.0	73.1	77.2	75.1	60.2	63.6	61.9
BBN	BBN	✓	92.6	89.6	90.3	92.0	91.1	85.7	87.3	86.5
WPB	BBN	✓	63.3	62.7	59.5	67.0	63.0	43.1	48.6	45.7
WP0	BBN	✓	84.3	74.2	80.3	78.2	79.2	70.7	68.8	69.7
WP1	BBN	✓	86.9	77.7	82.1	82.5	82.3	73.2	73.5	73.4
WP2	BBN	✓	87.4	78.1	82.1	83.4	82.7	73.4	74.6	74.0
WP3	BBN	✓	87.2	77.3	81.8	82.7	82.2	73.1	73.9	73.5
WP4	BBN	✓	85.8	77.0	80.8	82.0	81.4	71.8	72.9	72.3
WP2+CONLL	BBN	✓	89.1	79.5	83.4	85.2	84.3	74.8	76.4	75.6
WP2+BBN	BBN	✓	92.8	89.5	91.1	91.3	91.2	86.8	87.0	86.9

TABLE C.3: DEV performance with MISC

(a) TYPE scoring

TRAIN size (M tokens)	With MISC CONLL	With MISC BBN	No MISC MUC	No MISC CONLL	No MISC BBN
0.25	70.7	69.4	73.4	66.7	62.6
0.5	71.6	72.8	76.4	69.6	68.5
1.0	74.3	77.3	79.8	71.2	73.4
1.5	75.9	79.0	80.5	73.5	76.3
2.0	76.7	80.3	81.2	74.6	78.8
2.5	76.8	81.3	81.3	74.2	79.1
3.0	76.9	82.3	81.0	74.3	80.0
3.5	76.1	82.7	81.6	74.5	81.9
4.0	76.1	83.0	81.8	74.6	82.1
4.5	76.6	83.0	82.2	74.4	81.3
5.0	76.3	82.7	81.8	74.3	81.3
5.5	75.8	82.7	81.1	75.4	81.4
6.0	75.7	83.2	81.6	74.1	81.0
6.5	73.4	83.1	81.3	72.7	81.1

(b) EXACT scoring

With MISC CONLL	With MISC BBN	No MISC MUC	No MISC CONLL	No MISC BBN
59.8	53.1	66.2	59.2	46.8
62.5	58.8	69.8	64.2	58.6
66.6	65.6	73.3	66.0	64.5
68.1	68.1	75.5	68.4	68.3
69.4	70.5	75.8	69.4	71.2
69.5	72.0	76.6	69.5	72.4
69.9	73.5	77.0	69.5	73.2
69.0	74.0	76.6	69.4	75.1
69.1	74.3	76.7	69.5	75.2
69.5	74.3	77.2	69.2	73.8
68.9	73.6	76.3	68.6	73.7
68.2	74.2	75.4	70.0	74.0
68.2	75.0	76.7	68.6	73.9
65.4	74.9	76.0	66.2	73.8

TABLE C.4: Effect of changing Wikipedia corpus size on DEV performance. TRAIN corpora are from WP2 data.

TRAIN	TEST	With MISC	TEXT F	TYPE F	MUCEVAL P	MUCEVAL R	MUCEVAL F	EXACT P	EXACT R	EXACT F	TYPE LOC	TYPE MISC	TYPE ORG	TYPE PER	EXACT LOC	EXACT MISC	EXACT ORG	EXACT PER
MUC	MUC		83.4	78.7	78.9	83.2	81.0	71.7	75.5	73.5	77.3	—	76.0	85.8	75.1	—	67.4	82.8
CoNLL	MUC		80.0	72.4	71.6	81.5	76.2	61.9	70.4	65.9	69.8	—	67.2	86.0	66.7	—	56.9	79.9
BBN	MUC		84.6	82.7	79.7	88.0	83.6	74.2	82.0	77.9	83.8	—	80.2	86.1	81.8	—	72.5	83.1
WP2	MUC		82.0	80.5	85.3	77.6	81.3	80.7	73.4	76.8	80.5	—	74.8	90.8	77.9	—	70.3	87.0
MUC	CoNLL		76.2	60.8	72.4	64.9	68.5	58.7	52.6	55.5	64.6	—	41.0	73.9	62.3	—	34.0	66.9
CoNLL	CoNLL		91.0	84.8	86.3	89.6	87.9	80.6	83.6	82.1	83.5	—	80.4	90.8	82.2	—	75.0	89.3
BBN	CoNLL		72.3	64.7	69.6	67.5	68.5	54.7	53.1	53.9	67.3	—	47.4	78.5	65.4	—	38.0	56.6
WP2	CoNLL		71.8	65.4	82.2	58.9	68.6	73.7	52.8	61.5	69.0	—	40.0	80.6	66.4	—	33.9	77.0
MUC	BBN		79.3	75.9	75.7	79.6	77.6	65.8	69.2	67.5	73.4	—	77.8	74.5	70.1	—	64.4	71.2
CoNLL	BBN		76.7	71.5	71.7	76.7	74.1	60.3	64.6	62.4	65.1	—	75.5	71.5	62.5	—	62.3	62.3
BBN	BBN		92.0	91.8	89.9	93.9	91.9	86.5	90.4	88.4	89.9	—	92.4	92.4	88.5	—	87.5	90.5
WP2	BBN		79.2	75.4	82.2	72.8	77.2	74.4	65.9	69.9	80.8	—	68.7	82.0	77.1	—	60.2	80.9
CoNLL	CoNLL	✓	91.3	84.2	87.0	88.5	87.8	80.5	81.9	81.2	83.2	79.4	80.2	91.4	81.8	73.4	75.3	89.8
BBN	CoNLL	✓	73.5	65.1	73.0	66.0	69.3	57.6	52.0	54.7	67.2	63.0	49.5	78.2	65.4	58.6	39.3	56.5
WP2	CoNLL	✓	75.1	65.3	76.0	65.3	70.2	66.0	56.6	60.9	69.2	58.0	43.0	82.0	65.8	52.7	36.8	78.2
CoNLL	BBN	✓	79.8	70.1	73.2	76.9	75.0	60.8	63.8	62.3	66.3	50.3	75.2	73.6	63.6	47.3	62.7	66.8
BBN	BBN	✓	91.9	90.3	90.2	92.0	91.1	85.8	87.6	86.7	90.4	78.4	91.8	91.8	88.7	74.2	86.8	89.8
WP2	BBN	✓	83.3	75.0	78.4	79.9	79.1	68.6	69.9	69.3	79.4	56.9	74.4	79.3	75.9	53.3	65.1	77.5

TABLE C.5: TEST performance with break-down by entity class

Wordtype	LOC				MISC				ORG				PER			
	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN
Aaa	85.5	87.4	76.6	74.4	47.4	64.2	74.3	78.5	62.0	61.5	65.5	55.6	80.1	90.6	89.6	88.0
AA	0.7	2.3	16.9	0.4	3.3	5.1	11.6	1.2	7.1	18.1	20.0	10.9	0.2	0.3	1.6	0.4
A.A.	0.2	1.2	4.4	14.0	0.0	0.0	0.0	0.3	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0
Aaa Aaa.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.2	7.4	0.0	0.0	0.0	0.0
Aaa of Aaa	2.3	2.1	0.1	0.3	2.9	2.3	0.3	0.5	6.3	4.2	1.2	1.7	1.5	0.0	0.0	0.0
Aaa A. Aaa	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	2.0	0.9	0.1	5.9
Aaa aa	1.5	1.1	0.1	0.5	4.1	2.9	0.1	0.3	2.2	1.9	0.1	0.1	0.5	0.3	0.0	0.0
A. Aaa	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.1	1.2	0.6	3.9	0.5
The Aaa	0.2	0.2	0.3	0.0	3.7	2.2	0.0	0.8	2.6	1.3	0.2	0.3	0.1	0.1	0.0	0.0
Aaa.	0.0	0.6	0.1	3.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AA Aaa	0.1	0.1	0.0	0.0	1.4	1.0	0.4	0.1	1.8	1.4	2.8	1.2	0.1	0.1	0.0	0.0
A.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aaa AA	0.1	0.1	0.0	0.0	2.2	1.4	0.1	0.7	0.5	0.4	0.7	0.6	1.4	0.9	0.1	0.1
Aaa , Aaa	2.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.0	0.0	0.0
Aaa Aa.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8	0.0	0.0	0.0	0.4
Aaa-aa	0.1	0.1	0.0	1.4	0.1	0.3	1.8	1.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1
Aaa aa Aaa	0.4	0.3	0.1	0.0	0.2	0.2	0.1	0.1	0.3	0.3	0.2	0.2	1.7	0.8	0.6	0.2
Aaa 00	0.1	0.1	0.0	0.1	1.0	0.6	0.7	1.6	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Aaa & Aaa	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.1	0.3	0.2	0.2	1.4	0.0	0.0	0.0	0.0
A,	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aaa 's	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.3	0.1	0.3	1.3	0.6	0.0	0.0	0.0
'Aaa Aaa	0.1	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0
Aaa-Aaa	0.5	0.4	0.1	0.2	0.6	0.7	0.5	1.2	0.5	0.4	0.2	1.2	0.1	0.2	0.0	0.3
Aa Aaa	0.3	0.4	0.1	0.2	0.2	0.2	0.0	0.1	0.2	0.1	1.1	0.3	0.4	0.4	0.3	0.2
AaaAaa	0.0	0.0	0.0	0.0	0.6	0.7	0.1	0.2	0.4	0.5	0.3	1.0	0.0	0.0	0.0	0.0
The Aaa of Aaa	0.0	0.0	0.0	0.0	1.0	0.6	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TABLE C.6: Entity wordtype frequency comparison (see section 3.3.2) for four corpora. Each cell shows the percentage of a corpus's entities of a particular class having the row's pattern. Orthographic pattern types replace content words. Sequences of duplicate patterns are removed (Aaa Aaa → Aaa). Values dissimilar to CoNLL are blue, to BBN are red, and to both are purple. All rows where at least one cell contains over 1% are shown.

POS type	LOC				MISC				ORG				PER			
	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN
NNP	79.2	87.9	96.2	93.5	49.2	48.1	44.2	28.2	65.9	73.2	87.5	82.0	89.0	93.0	95.5	98.6
JJ	3.7	0.5	0.2	2.1	5.4	17.8	39.1	44.8	1.0	0.6	0.4	0.3	0.5	0.4	0.2	0.1
NNPS	0.7	0.4	0.1	0.2	1.4	3.7	3.2	9.6	0.5	1.5	0.6	0.3	0.2	0.2	0.2	0.5
NNP IN NNP	1.9	1.6	0.1	0.2	2.7	1.6	0.5	0.2	6.2	4.0	1.0	1.9	1.9	0.1	0.0	0.1
NNP NNPS	3.0	2.0	0.9	0.4	1.0	0.7	0.7	1.1	4.4	3.5	1.6	2.5	0.4	0.2	0.3	0.0
NNP CC NNP	0.3	0.3	0.0	0.0	0.8	0.6	0.0	0.2	0.9	0.7	0.6	3.4	0.2	0.1	0.0	0.0
NNS	0.6	0.8	0.6	0.2	1.7	3.1	2.0	2.9	0.9	2.5	1.5	0.3	0.6	1.3	0.6	0.1
DT NNP	0.3	0.2	0.2	0.0	2.8	1.6	0.0	0.7	1.7	0.9	0.2	0.3	0.1	0.1	0.1	0.0
NNP NN	1.1	0.8	0.1	0.5	2.8	1.7	0.2	0.3	1.6	1.3	0.2	0.1	0.4	0.2	0.0	0.0
NN	0.8	1.1	0.6	0.5	2.0	2.2	1.6	1.7	1.6	2.1	0.9	0.3	0.7	1.4	0.7	0.1
NNP , NNP	2.2	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.4	0.0	0.0	0.1	0.2	0.1	0.0	0.0
NNP CD	0.3	0.2	0.0	0.1	2.1	1.4	1.1	0.9	0.3	0.2	0.1	0.1	0.1	0.1	0.0	0.0
$	0.0	0.0	0.0	0.0	0.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JJ NNP	0.6	0.5	0.1	0.1	1.2	1.1	1.3	0.0	1.6	1.3	0.6	0.1	0.3	0.2	0.2	0.0
JJ NN	0.8	0.6	0.0	0.1	1.6	1.2	0.8	0.2	1.3	1.0	0.2	0.0	0.2	0.1	0.0	0.0
NNP NNPS NNP	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.4	0.3	0.4	1.5	0.0	0.0	0.0	0.0
NNP POS	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.4	0.2	0.3	1.4	0.7	0.0	0.0	0.0
NNP .	0.1	0.0	0.1	1.4	0.2	0.1	0.1	0.0	0.2	0.1	0.1	1.2	0.0	0.0	0.0	0.1
NN IN NNP	0.3	0.3	0.0	0.0	1.3	1.2	0.1	0.1	0.4	0.3	0.1	0.0	0.3	0.0	0.0	0.0
NN CD	0.0	0.0	0.0	0.0	0.1	0.0	0.1	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CD	0.0	0.0	0.0	0.0	0.2	0.2	0.1	1.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0

TABLE C.7: Entity POS type frequency comparison (see section 3.3.2) for four corpora. Each cell shows the percentage of a corpus's entities of a particular class having the row's POS tags. Sequences of duplicate POS tags are removed (NNP NNP → NNP). Values dissimilar to CoNLL are blue, to BBN are red, and to both are purple. All rows where at least one cell contains over 1% are shown.

Predicted	Correct: LOC				Correct: MISC				Correct: ORG				Correct: PER			
	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN
LOC	**1802**	**1903**	**1775**	**1967**	534	83	68	67	717	427	638	133	140	27	280	28
MISC	31	36	45	10	**137**	**617**	**473**	**615**	120	176	80	13	23	15	8	7
ORG	241	98	168	88	157	68	190	100	**2105**	**2043**	**2050**	**3010**	96	34	62	37
PER	15	21	58	16	34	49	62	32	105	214	236	64	**1494**	**1667**	**1381**	**1693**
O	7	24	16	14	21	54	58	48	127	237	65	35	4	22	23	6
ORG O	2	1	5	0	1	1	5	2	64	67	70	12	1	1	0	1
O ORG	1	0	0	0	5	2	9	5	10	25	20	11	0	0	0	0
PER ORG	0	0	1	0	0	0	0	0	7	9	43	9	1	0	3	5
LOC O	5	12	20	4	0	0	3	2	9	8	10	4	1	0	0	0
LOC ORG	0	0	0	0	0	0	0	0	12	8	25	11	0	0	0	0
PER O	1	3	0	0	1	1	1	0	4	18	4	3	2	4	8	1
O LOC	1	2	2	2	3	4	9	1	6	14	4	0	0	0	0	0
ORG PER	0	1	2	0	0	0	0	0	6	6	6	5	0	4	7	5
PER PER	0	1	0	0	2	0	0	0	2	2	0	0	22	4	0	0
O MISC	0	1	2	0	2	9	1	7	1	2	4	0	0	0	0	0
O PER	0	0	0	0	0	1	2	4	0	6	2	2	0	4	3	1
LOC O ORG	0	0	0	0	0	1	1	2	5	4	7	3	0	0	0	0
MISC ORG	0	0	0	0	0	0	1	0	0	4	9	8	0	0	0	0
LOC O LOC	0	1	2	0	0	0	2	1	4	6	5	1	0	0	0	0
ORG O ORG	0	0	0	0	0	0	3	0	4	10	2	2	0	0	1	0
MISC O	1	2	0	0	3	5	2	4	0	2	0	0	0	0	0	0
PER O PER	0	0	0	0	0	0	0	0	4	3	3	1	1	0	2	0
LOC PER	0	0	1	0	0	0	0	0	0	0	5	1	0	0	5	1
PER LOC	0	1	4	1	0	0	0	0	0	1	1	0	0	2	1	0
PER O ORG	0	0	0	0	0	0	1	0	4	5	0	0	0	0	0	0

TABLE C.8: CORR-PRED sequence confusion matrix (see section 3.3.3) comparing four training corpora on BBN DEV. Each cell shows the frequency of its row's tag sequence being predicted for a gold-standard entity of the column's class.
O represents one or more non-entity tokens. For example, correct: "[ORG InfoCorp .]", predicted "[ORG InfoCorp] ." is one of 64 entries in WPB's cell (col: ORG; row: ORG O).
Rows whose sum is ≥ 10 are shown. Cells are highlighted where $((f_{cell}/f_{BBN} \geq 2) \vee (f_{cell}/f_{BBN} < \frac{1}{2})) \wedge (|f_{cell} - f_{BBN}| > 4)$.

Correct	Predicted: LOC				Predicted: MISC				Predicted: ORG				Predicted: PER			
	WPB	WP2 CoNLL		BBN	WPB	WP2 CoNLL		BBN	WPB	WP2 CoNLL		BBN	WPB	WP2 CoNLL		BBN
LOC	**1378**	**1875**	**1799**	**1961**	12	26	52	7	55	70	98	58	4	23	62	17
MISC	476	85	92	77	**81**	**619**	**475**	**623**	101	54	170	87	21	50	66	34
ORG	669	466	722	164	81	188	102	22	**1919**	**2165**	**1980**	**2957**	128	284	331	90
PER	71	27	112	29	2	7	8	7	27	27	37	33	**758**	**1669**	**1219**	**1688**
O	621	109	189	45	492	159	144	47	396	112	213	139	84	47	27	10
O PER	27	1	170	0	9	4	0	0	19	6	16	7	734	12	188	9
ORG O	42	5	5	1	11	10	1	1	123	19	68	41	3	1	1	1
LOC O	97	23	9	14	2	3	0	1	26	8	6	13	0	1	1	0
LOC O LOC	120	3	0	0	2	0	0	0	29	0	3	2	1	1	0	0
ORG O ORG	8	2	0	0	3	1	0	0	28	12	61	20	0	0	0	0
O ORG	7	2	3	0	10	1	3	0	24	4	30	11	6	0	0	0
O MISC	9	0	0	0	35	2	3	2	11	2	4	3	7	0	0	0
ORG O LOC	8	0	0	0	1	1	0	0	26	6	22	5	0	0	0	0
ORG MISC	3	1	5	2	7	5	0	0	6	7	7	15	1	2	1	2
MISC O	21	0	0	0	5	1	3	4	6	1	6	2	1	0	0	0
O LOC	21	0	0	1	8	2	0	2	4	2	6	1	0	0	0	0
ORG O PER	8	0	0	0	4	1	0	0	21	1	7	4	0	0	0	0
PER O	11	1	0	0	0	0	1	1	1	1	1	0	19	3	3	3
LOC ORG	6	8	0	0	1	2	0	0	10	2	2	2	0	0	0	0
MISC O MISC	0	0	0	0	1	3	0	3	4	2	7	2	0	0	0	0
ORG O MISC	6	2	0	0	2	0	0	1	3	0	4	3	0	0	0	0
O PER O	0	0	1	0	3	0	0	0	4	0	1	0	12	0	0	0
LOC O ORG	6	3	0	0	0	0	0	0	10	0	0	0	0	0	0	0
LOC MISC	0	1	0	1	0	1	0	0	3	4	4	2	0	0	1	0
O PER O MISC O LOC	6	0	0	0	0	0	0	0	8	0	0	0	1	0	0	0
LOC O PER	6	0	0	0	0	1	0	0	4	0	3	0	0	0	0	0
ORG O LOC O LOC	1	0	0	0	0	0	0	0	6	0	6	1	0	0	0	0
O MISC O	0	1	0	0	6	1	0	0	1	0	0	0	3	0	0	0
PER LOC	0	1	0	0	0	0	0	0	0	2	1	0	1	2	2	2
ORG O LOC O	1	0	0	0	0	0	0	0	7	0	2	0	0	0	0	0

TABLE C.9: PRED-CORR sequence confusion matrix comparing four training corpora on BBN DEV. Each cell shows the frequency of its row's gold-standard tag sequence being tagged as an entity of the column's class.

Predicted	Correct: LOC				Correct: MISC				Correct: ORG				Correct: PER			
	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN	WPB	WP2	CoNLL	BBN
LOC	**1350**	**1422**	**1709**	**1324**	447	99	39	130	480	324	84	316	435	81	62	80
MISC	60	47	13	10	**184**	**554**	**731**	**470**	70	59	16	27	56	22	6	5
ORG	248	71	70	241	208	77	60	156	**591**	**487**	**1124**	**587**	173	43	36	181
PER	38	19	20	83	42	31	16	27	30	57	45	139	**975**	**1525**	**1715**	**1508**
O	121	266	9	160	29	128	25	84	135	372	18	194	114	138	6	48
PER PER	0	0	0	0	0	0	0	0	1	1	0	0	66	11	0	0
LOC O	0	1	3	2	3	6	2	9	1	4	9	17	0	0	0	1
O LOC	15	3	3	3	0	0	2	0	11	10	1	4	4	0	0	0
ORG O	0	0	2	1	1	8	6	16	1	0	3	7	0	0	1	0
O ORG	4	0	2	0	3	1	1	5	10	5	2	3	2	0	0	0
O PER	0	0	0	2	0	1	2	1	0	1	2	2	5	9	0	3
LOC MISC	0	1	0	0	0	3	12	4	0	0	0	4	0	0	0	0
O MISC	0	0	0	0	1	3	9	2	5	0	0	0	0	0	0	0
MISC MISC	0	3	0	0	0	5	4	0	0	4	0	0	2	0	0	1
PER O	0	0	0	0	0	0	2	0	0	1	2	3	0	0	2	6
PER ORG	0	0	0	1	0	0	0	0	1	0	5	2	1	2	2	2
PER LOC	0	0	2	8	0	0	0	0	0	0	0	2	3	0	0	1
ORG O LOC	0	0	0	0	0	1	0	1	0	4	1	9	0	0	0	0
MISC ORG	0	0	1	0	0	0	0	2	0	0	6	4	0	0	0	0
MISC O	0	2	0	0	0	0	2	2	0	0	4	2	0	0	0	0
LOC ORG	0	0	0	0	0	0	0	0	1	3	3	2	0	0	0	1

TABLE C.10: CORR-PRED sequence confusion matrix comparing four training corpora on CoNLL DEV.

Correct	Predicted: LOC				Predicted: MISC				Predicted: ORG				Predicted: PER			
	WPB	WP2	CONLL	BBN	WPB	WP2	CONLL	BBN	WPB	WP2	CONLL	BBN	WPB	WP2	CONLL	BBN
LOC	**880**	**1402**	**1701**	**1316**	1	27	15	10	48	58	48	136	3	17	21	51
MISC	373	92	58	149	**120**	**551**	**737**	**464**	97	65	61	129	28	27	22	28
ORG	194	211	101	293	16	44	28	37	**358**	**451**	**1088**	**493**	23	59	54	127
PER	218	75	66	79	19	11	8	4	35	33	25	106	**966**	**1536**	**1713**	**1204**
O	322	123	31	32	369	73	45	39	183	43	32	88	47	9	11	17
PER O	23	1	1	3	17	5	0	0	16	4	3	8	37	25	11	256
ORG O	103	71	0	19	31	6	1	1	39	19	15	32	7	2	0	13
LOC O	153	11	8	12	26	2	0	0	56	3	2	2	1	0	0	2
PER O LOC O	131	0	0	0	12	0	0	0	55	0	0	6	26	0	0	0
MISC O	34	6	0	0	22	8	14	10	36	8	0	27	6	1	1	2
O PER	13	3	3	0	4	6	1	2	12	4	3	4	63	3	9	12
ORG O ORG O	41	19	0	7	6	4	0	0	37	1	0	2	0	0	0	0
ORG O ORG	20	9	0	16	0	0	0	0	17	10	13	21	0	0	0	4
PER O LOC	6	0	0	0	0	0	0	0	3	1	0	30	1	0	0	40
ORG O LOC	10	2	0	4	2	12	0	0	25	4	8	13	0	0	0	0
O ORG	17	3	0	2	8	1	0	0	17	1	9	11	3	0	1	2
O MISC	9	2	0	0	24	7	3	4	14	2	2	4	1	0	0	0
LOC O LOC	31	2	2	1	0	1	0	0	3	1	1	15	0	0	0	0
O LOC	11	1	3	3	9	10	0	1	3	1	8	2	2	0	0	0
O LOC O	28	0	0	0	0	1	0	0	6	0	0	0	1	0	0	0
LOC O PER	12	3	1	0	1	1	0	0	4	0	0	1	3	1	0	0
ORG O PER	4	0	0	0	0	0	0	1	8	2	5	1	0	0	0	1
O ORG O	2	2	0	3	3	0	0	2	6	0	1	2	0	0	0	0
PER O PER	0	0	0	0	1	0	0	0	3	0	0	1	6	0	0	7
MISC ORG	1	4	0	1	0	0	0	0	6	1	2	1	0	0	1	0
MISC PER	1	0	0	0	0	0	0	0	5	2	1	1	2	5	0	0
LOC O LOC O	10	0	0	0	0	0	0	0	4	0	0	3	0	0	0	0
O MISC O	0	0	0	0	7	0	0	1	9	0	0	0	0	0	0	0
MISC O PER	1	0	0	0	1	0	0	0	3	4	0	3	3	0	0	0
PER O PER O	3	0	0	0	0	0	0	0	3	0	0	2	4	0	0	0
MISC MISC	1	0	0	1	0	1	4	0	2	1	0	1	1	0	0	0
LOC MISC	0	0	0	0	0	1	1	0	3	0	0	4	0	1	2	0
MISC O LOC	6	1	0	0	1	0	0	0	2	1	0	1	0	0	0	0
MISC O MISC	4	1	0	0	1	0	1	3	1	0	0	1	0	0	0	0
LOC O LOC O MISC O	3	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
MISC LOC	2	2	0	1	0	0	1	0	1	1	1	1	0	0	0	0
PER O ORG	1	0	0	1	0	0	0	0	2	0	1	4	1	0	0	0

TABLE C.11: PRED-CORR sequence confusion matrix comparing four training corpora on CONLL DEV.