Learning multilingual named entity recognition from Wikipedia

Joel Nothman[a,b] Nicky Ringland[a] Will Radford[a,b] Tara Murphy[a]
James R. Curran[a,b]

[a] *School of Information Technologies*
University of Sydney
NSW 2006, Australia

[b] *Capital Markets CRC*
55 Harrington Street
NSW 2000, Australia

Abstract

We automatically create enormous, free and multilingual *silver-standard* training annotations for named entity recognition (NER) by exploiting the text and structure of Wikipedia. Most NER systems rely on statistical models of annotated data to identify and classify names of people, locations and organisations in text. This dependence on expensive annotation is the knowledge bottleneck our work overcomes.

We first classify each Wikipedia article into named entity (NE) types, training and evaluating on 7,200 manually-labelled Wikipedia articles across nine languages. Our cross-lingual approach achieves up to 95% accuracy.

We transform the links between articles into NE annotations by projecting the target article's classifications onto the anchor text. This approach yields reasonable annotations, but does not immediately compete with existing gold-standard data. By inferring additional links and heuristically tweaking the Wikipedia corpora, we better align our automatic annotations to gold standards.

We annotate millions of words in nine languages, evaluating English, German, Spanish, Dutch and Russian Wikipedia-trained models against CoNLL Shared Task data and other gold-standard corpora. Our approach outperforms other approaches to automatic NE annotation (Richman and Schone, 2008; Mika et al., 2008); competes with gold-standard training when tested on an evaluation corpus from a different source; and performs 10% better than newswire-trained models on manually-annotated Wikipedia text.

Keywords: Named Entity Recognition, Information Extraction, Wikipedia, Semi-structured resources, Annotated corpora, Semi-supervised learning

1. Introduction

Named Entity Recognition (NER) is the information extraction task of identifying and classifying mentions of people, organisations, locations and other named entities (NEs) within text. It is a core component in many natural language processing (NLP) applications, including question answering, summarisation, and machine translation.

Manually annotated newswire has played a defining role in NER, starting with the Message Understanding Conference (MUC) 6 and 7 evaluations (Chinchor, 1998b) and continuing with the Conference on Natural Language Learning (CoNLL) shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) held in Spanish, Dutch, German and English. More recently, the BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005), added detailed NE annotations to the Penn Treebank (Marcus et al., 1993).

With a substantial amount of annotated data and a strong evaluation methodology in place, the focus of research in this area has almost entirely been on developing language-independent systems that learn statistical models for NER. The competing systems extract terms and patterns indicative of particular NE types, making use of many types of contextual, orthographic, linguistic and external evidence.

Unfortunately, the need for time-consuming and expensive expert annotation hinders the creation of high-performance NE recognisers for most languages and domains. This data dependence has impeded the adaptation or *porting* of existing NER systems to new domains such as scientific or biomedical text, e.g. Nobata et al. (2000). The adaptation penalty is still apparent even when the same NE types are used in text from similar domains (Ciaramita and Altun, 2005).

Differing conventions on entity types and boundaries complicate evaluation, as one model may give reasonable results that do not exactly match the test corpus. Even within CoNLL there is substantial variability: nationalities are tagged as MISC in Dutch, German and English, but not in Spanish. Without fine-tuning types and boundaries for each corpus individually, which requires language-specific knowledge, systems that produce different but equally valid results will be penalised.

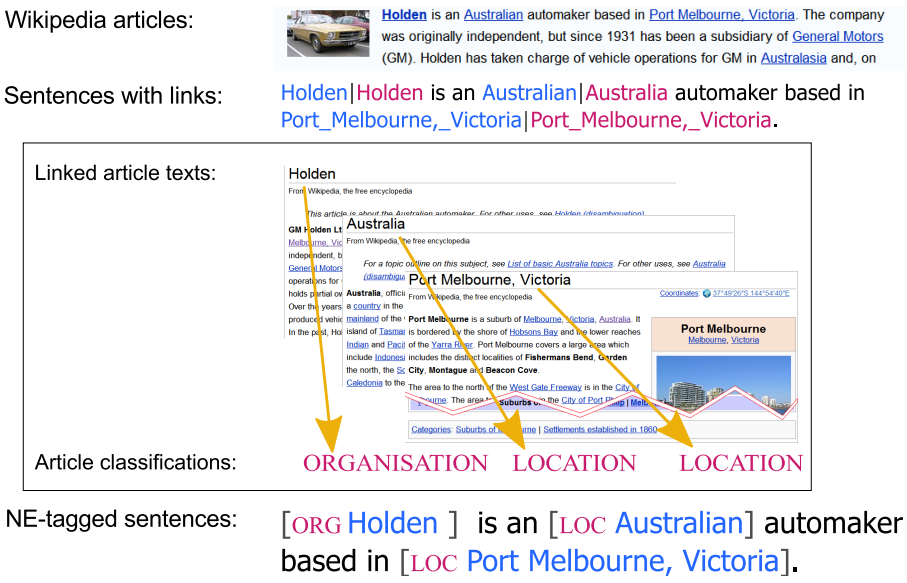

Figure 1: Deriving training sentences from Wikipedia text: sentences are extracted from articles; links to other articles are then translated to NE categories.

We process Wikipedia[1]—a free, enormous, multilingual online encyclopaedia—to create NE annotated corpora. Wikipedia is constantly being extended and maintained by thousands of users and currently includes over 3.6 million articles in English alone. When terms or names are first mentioned in a Wikipedia article they are often linked to the corresponding article. Our method transforms these links into NE annotations.

In Figure 1, a passage about Holden, an Australian automobile manufacturer, links both Australian and Port Melbourne, Victoria to their respective Wikipedia articles. The content of these linked articles suggest they are both locations. The two mentions can then be automatically annotated with the corresponding NE type (LOC). Millions of sentences may be annotated like this to create enormous *silver-standard* corpora—lower quality than manually-annotated gold standards, but suitable for training supervised NER systems for many more languages and domains.

We exploit the text, document structure and meta-data of Wikipedia, including the titles, links, categories, templates, infoboxes and disambiguation

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

data. We utilise the inter-language links to project article classifications into other languages, enabling us to develop NE corpora for eight non-English languages. Our approach can arguably be seen as the most intensive use of Wikipedia's structured and unstructured information to date.

1.1. Contributions

This paper collects together our work on: transforming Wikipedia into NE training data (Nothman et al., 2008); analysing and evaluating corpora used for NER training (Nothman et al., 2009); classifying articles in English (Tardif et al., 2009) and German Wikipedia (Ringland et al., 2009); and evaluating on a gold-standard Wikipedia NER corpus (Balasuriya et al., 2009). In this paper, we extend our previous work to a largely language-independent approach across nine of the largest Wikipedias (by number of articles): English, German, French, Polish, Italian, Spanish, Dutch, Portuguese and Russian.

We have developed a system for extracting NE data from Wikipedia that performs the following steps:

1. Classifies each Wikipedia article into an entity type;
2. Projects the classifications across languages using inter-language links;
3. Extracts article text with outgoing links;
4. Labels each link according to its target article's entity type;
5. Maps our fine-grained entity ontology into the target NE scheme;
6. Adjusts the entity boundaries to match the target NE scheme;
7. Selects portions for inclusion in a corpus.

Using this process, free, enormous NE-annotated corpora may be engineered for various applications across many languages.

We have developed a hierarchical classification scheme for named entities, extending on the BBN scheme (Brunstein, 2002), and have manually labelled over 4,800 English Wikipedia pages. We use inter-language links to project these labels into the eight other languages. To evaluate the accuracy of this method we label an additional 200–870 pages in the other eight languages using native or university-level fluent speakers.[2]

Our logistic regression classifier for Wikipedia articles uses both textual and document structure features, and achieves a state-of-the-art accuracy of 95% (coarse-grained) when evaluating on popular articles.

[2]These and related resources are available from `http://schwa.org/resources`.

We train the C&C tagger (Curran and Clark, 2003) on our Wikipedia-derived silver-standard and compare the performance with systems trained on newswire text in English, German, Dutch, Spanish and Russian. While our Wikipedia models do not outperform gold-standard systems on test data from the same corpus, they perform as well as gold models on non-corresponding test sets. Moreover, our models achieve comparable performance in all languages.

Evaluations on silver-standard test corpora suggest our automatic annotations are as predictable as manual annotations, and—where comparable—are better than those produced by Richman and Schone (2008).

We have created our own "Wikipedia gold" corpus (WIKIGOLD) by manually annotating 39,000 words of English Wikipedia with coarse-grained NE tags. Corroborating our results on newswire, our silver-standard English Wikipedia model outperforms gold-standard models on WIKIGOLD by 10% F-score, in contrast to Mika et al. (2008) whose automatic training did not exceed gold performance on Wikipedia.

We begin by reviewing Wikipedia's utilisation for NER, for language models and for multilingual NLP in the following section. In section 3 we describe our Wikipedia processing framework and characteristics of the Wikipedia data, and then proceed to evaluate new methods for classifying articles across nine Wikipedia languages in section 4. This classification provides distant supervision to our corpus derivation process, which is refined to suit the target evaluation corpora as detailed in section 5. We introduce our evaluation methodology in 6, providing results and discussion in the following sections, which together indicate Wikipedia's versatility for creating high-performance NER training data in many languages.

2. Background

Named entity recognition (NER), as first defined by the Message Understanding Conferences (MUC) in the 1990s, sets out to identify and classify proper-noun mentions of predefined entity types in text. For example, in

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

the word Paris is a personal name, a location, and an attribute of a hotel or organisation. Resolving these ambiguities makes NER a challenging semantic processing task. Approaches to NER are surveyed by Nadeau and Sekine (2007).

Part of the challenge is developing NER systems across different domains and languages, first evaluated in the Multilingual Entity Task (Merchant et al., 1996). The CoNLL NER shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) focused on language-independent machine-learning approaches to identifying persons (PER), locations (LOC), organisations (ORG) and other miscellaneous entities (MISC), such as events, artworks and nationalities, in English, German, Dutch and Spanish. Our work compares using these and other manually-annotated corpora against harnessing the knowledge contained in Wikipedia.

2.1. External knowledge and Named Entity Recognition

World knowledge is often incorporated into NER systems using *gazetteers*: categorised lists of names or common words. While extensive gazetteers of names in each entity type may be extracted automatically from the web (Etzioni et al., 2005) or from Wikipedia (Toral et al., 2011), Mikheev et al. (1999) and others have shown that relying on large gazetteers for NER does not necessarily correspond to increased NER performance: such lists can never be exhaustive of all naming variations, nor free from ambiguity. Experimentally, Mikheev et al. (1999) showed that reducing a 25,000-term gazetteer to 9,000 gave only a small performance loss, while carefully selecting 42 entries resulted in a dramatic improvement.

Kazama and Torisawa (2007) report an F-score increase of 3% by including many Wikipedia-derived gazetteer features in their NER system, although deriving gazetteers by clustering words in unstructured text yielded higher gains (Kazama and Torisawa, 2008). A state-of-the-art English CoNLL entity recogniser (Ratinov and Roth, 2009) similarly incorporates 16 Wikipedia-derived gazetteers. Unfortunately, gazetteers do not provide the crucial contextual evidence available in annotated corpora.

2.2. Semi-supervision and low-effort annotation

NER approaches seeking to overcome costly corpus annotation include automatic creation of silver-standard corpora and semi-supervised methods.

Prior to Wikipedia's prominence, An et al. (2003) created NE annotations by collecting sentences from the web containing gazetteered entities, producing a 1.8 million word Korean corpus that gave similar results to manually-annotated data. Urbansky et al. (2011) similarly describe a system to learn NER from fragmentary training instances on the web. In their evaluation

on English CoNLL-03 data, they achieve an F-score 27% lower (absolute difference with the MucEval metric) with automatic training than the same system trained on CoNLL training data. Nadeau et al. (2006) perform NER on the MUC-7 corpus with minimal supervision—a short list of names for each NE type—performing 16% lower than a state-of-the-art system in the MUC-7 evaluation. Like gazetteer methods, these approaches benefit from being largely robust to new and fine-grained entity types.

Other semi-supervised approaches improve performance by incorporating knowledge from unlabelled text in a supervised NER system, through: highly-predictive features from related tasks (Ando and Zhang, 2005); selected output of a supervised system (Wong and Ng, 2007; Wu et al., 2009; Liao and Veeramachaneni, 2009); jointly modelling labelled and unlabelled (Suzuki and Isozaki, 2008) or partially-labelled (Fernandes and Brefeld, 2011) language; or induced word class features (Kazama and Torisawa, 2008; Ratinov and Roth, 2009).

Given a high-performance NER system, phrase-aligned corpora and machine translation may enable the transference of NE knowledge from well-resourced languages to others (Yarowsky et al., 2001; Samy et al., 2005; Shah et al., 2010; Ma, 2010; Fu et al., 2011; Ehrmann et al., 2011).

Another alternative to expensive corpus annotation is to use crowdsourced annotation decisions, which Voyer et al. (2010) and Lawson et al. (2010) find successful for NER; Laws et al. (2011) show that crowdsourced annotation efficiency can be improved through active learning.

Unlike these approaches, our method harnesses the complete, native sentences with partial annotation provided by Wikipedia authors.

2.3. Learning Wikipedia's language

While solutions to NER and related tasks, e.g. NE linking (Bunescu and Paşca, 2006; Cucerzan, 2007; Mihalcea and Csomai, 2007) and document classification (Gabrilovich and Markovitch, 2006; Schönhofen, 2006) rely on Wikipedia as a large source of world knowledge, fewer applications exploit both its text and structured features. Wu and Weld (2007) learn the relationship between information in Wikipedia's infoboxes and the associated article text, and use it to extract similar types of information from the web. Biadsy et al. (2008) exploit the sentence ordering in Wikipedia's articles about people, harnessing it for biographical summarisation.

Wikipedia's potential as a source of silver-standard NE annotations has been recognised by Richman and Schone (2008), Mika et al. (2008), Nothman

et al. (2008) and others.

Richman and Schone (2008) and Nothman et al. (2008) classify Wikipedia's articles into NE types and label each outgoing link with the target article type. This approach does not label a sufficient portion of Wikipedia's sentences, since only first mentions are typically linked in Wikipedia, so both develop methods of annotating additional mentions within the same article.

Richman and Schone (2008) create NER models for six languages, evaluated against the automatically-derived annotations of Wikipedia and on manually-annotated Spanish, French and Ukrainian newswire. Their evaluation uses Automatic Content Extraction entity types (LDC, 2005), as well as MUC-style (Chinchor and Robinson, 1998) numerical and temporal annotations that are largely not derived from Wikipedia. Their results with a Spanish corpus built from over 50,000 Wikipedia articles are comparable to 20,000–40,000 words of gold-standard training data.

In Nothman et al. (2008), we produce silver-standard CoNLL annotations from English Wikipedia, and show that Wikipedia training can perform better on manually-annotated news text than a gold-standard model trained on a different news source. We also show that our Wikipedia-trained model outperforms newswire models on a manually-annotated corpus of Wikipedia text (Balasuriya et al., 2009).

Mika et al. (2008) use infobox information, rather than outgoing links, to derive their NE annotations. They treat the infobox summary as a list of key-value pairs, e.g. values Nicole Kidman and Katie Holmes for the spouse key in the Tom Cruise infobox, and their system finds instances of each value in the article's text, and labels it with the corresponding key.

They learn associations between NE types and infobox keys by tagging English Wikipedia text with a CoNLL-trained NER system. This mapping is then used to project NE types onto the labelled instances which are used as NER training data. They perform a manual evaluation on Wikipedia, with each sentence's annotations judged acceptable or unacceptable, avoiding the complications of automatic NER evaluation (see section 6.2). They find that a Wikipedia-trained model does not outperform CoNLL training, but combining automatic and gold-standard annotations in training exceeds the gold-standard model alone.

Fernandes and Brefeld (2011) similarly use Wikipedia links with automatic NE tags as training data, but use a perceptron model specialised for partial annotations to augment CoNLL training, producing a small but significant increase in performance.

2.4. Multilingual processing in Wikipedia

Wikipedia is a valuable resource for multilingual NLP with over 100,000 articles in each of 37 languages, and *inter-language links* associating articles on the same topic across languages. Wentland et al. (2008) refine these links into a resource for named entity translation, while other work integrates language-internal data and external resources such as WordNet to produce multilingual concept networks (Nastase et al., 2010; Navigli and Ponzetto, 2010; de Melo and Weikum, 2010). Richman and Schone (2008) and Fernandes and Brefeld (2011) use inter-language links to transfer English article classifications to other languages.

Approaches to cross-lingual information retrieval, e.g. Potthast et al. (2008); Schönhofen et al. (2008), or question answering (Ferrández et al., 2007) have mapped a query or document to a set of Wikipedia articles, and use inter-language links to translate the query. Attempts to automatically align sentences from inter-language linked articles have not given strong results (Adafre and de Rijke, 2006), probably because each Wikipedia language is developed largely independently; Filatova (2009) suggests exploiting this asymmetry for selecting information in summarisation. Adar et al. (2009) and Bouma et al. (2009) translate information between infoboxes in language-linked articles, finding discrepancies and filling in missing values. Thus NLP is able to both improve Wikipedia and to harness its content and structure.

3. Processing Wikipedia

Wikipedia's articles are written using MediaWiki markup[3], a markup language developed for use in Wikipedia. The raw markup is available in frequent XML database snapshots. We parse the MediaWiki markup, filter noisy non-sentential text (e.g. table cells and embedded HTML), split the text into sentences, and tokenise it.

MediaWiki allows nestable *templates* to be included with substitutable arguments. Wikipedia makes heavy use of templates for generating specialised formats, e.g. dates and geographic coordinates, and larger document structures, e.g. tables of contents and information boxes. We recursively expand all templates in each article and parse the markup using `mwlib`[4], a Python

[3] `http://www.mediawiki.org/wiki/Markup_spec`

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

library for parsing MediaWiki markup. We extract structured features and text from the parse tree, as follows.

3.1. Structured features

We extract each article's section headings, category labels, inter-language links, and the names and arguments of included templates. We also extract every outgoing link with its anchor text, resolving any redirects.

Further processing is required for *disambiguation pages*, Wikipedia pages that list the various referents of an ambiguous name. The structure of these pages is regular, but not always consistent. Candidate referents are organised in lists by entity type, with links to the corresponding articles. We extract these links when they appear zero or one word(s) after the list item marker. We apply this process to any page labelled with a descendant of the English Wikipedia Disambiguation pages category or an inter-language equivalent.

We then use information from cross-referenced articles to build reverse indices of incoming links, disambiguation links, and redirects for each article.

3.2. Unstructured text

All the paragraph nodes extracted by `mwlib` are considered body text, thus excluding lists and tables. Descending the parse tree under paragraphs, we extract all text nodes except those within references, images, math, indented portions, or material marked by HTML classes like `noprint`. We split paragraph nodes into sentences using Punkt (Kiss and Strunk, 2006), an unsupervised, language-independent algorithm. Our Punkt parameters are learnt from at least 10 million words of Wikipedia text in each language.

Tokenisation is then performed in the parse tree, enabling token offsets to be recorded for various markup features, particularly outgoing links. We slightly modify our Penn Treebank-style tokeniser to handle French and Italian clitics, and non-English punctuation. In Russian, we treat hyphens as separate tokens to match our evaluation corpus.

3.3. Wikipedia in nine languages

We use the English Wikipedia snapshot from 30 July, 2010, and the subsequent snapshot for the other eight languages[5], together constituting the ten largest Wikipedias excluding Japanese (to avoid word segmentation).

[5] All accessed from `http://download.wikimedia.org/backup-index.html`

Wiki	Language	Snapshot	Articles	Disamb.	Categ.	Tokens
EN	English	2010-07-30	3 398 404	200 113	605 912	1 205 569 685
DE	German	2010-08-15	1 123 266	114 404	89 890	389 974 559
FR	French	2010-08-02	980 773	61 678	150 920	293 287 033
IT	Italian	2010-08-10	723 722	45 253	106 902	211 519 924
PL	Polish	2010-08-03	721 720	40 203	69 744	126 654 300
ES	Spanish	2010-08-06	632 400	27 400	119 421	254 787 200
NL	Dutch	2010-08-04	617 469	37 447	53 242	123 047 016
PT	Portuguese	2010-08-04	598 446	21 065	94 117	120 137 554
RU	Russian	2010-08-10	572 625	44 153	140 270	156 527 612

Table 1: Summary of Wikipedias used in our analysis. Columns show the total number of articles, how many of them are disambiguation pages, the number of category pages (though not all contain articles), and the number of body text tokens.

Feature \ Lang.	EN		DE		ES		NL		RU	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Incoming links	67.9	11	38.4	8	36.2	5	41.0	7	46.18	6
Outgoing links	73.8	30	43.3	24	41.2	29	46.8	23	55.6	29
Redirects	1.2	0	0.7	0	1.8	1	0.4	0	1.2	0
Categories	5.6	4	3.5	3	2.8	2	2.0	2	4.3	3
Templates	7.9	4	3.6	2	3.7	2	5.0	2	8.3	4
Tokens	354.8	135	347.2	196	402.9	177	199.3	95	273.4	111
Sentences	14.8	6	17.6	10	14.8	7	10.6	5	14.5	7
Paragraphs	5.3	3	6.0	4	6.2	3	3.9	3	5.6	3

Table 2: Mean and median feature counts per article for selected Wikipedias.

The languages, snapshot dates and statistics are shown in Table 1. English Wikipedia at 3.4 million articles is about six times larger than Russian, our smallest Wikipedia. All of the languages have at least 100 million words—comparable in size to the British National Corpus (BNC, 2007).

These statistics also highlight disparities in language and editorial approach. For instance, German has substantially fewer, and Russian substantially more, category pages per article; the reverse is true for disambiguation pages, with one for every 9.8 articles in German.

Table 2 shows mean and median statistics for selected structured and text content in Wikipedia articles. English articles include substantially more categories, incoming and outgoing links on average than other languages, which together with its size highlights its greater development and diversity of contributors than other Wikipedias.

4. Classifying Wikipedia articles

We first classify Wikipedia's articles into a fixed set of entity types, which can then label links to those articles. Since classification errors transfer into our NER models, high accuracy is essential. To facilitate this, we reimplement three classification approaches from the literature, extending our state-of-the-art method to nine languages, including novel multilingual features (Section 4.2). We use two article sampling approaches to create collections of manually-classified Wikipedia articles (Section 4.3); Section 4.4 considers the projection of this data to other Wikipedia versions and languages.

4.1. Background

Wikipedia's category hierarchy is a *folksonomy* (Strube and Ponzetto, 2006), making it unsuitable for many semantic applications. Suchanek et al. (2007) class each Wikipedia category as either *conceptual*—Holden is a Motor vehicle company; *relational*—Holden was established in 1856; *thematic*—Holden has theme Holden; or *administrative*—Date of birth missing. Non-conceptual categories may include articles of many different types. For example, products (Apple III), fictional characters (Yoda) and facilities (Cairns Tropical Zoo) are all members of the 1980 introductions category. Infoboxes are strongly correlated to entity type, but only have high coverage on LOC and PER articles.

Since Wikipedia does not have a direct source of entity types, there has been interest in mapping articles to existing ontologies such as WordNet (Ruiz-Casado et al., 2005; Suchanek et al., 2008; Ponzetto and Navigli, 2009) and Cyc (Medelyan and Legg, 2008), or classifying them into coarser schemes using heuristics (Toral and Muñoz, 2006; Bhole et al., 2007; Richman and Schone, 2008) and semi-supervised (Watanabe et al., 2007; Dakka and Cucerzan, 2008; Nothman et al., 2008) or fully supervised modelling approaches (Bhole et al., 2007; Dakka and Cucerzan, 2008; Tardif et al., 2009; Tkatchenko et al., 2011).

4.2. Article classification approaches

We compare a baseline heuristic, a semi-supervised and a fully-supervised monolingual classification approach from the literature. We then provide three ways to extend the latter approach to multiple languages.

NE type	Keyword example	Quantity
LOC	Rivers of, Towns	30
ORG	Organizations, musical groups	27
PER	Living People, Year of birth	36
MISC	Television series, discographies	27
NON	Years, Wikipedia	18
DAB	Disambiguation	3

Table 3: Examples and quantity of category keywords for each coarse-grained type.

4.2.1. Classification with category keyword heuristics

Richman and Schone (2008) produced a set of key phrases from English Wikipedia category names that correspond to PER, LOC, ORG and other entity types (but not MISC or non-entities). When classifying, each article's categories are matched against the phrases, backing off to parents and grandparents of those categories, until support for a particular type exceeds a threshold. If the threshold is not met, the article's type remains *unknown*. Each key phrase votes with a manually set weight (Richman, 2010).

For example, Queanbeyan has categories Cities in New South Wales, Populated places established in 1838, Queanbeyan and Australian Aboriginal placenames. The key phrase *Cities* might vote for type LOC, but the other categories do not match any keywords directly. This may not exceed the threshold, so the parents of unmatched categories are also considered. The Queanbeyan category has parent categories Cities in New South Wales and Categories named after populated places in Australia, so *Cities* again votes for Queanbeyan as a LOC.

We attempt to replicate Richman and Schone (2008), but the key phrases were unavailable and many of the details were underspecified, so our replica is approximate. For instance, in the case of a tie between types, we randomly choose a type, and we use a support threshold of one to discourage unknowns.

We have created our own list of key phrases, starting with their published examples and adding phrases from large type-homogeneous categories, if the other categories matching those phrases are also homogeneous. We have also added phrases for matching MISC, non-entities (NON), disambiguation pages (DAB). Table 3 shows some examples of the 141 keywords, with the full list in Appendix A.

4.2.2. Classification with keyword bootstrapping

Nothman et al. (2008) developed a semi-supervised approach to classify

English Wikipedia articles with relatively few labelled instances.[6] A small number of structural features are extracted from each article. Iteratively, confident mappings from feature to NE type are inferred from classified articles, and the classifier is again applied to all of Wikipedia. Over three iterations (empirically selected), the mapped feature space grows, and the proportion of unknown articles decreases.

The following features are used in bootstrapping:

- *Plural category heads:* Suchanek et al. (2007) suggested that categories with plural head nouns are usually conceptual, such as *cities*, *places* and *placenames*—but not *Queanbeyan*—in the Queanbeyan example above. We extract head unigrams and collocated bigrams.

- *Definition noun:* Since many of Wikipedia's articles begin with a definition, we extract the head unigram or bigram following a copula, if any, from the first sentence, following Kazama and Torisawa (2007).

An article is assigned the type most supported by its features, remaining unknown in a tie. Specialised heuristics identify non-entity articles (NON and DAB), including the capitalisation of incoming anchor text and title keyword matching for disambiguation and list pages.

4.2.3. Classification as text categorisation with structured features

The approaches above, along with many in the literature, have relied on the precision of Wikipedia's structured features. However, the most successful have used statistical models of its body text (Dakka and Cucerzan, 2008), which may also be more readily ported to new languages.

In Tardif et al. (2009), we compare Naïve Bayes (NB) and Support Vector Machines (SVM) for classifying Wikipedia articles using bag-of-words and structured features. Here we use the `liblinear` (Fan et al., 2008) in the logistic regression with L2 regularization mode.

Dakka and Cucerzan (2008) suggest that most humans will be able to classify an article after reading its first paragraph. We therefore use the words of the first paragraph, first sentence and title as separate feature groups. In addition, we use template names, and the contents of *infobox*, *sidebar* and *taxobox* templates. These templates often contain a condensed set of

[6] We extended this method to German in Ringland et al. (2009).

important facts relating to the article, and so are powerful additions to the bag-of-words representation of an article.

Monolingual classification. Having projected our gold-standard classifications to nine other languages via inter-language links, we train monolingual article classifiers for each language.

Multilingual classification. Each topic is likely to have different coverage in different Wikipedias. We therefore present two methods for combining the knowledge found in equivalent articles in multiple languages:

VOTED We learn monolingual classifiers for each language, and classify an article as the most popular vote of its inter-language equivalents, backing off to English (our best-performing monolingual model) in a tie.

UBER We merge the feature spaces of language-linked articles across the nine languages, prefixing each feature name with the language it came from. We model this extended feature space, and classify each article using features from it and its cross-lingual equivalents.

4.3. Annotating gold-standard classifications

We use manual classifications of Wikipedia pages as indirect supervision for NER and to evaluate our classifiers. However, it is unclear how best to sample articles. Random sampling produces more challenging instances for evaluation, but Nothman et al. (2008) found it under-samples entity types that have few instances but are essential to NER, such as countries. Selecting only popular articles provides advantages for multilingual processing, and should assist with classifying the entities most frequent in text. We therefore present two sets of labelled articles, POPULAR and RANDOM. Both are available for download.[7]

4.3.1. POPULAR *labelled corpus*

As previously presented (Tardif et al., 2009; Ringland et al., 2009), we produced a corpus of approximately 2,300 English Wikipedia articles (March 2009 snapshot), including the 1,000 most frequently-accessed pages of August 2008[8] and otherwise the pages with most incoming links. We required

[7]From `http://schwa.org/resources`.

[8]According to the Wikipedia proxy logs from `http://dammit.lt/wikistats`.

that each article include inter-language links to all ten largest language Wikipedias. This favoured typically longer, high-quality articles and about popular and useful subjects. It also largely avoided stubs and automatically-generated pages (Ringland et al., 2009).

Each article was double-annotated with a single fine-grained type. We extended the hierarchical scheme from BBN (Brunstein, 2002), allowing us to use BBN in later NER evaluations. However, Sekine's (2002) scheme would have been equally suitable. In order to get an estimate of inter-annotator agreement, about 1000 articles were annotated independently, achieving 97.5% agreement, calculated over a finer type schema than used in the experiments below (agreement on coarse-grained NE types was 99.5%). Subsequently, annotation was periodically paused to resolve conflicts.

4.3.2. RANDOM *labelled corpus*

The articles in POPULAR are not representative of Wikipedia's long tail of obscure articles, stubs, and automatically-generated pages. We therefore annotated a random sample of Wikipedia's articles to more accurately reflect its make-up: 2,500 in English, 850 in German, and 200 in each of the seven other languages. We annotated a few extra articles to allow for MediaWiki extraction errors.

Each article was classified by at least two annotators, of whom at least one was a native speaker or had university-level language skills in the appropriate language. RANDOM presented many more edge cases for classification than POPULAR, making its annotation more time consuming. Nonetheless, all discrepancies were resolved at the NE type granularity used in the present work.

The annotation followed the method we developed in Tardif et al. (2009): annotators were able to add fine-grained types to the hierarchy as required, leading to very fine distinctions; SUBURB, ADMIN DISTRICT and STATE are all subtypes of LOC:GPE. This resulted in 154 types, which were grouped together to create 62 very fine-grained types, 19 fine-grained types and 6 coarse-grained types. Of the original 154 categories, 67 map to NON, 29 to LOC, 14 to ORG, 4 to PER, and 37 to MISC. Table 4 gives examples from POPULAR and RANDOM; the mappings are available for download.[9]

For languages where two fluent speakers were not available, we used

[9]From `http://schwa.org/resources`.

Fine-grained NE type	POPULAR example	RANDOM example
LOCATION (LOC):		
TOWN/CITY	Bangkok	Terese, California
GPE	Aceh	Castel di Judica
FACILITY	Beijing National Stadium	Urashuku Station
OTHER	Great Wall of China	Bressay
ORGANISATION (ORG):		
BAND	Blink-182	Transitional (band)
CORPORATION	Atari	Logitech
OTHER	Interpol	Manchester A's
PERSON (PER):		
PERSON	John F. Kennedy	Peter McConnell
OTHER	Yoda	Bold Reason
OTHER (MISC):		
EVENT	2008 South Ossetia war	2006 J&S Cup
NORP	Hungarian People	Norts
WORKOFART	Entourage (TV series)	Man of the Hour
PRODUCT	AK-47	Bugatti Type 53
MISCELLANEOUS	Capoeira	World Habitat Awards
NON-ENTITY (NON):		
LIFE	Capsicum	Platysilurus
SUBSTANCE	DNA	Mango oil
OTHER	Blitzkrieg	Canadian units
DISAMBIGUATION (DAB)	California (disambiguation)	Lip (disambiguation)

Table 4: Fine-grained NE types with examples from POPULAR and RANDOM collections.

Google Translate[10] to assist in classification decisions. This approach makes subtle, very fine-grained distinctions difficult. For example, the German word Gemeinde translates to town, borough, or parish depending on use, each of which may belong in a different LOC subtype.

In other cases, the extremely fine granularity created annotation disputes. For example, annotators disagreed on whether Manhattan, an island borough of New York City, should be classified as its own independent city/town, a suburb, or an island. The annotators resolved their disagreements and annotation guidelines were updated continuously.

[10] http://translate.google.com

Corpus		No. of articles	% inter-lang Any	% inter-lang EN	Coarse type distribution (%) LOC	ORG	PER	MISC	NON	DAB
POPULAR	EN	2,322	100	-	28	11	11	16	30	4
RANDOM	EN	2,531	46	-	20	10	26	18	16	10
RANDOM	DE	872	57	49	19	11	33	13	12	12
RANDOM	ES	203	58	51	28	10	19	19	20	4
RANDOM	FR	210	61	54	22	5	25	20	20	8
RANDOM	IT	203	71	64	30	4	23	19	18	6
RANDOM	NL	286	73	63	34	9	17	15	17	8
RANDOM	PL	210	68	60	36	4	30	13	11	6
RANDOM	PT	202	72	66	38	6	17	15	19	5
RANDOM	RU	223	62	51	30	8	26	14	13	9

Table 5: Gold-standard classification statistics per corpus: size; percentage of articles with inter-language links to any/English Wikipedia; distribution of coarse entity types, disambiguation pages (DAB) and non-entities (NON).

Table 5 compares the final sizes of POPULAR and RANDOM samples, and their distributions over coarse-grained entity types. Within English Wikipedia, POPULAR contains far more LOC and NON articles, and RANDOM is skewed more toward PER and MISC. The RANDOM type distribution varies greatly between languages; however, for most, the sample size is small.

4.4. Projecting data between Wikipedia versions

Wikipedia articles are referred to by title, which does not ensure accurate linking since articles may be renamed over time. Our data maps Wikipedia titles from 2008-10 Wikipedia snapshots to NE types, and we need to transfer these types to newer Wikipedia snapshots, and across inter-language links.

Sorg and Cimiano (2008) analysed the coverage of inter-language links between English and German Wikipedias from October 2007: 46% of German pages linked to English, and 14% of English pages had German links. Of the links present, around 95% were bijective, i.e. linking from EN to its DE equivalent, and back to the same EN page. Table 5 gives the proportion of each language's articles with inter-language links. Ringland et al. (2009) checked the integrity of a sample of English-German links, and found very few were erroneous.[11] Confusion between an entity article and a disambiguation

[11]Bijective links may still have errors, since editors may insert language links without ensuring that the target page exists, or before it is created. The titles may be translations, but the articles on different topics (commonly one is a disambiguation page and the other

TEXTCAT	Coarse-grained			Fine-grained		
classifier	Precision	Recall	F-score	Precision	Recall	F-score
EN	94.6	94.6	94.6	89.9	89.7	89.8
DE	94.1	93.9	94.0	89.7	88.6	89.2
ES	93.9	93.7	93.8	88.6	87.9	88.2
FR	93.9	93.7	93.8	89.8	88.7	89.3
IT	93.9	93.7	93.8	89.6	88.7	89.2
NL	94.0	93.8	93.9	89.1	88.1	88.6
PL	93.1	92.7	92.9	88.9	87.7	88.3
PT	93.2	93.0	93.1	88.5	87.1	87.8
RU	93.6	93.3	93.5	88.0	87.1	87.6
VOTED	94.9	94.8	94.9	89.9	88.9	89.4
UBER	94.9	94.8	94.8	89.9	89.3	89.6

Table 6: Coarse and fine-grained results over POPULAR for multilingual text categorisation.

page of the same title are a common source of error.

We assume that NE type is maintained across an inter-language link and for an article with the same name in different snapshots of Wikipedia. We do not manually check this, instead applying a naive approach: look up the title, following any redirects; if no such page exists, or the target is a section (not a full article), remove the instance.

For example, EN Yoda links to the Yoda section of DE Star Wars Characters, and so is discarded in DE. In some cases, two different articles link to the same title in another language, which is especially problematic when their types differ; Gulf Coast Wing (ORG) and Aviation (NON) both appear in POPULAR, but both link to Aviation in other languages.

Changes over time are handled similarly: Anglesey now redirects to Isle of Anglesey, but the projected type is still valid. Death (band) now redirects to the subsection *Music* of Death (disambiguation), and so is discarded.

In the present work, we do not project across RANDOM language links for classification.

4.5. Results and discussion

We report 10-fold cross-validated precision, recall and F-score, evaluating over: language; classification approach; use of POPULAR, RANDOM or their combination; and fine (18 types) vs coarse (6) entity types.

not). Further, bots exist to check for or ensure bijectivity.

Train \ Test	POPULAR	RANDOM	POP+RAND
POPULAR	94.6	75.1	83.5
RANDOM	91.7	90.4	90.7
POP+RAND	95.5	90.7	93.1

Table 7: Coarse-grained English TEXTCAT classification F-score when training and testing over different datasets.

The results in Table 6 extend Tardif et al.'s (2009) approach to 9 languages, relying on POPULAR's full complement of inter-language links. The high coarse-grained performance (94.6%) on English is similar to that previously reported on an older snapshot of Wikipedia; other languages' monolingual classifiers perform less than 2% worse, proving this approach is effective independent of language. VOTED and UBER results are almost identical, and only differ marginally from the English monolingual result, but are often better than other monolingual results. Fine-grained F-scores are 4-6% lower than the coarse equivalents.

Although results on POPULAR are promising in all languages, it is not clear how this applies to Wikipedia's long tail. To explore this, we consider every train-test combination of POPULAR, RANDOM and their union (POP+RAND), with coarse-grained English results shown in Table 7. POPULAR alone is very poor training for RANDOM, achieving only 75%, while top performance on RANDOM is about 5% lower than on POPULAR. Independent of the test corpus, performance is best when trained with POP+RAND.

This result may be surprising when evaluating on POPULAR, given how much noise may be introduced by RANDOM. However, the combined dataset is about twice as large, and consists of both the longer, better-edited pages with richer features from POPULAR and the variety of RANDOM. We select POP+RAND for the remaining experiments, given its high performance and its relative suitability for NER.

Table 8 compares the coarse-grained performance of the three approaches. TEXTCAT significantly outperforms the BOOTSTRAP approach and the KEYWORD baseline, and has the most uniform distribution of performance over types. KEYWORD performs particularly poorly on the most diverse types, MISC and NON, though Richman and Schone (2008) did not develop classifiers for these types. BOOTSTRAP performance is close to TEXTCAT on PER and ORG, but is greatly exceeded on LOC, NON and DAB. Overall, PER, LOC and DAB are easiest to classify, while ORG and MISC are the hardest, a trend

NE type	KEYWORD	BOOTSTRAP	TEXTCAT	VOTED	UBER
LOC	57.8	89.7	96.8	96.6	96.5
ORG	58.1	84.1	87.5	87.3	86.4
PER	86.7	97.0	97.2	97.5	97.2
MISC	45.9	80.7	87.5	87.8	86.8
NON	45.3	83.1	91.7	91.6	92.0
DAB	80.8	77.4	94.5	93.9	94.3
Total	64.6	87.0	93.1	93.1	92.9

Table 8: English coarse-grained classification F-score over POP+RAND.

NE type	EN	DE	ES	FR	IT	NL	PL	PT	RU
LOC	96.8	96.9	97.8	97.8	97.4	97.7	97.6	98.1	97.8
ORG	87.5	87.4	88.0	89.0	90.2	89.5	91.1	89.9	89.3
PER	97.2	97.5	95.3	95.5	97.8	96.0	94.4	94.0	96.3
MISC	87.5	83.5	86.0	86.2	85.3	84.5	84.0	83.8	84.9
NON	91.7	91.5	92.7	93.0	92.5	92.1	91.2	91.7	92.0
DAB	94.5	95.7	97.7	92.2	93.5	92.6	95.6	94.9	93.2
Total	93.1	92.8	93.4	93.4	93.5	93.0	92.8	92.9	93.2

Table 9: Coarse-grained classification F-score for monolingual TEXTCAT over POP+RAND.

which continues across all languages (Table 9).

In Table 10 we show fine-grained classification results in five languages,[12] VOTED and UBER. Performance is low for types which have few training instances, are diverse, and lack defining article structure (such as infoboxes, categories, or geographical coordinates). NON-ENTITY acts as the default type due to its diversity and high frequency: for every classifier, instances of each other type are misclassified as NON-ENTITY, including Bugatti Type 53 (PRODUCT), British Japan Consular Service (ORG:OTHER), Battle of Pistoria (EVENT) and The Star-Spangled Banner (WORKOFART). NORP[13] is difficult to identify in all classifiers, and in RU all NORP articles are classified as NON-ENTITY.

Entities which function as multiple types challenge our single-label classifiers. While the Popeye and James Bond articles specify that they are about

[12]We use these languages for NER evaluation due to available gold-standard corpora.

[13]NORP is a term used by BBN (Brunstein, 2002) to refer to national, organisational, religious, or political affiliations in an adjectival form. We use it for nationalities and other non-organisational named groups of people, which are generally considered MISC in CONLL NER.

NE TYPE	Count	EN	DE	ES	NL	RU	VOTED	UBER
LOC:TOWN/CITY	568	94.7	96.4	95.4	95.8	95.8	95.2	95.4
LOC:GPE	345	86.9	89.9	88.3	89.8	89.9	88.0	87.7
FACILITY	141	79.9	71.8	31.2	61.5	37.0	76.7	79.3
LOC:OTHER	134	82.9	73.7	55.3	55.4	78.7	82.5	82.9
ORG:BAND	101	93.8	97.1	98.0	98.8	98.7	94.4	92.2
ORG:CORPORATION	158	87.4	87.3	92.0	87.7	91.0	88.1	87.7
ORG:OTHER	218	76.5	64.2	64.9	59.3	55.5	75.4	74.9
PER:PERSON	871	97.1	99.4	96.9	96.0	97.6	97.6	96.4
PER:OTHER	66	61.1	54.3	69.2	66.7	64.9	64.2	58.8
EVENT	138	80.6	77.9	68.5	71.0	75.3	77.3	78.1
NORP	32	41.9	37.0	56.0	51.9	0.0	35.9	41.9
WORKOFART	359	89.3	86.3	87.8	87.7	91.5	89.0	87.9
PRODUCT	228	87.6	84.8	89.2	87.4	83.9	87.1	86.8
MISCELLANEOUS	65	50.5	5.0	24.4	37.2	29.3	43.7	50.0
NON-ENTITY:LIFE	276	95.0	94.0	93.8	92.3	93.7	94.6	95.0
NON-ENTITY:SUBSTANCE	111	73.2	70.7	70.1	73.3	69.9	67.1	74.4
NON-ENTITY	711	83.7	81.5	82.6	81.5	80.7	81.1	83.4
DAB	321	94.6	95.2	98.2	92.6	93.8	93.7	94.3
TOTAL	4,843	88.7	88.4	87.6	87.4	87.4	88.3	88.4

Table 10: Fine-grained TEXTCAT classification F-score for five monolingual models, VOTED and UBER (evaluating for EN), over POP+RAND. Count is the total number of gold instances of each type, though fewer are available in each language.

fictional characters (PER:OTHER), they also discuss the related media franchises, so both are incorrectly classified WORKOFART. Similarly, FACILITY articles are often confused with LOC and ORG types.

Some misclassifications arise from debatable down-mappings of our annotation types. For instance, we group disambiguation and list pages together as DAB, but many list pages include additional content that makes them more similar to NON than the largely-fixed structure of DAB pages.

Other mistakes are due to our naive approach to modifications of Wikipedia (see Section 4.4); Eagles now is a redirect to the animal Eagle, whereas when the page was annotated, it described the band, The Eagles.

Our overall results for fine-grained classification of English Wikipedia articles compare favourably to Tkatchenko et al. (2011) who report approximately 75% accuracy over randomly-sampled articles labelled with 18 types; we attain 85% accuracy for cross-validation on RANDOM.

4.6. Summary

We have developed accurate coarse- and fine-grained Wikipedia article classifiers for nine languages. These have been evaluated on both a high-quality POPULAR gold standard and a noisier but more representative RANDOM gold standard. We find that the combination of POPULAR and RANDOM training data produces the best results. This combined data set trains our UBER multilingual text-categorisation approach, allowing us to classify all Wikipedia articles and label links to them as NE tags.

5. Designing a training corpus

Under the broad definition of NER, our basic approach to creating a Wikipedia-derived NE-annotated corpus described in Section 1 produces reasonable annotations. However, in order to automatically produce a corpus comparable to existing gold standards, heuristic selection and further refinement of the annotations is required.

While both gold-standard corpora and Wikipedia have some inconsistencies in their markup (Nothman et al., 2009), the former are generally created with strict annotation guidelines, by a small number of annotators, and for the precise purpose of NER. Not surprisingly, Wikipedia's link spans and targets often do not directly correspond to the NE annotation scheme of a particular evaluation corpus. Through a set of heuristics, we design Wikipedia corpora that better approximate existing gold standards.

In this section, we describe methods we apply to reduce the differences between Wikipedia and gold-standard NER corpora, beginning with an overview of our approach to identifying these differences.

5.1. Comparing NER *corpora*

In Nothman et al. (2009) we describe three approaches for identifying inconsistencies within and between corpora with phrasal annotations:

N-gram tag variation: search for internal variations, where the same text span with different tags but identical context appears multiple times in the corpus, as proposed by Dickinson and Meurers (2003).

Type frequency: compare the entity type distribution across corpora, by extracting all entity mentions, representing them by their orthography or POS-tag sequences, and comparing aggregates over each type.

Tag sequence confusion: as a simple confusion matrix cannot be applied to phrasal tagging, analyse confusion between the type of each predicted entity and the corresponding gold-standard tag sequence (which may include entity and non-entity portions), and between each gold-standard entity and the corresponding predicted tag sequence.

We apply these methods systematically to derive an annotated corpus from English Wikipedia, by comparing to CoNLL and BBN gold-standard annotations. Aware of key issues from our work in English, we mostly use direct inspection to apply similar methods in other languages. This analysis was performed by the authors (native English speakers) with contributions from volunteers familiar with the Cyrillic alphabet; a second-language speaker of German with some Dutch knowledge; and a native speaker of Spanish.

5.2. Selection approach

We include portions of articles in our training corpus using criteria based on *confidence* that we have correctly identified all entities within that portion, and on its *utility* for learning NER. The size and redundancy of Wikipedia's content allows us to discard large portions of the available data.

We consider the following baseline criteria:

Confidence: all capitalised words are linked to articles of known entity type.

Utility: at least one entity is marked.

This confidence criterion was designed for general-domain NER in English where capitalisation usually corresponds closely to NEs.

In prior work, we applied our baseline criteria to each sentence in Wikipedia. We now consider two additional approaches: (a) upon identifying a token which fails the criteria, remove the containing parenthesised expression, or the whole sentence if not in parentheses; (b) do not require whole sentences, instead selecting the longest confident fragment of some utility from each sentence, following Mika et al. (2008). Often Wikipedia's parenthesised expressions contain glosses into other languages and other noisy material, removed by (a). Using sentence fragments slightly reduced our NER performance, while parenthesis removal improved performance and is used below.

Our confidence criterion is overly restrictive since: it extracts a low proportion of sentences per article; it is biased towards short sentences; and each

entity mention is often linked only on its first appearance in an article, so we are more likely to include fully-qualified names than shorter referential forms (surnames, acronyms, etc.) found later in the article. Many conventionally capitalised words, which do not correspond to entities, still cause problems and are discussed below.

5.3. Inferring additional links

In order to increase our coverage of Wikipedia sentences, our system infers additional links. Since Wikipedia style dictates that only the first mention of an entity should be linked in each article, we try to identify other mentions of that entity in the same article. We begin by compiling a list of aliases for each article. Then for any article in which we are attempting to infer links, we produce a trie containing the aliases of the current article and all outgoing links. We attempt to find the longest matching string within the trie, starting at each unlinked token in an article, and assign its entity type to the matching text. Aliases for an article A include:

Type 1 The title of A and those of redirects[14] to A (with expressions following a comma or within parentheses removed);

Type 2 Titles (and redirect titles) of disambiguation pages linking to A, enabling, e.g.: AMP as an alias for AMP Limited and Ampere, and Howard an alias for Howard Dean and John Howard;

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

We vary the level of inference (e.g. level 2 consists of types 1 and 2) below. The following exceptions help avoid over-generation and noisy links:

- aliases matching a list of stop-words from NLTK (Loper and Bird, 2002);
- aliases whose link boundaries would be adjusted (Section 5.6); and
- aliases which are the concatenation of another alias with lowercase words (e.g. Australian is a better match than Australian people, though both are redirect aliases for Australia).

Although it introduces many spurious links due to noisy data sources and ambiguity, this additional link inference allows for more variation in how our NE-annotated corpus refers to an entity.

[14]Redirect pages make articles accessible through non-canonical titles.

Capitalised?	EN	DE	ES	NL	RU
Most entities	Y	Y	Y	Y	Y
Sentence initial	Y	Y	Y	Y	Y
Common nouns	N	Y	N	N	N
Days and months	Y	S	N	N	N
Personal titles	Y	Y	Y	S	N
Acronyms	Y	S	Y	S	Y
Roman numerals	–	–	–	Y	Y
Adjectival entities	Y	N	N	S	N

Table 11: A summary of conventional capitalisation applying, as a (breakable) rule, in our evaluation languages. Y = yes, capitalised; N = not capitalised; S = sometimes; – = used infrequently outside of entities, so largely irrelevant.

5.4. Anomalous capitalisation

Non-entity links which are capitalised and all-lowercase entity links may be problematic as NE annotations. They often result from mis-classification, or to a link including a NE in its title, e.g. Greek alphabet or Jim Crow laws, in which case it would be incorrect to leave the reference untagged. Lowercase entity links result from common noun phrase references, e.g. in In the Ukraine, anarchists fought in the civil war ..., the anchor civil war links to Russian Civil War. Text containing capitalised NON links (except in German) or lowercase entity links is discarded, except for entities like gzip that Wikipedia explicitly marks as a lowercase title.

5.5. Conventional capitalisation

European orthographic systems that distinguish alphabet case do so in different contexts, as summarised according to our analysis in Table 11. As an exception to our confidence criterion, we attempt to identify non-entity capitalised words for inclusion in our corpus.

Sentence initial. If a word which begins a sentence or follows some punctuation (semicolon, left-quote, etc.) is capitalised and unlinked, we consider it safe for inclusion if it is linked to a non-entity (NON) article, or is found on a list of commonly lowercase words. For each language, this list consists of frequent sentence starters from our sentence boundary detection models (Kiss and Strunk, 2006), and a list of words which occurred at least 50 times lowercase, and at least 50 times sentence-initially, in 100,000 Wikipedia articles.

Common nouns. In German, all nouns are capitalised, presenting a challenge for our confidence criterion. We compile a list of common nouns using dict.cc, a collaboratively-constructed German-English dictionary. We utilise its translation database[15] to ignore German unigrams that have only lowercase glosses. Of 251,846 such German entries, we found 230,489 had only lowercase translations, while 17,538 had only capitalised translations.

German CoNLL frequently mentions the German currency, the *Mark*, which is not identified as a common noun because it is identical to a personal name. We use a list of currency names from German Wikipedia and label them as common nouns when the prior two tokens contain a cardinal. We also mark the word I in English, assuming it is the personal pronoun.

Days and months. EN and DE month and day names are marked for inclusion.

Personal titles. Personal titles (e.g. Dr., Brig. Gen., Prime Minister-elect) are conventionally capitalised in English and other languages, but are non-entities in CoNLL-style NER (although some are included in BBN). Titles are sometimes linked in Wikipedia, but allowing a link text like U.S. President as a non-entity would leave the entity U.S. unlabelled. Titles often appear immediately before PER mentions, so the most frequent instances can be compiled into a list of known titles.

In English, these are manually filtered—removing LOC or ORG mentions—and supplemented with abbreviated titles extracted from BBN, producing a list of 384 base forms, 11 prefixes (e.g. Vice) and 3 suffixes (e.g. -elect). Using this gazetteer, titles are identified and stripped of erroneous NE tags. In German, we extracted 203 base titles from Wikipedia, with 5 morphological suffixes and 1 prefix.

Acronyms. Our initial approach to acronyms—including all unlinked uppercase words—degraded performance, but we found that including all-uppercase words linked to non-entity articles was successful.

Roman numerals. Russian and Spanish make extensive use of capitalised Roman numerals. We identify them with a regular expression and include them in our corpus when unlinked.

[15] From `http://www1.dict.cc/translation_file_request.php`, accessed 2010-10-21

Adjectival forms. In English, the adjectival forms of entities, such as nationalities or religions e.g. American or Islamic, are capitalised. Both CoNLL and BBN (see Section 6.1) annotate them as MISC. In Wikipedia, nationalities often link to LOC articles.

For EN, DE, NL and RU, we compile a list of nationalities from Wikipedia words that our POS tagger marks as an adjective, and morphological variants to handle cases like Americans.[16] Each link matching this list is relabelled MISC. In German and Russian (and Dutch to a lesser extent), where adjectival forms are lowercase, but nominal forms of nationalities are capitalised, we include the lowercase forms in our corpus when linked, and remove sentences where they appear unlinked.

5.6. Adjusting link boundaries

We unlink certain strings when found at the end of link text: parenthesised expressions; text following a comma for LOC, ORG and PER; possessive 's in English; or other punctuation. For example, [LOC Sydney, Australia] is adjusted to [LOC Sydney], Australia, and may become [LOC Sydney], [LOC Australia] after link inference to match CoNLL and BBN annotations.

5.7. Miscellaneous changes

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

Removing rare cases. Personal title abbreviations (e.g. Mr.) are rare in Wikipedia compared to newswire text, so their appearance in entity names can lead to frequent tagging errors. In order to ensure against learning and overgenerating these rare entity cases, we explicitly remove English sentences containing title abbreviations appearing in non-PER entities such as movie titles. We also exclude personal names containing of, which are much more common in English Wikipedia's historical content than in newswire.

[16] While this requires highly language-dependent resources, it is entirely reasonable to consider another entity scheme in which Americans is marked LOC, as our default automatic annotation approach would label it. We require this more language-intensive approach only because we need to match an existing scheme.

Our initial Wikipedia models evaluated on German CoNLL generated long, spurious MISC entities containing punctuation. We therefore exclude MISC entities containing quotation marks and other punctuation from German.

Truncated conjunctions. For German and Dutch prefix coordinations like [LOC Under-] und [LOC Ober-Lais], we ensure the first prefix is tagged identically to the coordinated term, which is more likely to have an inferred link.

Fixing tokenisation. While Penn Treebank and CoNLL tokenisation consider hyphenated terms (e.g. Sydney-based) as single tokens, it is rare to infer links to hyphenated terms. We therefore split hyphenated terms into separate tokens before link inference in English, and rejoin them prior to training a model, excluding the sentence if the constituent entity types differ.

This approach does not readily apply to languages like German, where hyphenation represents compound nouns like Anne-Frank-Schule. For Russian, we treat the hyphen as a separate token since the evaluation data contains a number of hyphens between entities of different types.

6. Evaluation

To evaluate our automatically-annotated corpora, we train the C&C tagger[17] (a) with Wikipedia data; (b) with hand-annotated training data; and (c) with both combined, comparing the tagging results of each NER model on gold-standard test data.

We apply out-of-the-box the C&C Maximum Entropy NER tagger with default orthographic, contextual, in-document and personal name gazetteer features (Curran and Clark, 2003) in English. In other languages, we replaced this gazetteer with names extracted from Wikipedia articles listing of common names in various languages, and also supplemented this with first and last names from Wikipedia articles of that language that our classification model identified as PER more than 100 times. C&C optimises the Maximum Entropy task using Generalised Iterative Scaling over 200 iterations, with smoothing parameter $\sigma = \sqrt{2}$.

The text is tagged using the Penn Treebank-trained C&C POS tagger for English, and TreeTagger (Schmid, 1994) with default parameters for German,

[17] `http://schwa.org/candc`

Language	Corpus	Text	Number of tokens			MISC
			TRAIN	DEV	TEST	
English	CONLL-03	Reuters 1996	203 621	51 362	46 435	Yes
	BBN	WSJ 1998	901 894	142 218	129 654	Yes
	WIKIGOLD	Wikipedia 2008			39 007	Yes
German	CONLL-03	Frankf. Rundschau '92	206 931	51 444	51 943	Yes
	EUROPARL	EuroParl 1996		89 708	20 697	Yes
Dutch	CONLL-02	De Morgen 2000	199 069	36 908	67 473	Yes
Spanish	CONLL-02	EFE 2000	264 715	52 923	51 533	Yes
Russian	ABH	Various news	459 125	58 751	58,250	No

Table 12: Gold standard entity-annotated corpora. The MISC column indicates whether the corpus annotates all entities marked in CONLL as MISC.

Dutch and Spanish, and "small tagset"[18] parameters for Russian. NER tags are universally represented in IOB1 as used in CoNLL-03 corpora.

Our experiments use 3.5 million tokens of Wikipedia-derived training data.[19] Although much greater quantities are available, we are limited by the time and memory required to train a model. To conserve space, we only report results in languages where gold-standard corpora are available.

6.1. Evaluation corpora

Our primary evaluation uses CoNLL 2002-3 shared task NER annotations on English, German, Dutch and Spanish news text. In addition, we use other newswire corpora available for purchase—English BBN from the LDC and Russian from Appen Butler Hill (ABH)[20]; a European Parliament transcript (Faruqui and Padó, 2010); and a collection of Wikipedia pages with gold-standard NE annotations. Where standard train (TRAIN), development (DEV) and final evaluation (TEST) divisions are not provided, we have split the corpora, resulting in the sizes shown in Table 12.

We map BBN and ABH annotations to CoNLL entity types (PER, LOC, ORG, MISC). There are many stylistic and genre differences between the source texts and their annotation. For example, the English CoNLL corpus

[18]The default Russian TreeTagger tagset has 717 entries, including detailed morphology. Since C&C only uses POS as a discrete feature, coarse tags are more appropriate.

[19]Nothman (2008) reported performance over training corpora from 0.25 to 6.5 million tokens (section 7.2), finding it plateaued at around 3-4 million tokens.

[20]`http://www.appenbutlerhill.com/`

formats headlines in all-caps, and includes non-sentential data such as tables of sports scores. We now describe each corpus and its preprocessing.

The CoNLL NER shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) evaluated machine learning approaches to multilingual NER, on Spanish and Dutch (2002) and German and English (2003). The entity types are common but each language has an idiosyncratic annotation and genre. For instance, Spanish marks no lowercase adjectival nationalities and includes 192 instances where surrounding quotes are included in the entity annotation; Dutch annotates as PER the initials of photographers; and English has lots of financial and sports data in tables.

The BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005) annotates the entire Penn Treebank corpus with 105 fine-grained tags (Brunstein, 2002): 54 corresponding to CoNLL entities; 21 for numerical and time data; and 30 common noun types. We map BBN tags to CoNLL equivalents, removing extra tags.[21] We use sections 03–21 for TRAIN, 00–02 for DEV and 22-24 for TEST.

Appen Butler Hill (ABH) has produced approximately 500,000 tokens of NE-annotated data in a number of languages. We use their Russian (RUS_NER001) corpus, consisting of news reports from various sources. ABH's NE annotations divide CoNLL's LOC into locations, geopolitical entities and facilities, mark CoNLLMISC entities only when they are nationalities or religions (though religious organisations may be ORG in CoNLL), and mark titles and quantities left unannotated in CoNLL. Because not all MISC entities are marked, we evaluate on ABH without MISC. Of the 1,869 articles in the Russian ABH corpus, we used 0001–1489 as TRAIN, 1490–1679 as DEV, and 1680–1869 as TEST. We apply the sentence boundary detector and tokeniser used on our Russian Wikipedia data.

Faruqui and Padó (EUROPARL, 2010) present an out-of-domain evaluation for a CoNLL-trained German NER system on the first two German Europarl session transcripts, with CoNLL-style annotation. We used the larger transcript as DEV and the other as final TEST.

For an in-domain evaluation, we use our CoNLL-style Wikipedia corpus (WIKIGOLD, Balasuriya et al., 2009). 149 articles from a 2008 snapshot of English Wikipedia were annotated by three annotators, achieving a Fleiss'

[21] We map: LOC := FAC ∪ GPE ∪ LOCATION; ORG := ORGANIZATION; PER := PERSON; MISC := EVENT ∪ LANGUAGE ∪ LAW ∪ NORP ∪ PRODUCT ∪ WORK_OF_ART.

Kappa of 0.83 on NE tokens only, and 0.92 overall. We ensure that none of our English Wikipedia training corpora use the articles included in WIKIGOLD.

For each evaluation language we also hold out just over 100,000 tokens of silver-standard Wikipedia annotations, derived using the same method as our training data, but from different Wikipedia articles. We use this to evaluate how predictable our Wikipedia-derived data is in comparison to gold-standard corpora.

6.2. Evaluating NER *performance*

Establishing a sensible evaluation metric for NER is challenging (Nadeau and Sekine, 2007). Both the span and type of an entity may be mismatched, and the severity of an error depends on the specific instance being evaluated.

MUC (Chinchor, 1998b) awards equal score for matching an entity's *type* when at least one boundary is correct, and *text*, where an entity's boundaries are matched correctly, irrespective of classification.[22] This equal weighting is unrealistic, as some boundary errors are highly significant, while others are arbitrary (for example, the inclusion of punctuation, Mr. or the).

CoNLL only awards EXACT phrasal matches—requiring correct type and text—providing a lower-bound measure of NER performance. Manning (2006) argues that this style of evaluation favours systems that leave entities with ambiguous boundaries untagged, since boundary errors incur false positives and false negatives simultaneously.

We present our results using micro-averaged F-score for both metrics, for comparability to the CoNLL Shared Task literature and the MucEval results reported by Richman and Schone (2008).

6.3. Statistical significance of results

For each pair of systems (trained on different corpora), we consider the null hypothesis that their F-scores differ only by chance. We apply approximate randomisation (Noreen, 1989), in which we randomly keep or swap the

[22]The MUC scorer maximises F-score over the possible mappings between gold and predicted entity mentions. Each of $\{type, text\}$ for mapped mention pairs is marked as *correct* (C) or *incorrect* (I). Those which cannot be mapped are considered *spurious* (S) predictions, or gold mentions *missing* (M) in prediction. Then $P := \frac{|C|}{|C|+|I|+|S|}$ and $R := \frac{|C|}{|C|+|I|+|M|}$ (Chinchor, 1998a). Note that when evaluating performance for a particular entity type t, $I :=$(gold t mentions mapped to predictions $\neg t$). As a result, per-type precision fails to account for false positives with the same span as a gold mention.

Language	Corpus	Million tokens		Thousand sentences		Articles
		Initial	Selected	Initial	Selected	Initial
EN	WIKI-0	66.4	3.5	2 561	150	35 735
	WIKI-3	14.1	3.5	540	142	6 452
DE	WIKI-0	156.5	3.5	8 426	252	777 797
	WIKI-3	39.6	3.5	2 087	237	146 425
ES	WIKI-0	58.1	3.5	2 048	137	67 980
	WIKI-3	18.0	3.5	631	128	15 585
NL	WIKI-0	50.7	3.5	2 666	212	194 024
	WIKI-3	14.7	3.5	745	181	25 984
RU	WIKI-0	69.2	3.5	3 569	222	191 251
	WIKI-3	56.0	3.5	2 851	211	139 941
FR	WIKI-0	47.1	3.5	1 756	141	56 212
	WIKI-3	15.5	3.5	572	134	15 277
IT	WIKI-0	57.0	3.5	1 967	134	142 485
	WIKI-3	18.1	3.5	612	128	22 457
PL	WIKI-0	52.8	3.5	3 004	235	282 839
	WIKI-3	19.9	3.5	1 101	203	50 617
PT	WIKI-0	69.5	3.5	2 766	160	255 776
	WIKI-3	20.9	3.5	803	143	32 325

Table 13: The sizes of some of our Wikipedia-derived silver-standard corpora, with the quantities of initial Wikipedia data used to produce them. We compare corpora above the double line to gold-standard NE annotations.

outputs of the two systems for each sentence, and reevaluate the difference between the resulting pseudo-systems' EXACT overall F-scores. If the difference between F-scores is greater than the original in less than 50 of 9,999 such trials (i.e. $p \leq 0.005$), we reject the null hypothesis and consider the results significantly different.

7. Results

7.1. Wikipedia-derived training corpora

For each evaluation language we classify all articles with an UBER model in each language trained on POP+RAND, and set a hand-picked threshold of 0.5 on `liblinear`'s confidence, below which we consider an article's classification

unknown (UNK). This threshold gives reasonable coverage in all entity types, but exploits Wikipedia's redundancy by discarding entities with doubtful classifications. In English we retain classifications (as NE types or NON) for 96% of all articles after applying this threshold.

We produce five training corpora, each of around 3.5M tokens, for each target language:

- WIKI-BASE applies NE types to links, and uses our basic utility and confidence criteria, loosened to allow capitalised sentence starters and German common nouns.
- WIKI-0 applies all enhancements, but performs no link inference.
- WIKI-1 adds link inference with title and redirect aliases.
- WIKI-2 adds link inference with disambiguation aliases.
- WIKI-3 adds link inference with link text aliases.

All sentences passing our criteria are included, in the order of the Wikipedia snapshot, until the target 3.5M tokens is exceeded. Although each corpus' size in tokens is similar, the quantity of tokens or sentences discarded and the total number of articles processed varies greatly (Table 13). Link inference reduces the initial data required to produce a 3.5M token corpus by nearly 5 times in English. Sentences where not all proper names are labelled at lower inference levels may be included at higher levels, resulting in longer sentences on average. However, link inference has little impact in Russian.

Table 14 lists the top three entity mentions per type in each WIKI-2 corpus. The most frequent entities in each corpus are locations and nationalities, reflecting their regular appearance on Wikipedia's most frequent types of article, person and location. In all languages but Italian, the equivalent of World War II is among the 20 most frequent entity texts. German's top entity mentions exhibit some high-profile classification errors, such as [ORG DDR] which should be LOC, and non-entity [PER griechischen Mythologie]; many models classify Soviet Union as an ORG, while CONLL considers it a LOC.

7.2. Selecting an English Wikipedia model

Our English DEV results in Table 15 indicate the effectiveness of link inference, which raises F-score significantly, by up to 4%,[23] and that our

[23]Reported differences in F-score are absolute.

Language	LOC	ORG	PER	MISC
EN	U.S.	EU	Henry	American
	United States	European Union	Alexander	World War II
	Germany	NFL	Jesus	French
DE	Berlin	SPD	Johannes Paul II.	deutscher
	Deutschland	DDR	griechischen Mythologie	US-amerikanischer
	München	CDU	Paul VI.	Schweizer
ES	España	Unión Europea	Hitler	Segunda Guerra Mundial
	Estados Unidos	Unión Soviética	Zeus	Internet
	Francia	Microsoft	Jesús	Primera Guerra Mundial
NL	Nederland	Sovjet-Unie	Hitler	Nederlandse
	België	Europese Unie	Jezus	Duitse
	Duitsland	PvdA	Napoleon	Tweede Wereldoorlog
RU	США	СССР	Петра I	Великой Отечественной войны
	России	Microsoft	Пётр I	Второй мировой войны
	Москве	IBM	Гальдер Франц	Первой мировой войны
FR	France	UMP	Platon	Seconde Guerre mondiale
	États-Unis	URSS	Nietzsche	Internet
	Paris	Microsoft	Jung	Première Guerre mondiale
IT	Italia	Apple	Dante	Internet
	Roma	Unione Sovietica	Hitler	Seconda guerra mondiale
	Stati Uniti	Formula 1	Napoleone	Linux
PL	Włoszech	ZSRR	J. R. R. Tolkiena	II wojny światowej
	Francji	PRL	Hitler	I wojny światowej
	USA	PZPR	Peter Jackson	II wojnie światowej
PT	Brasil	União Soviética	Aníbal	Segunda Guerra Mundial
	Estados Unidos	URSS	Hitler	Internet
	França	União Europeia	Jesus	Primeira Guerra Mundial

Table 14: The three most frequent entity mentions for each type, for WIKI-2 in each language.

Train \ Test	Exact		MucEval	
	CONLL	BBN	CONLL	BBN
CONLL	*89.6*	69.4	*93.1*	79.9
BBN	65.0	*88.6*	75.4	*92.3*
WIKI-BASE	55.2	50.2	68.7	67.1
WIKI (WIKI-0)	64.2	69.1	75.3	79.9
+ page & redirect titles (WIKI-1)	67.3	71.7	77.7	81.9
+ DAB page titles (WIKI-2)	**67.9**	71.6	77.9	81.9
+ link text (WIKI-3)	67.6	**71.9**	**78.2**	**82.1**

Table 15: English DEV results with Wikipedia and gold-standard training corpora.

Train \ Test	Exact			MucEval		
	CONLL	BBN	WIKIGOLD	CONLL	BBN	WIKIGOLD
CONLL	*85.2*	68.3	55.2	*89.9*	78.7	68.6
BBN	**61.3**	*89.1*	56.7	72.0	*92.4*	70.6
WIKI-2	**61.3**	**69.5**	**66.6**	**73.0**	**80.5**	**78.1**

Table 16: English TEST results with our best Wikipedia model.

other refinements provide a substantial 9% increase over WIKI-BASE. At the baseline, Exact and MucEval performance differs by 13-17%, while our enhancements reduce this gap to around 10%, suggesting that many baseline errors relate to incorrect entity boundaries, or incorrect entity types where boundaries are correctly identified. Results over the three levels of link inference are insignificantly different, whether testing on CoNLL or BBN; we select WIKI-2 for final English testing.

7.3. Comparing English Wikipedia to gold-standard training

Our DEV (Table 15) and TEST (Table 16) results confirm that none of our English Wikipedia models approach the NER performance of a CoNLL-trained model evaluated on CoNLL, or BBN on BBN. We italicise such intra-corpus results in our tables and—where appropriate and not captioned otherwise—mark the highest inter-corpus (non-italic) performance in bold.

The 19–25% mismatch between training and evaluation data suggests that the training corpus is an important performance factor, cf. Ciaramita and Altun (2005). However, our final model performs as well as BBN training

Train \ Language	Exact				MucEval			
	EN	DE	ES	NL	EN	DE	ES	NL
CONLL	*89.6*	*63.6*	*77.6*	*76.5*	*93.1*	*71.0*	*85.7*	*85.3*
WIKI-BASE	55.2	53.5	54.5	55.9	68.7	64.1	71.8	71.7
WIKI-0	64.2	56.7	56.1	61.4	75.3	67.2	72.6	73.4
WIKI-2	67.9	60.9	60.7	62.2	77.9	70.9	75.2	75.9
WIKI-2 + CONLL	87.9	**67.7**	73.0	72.7	92.2	**75.9**	83.5	83.4

Table 17: Results on CONLL 2002–3 DEV corpora when training on the corresponding CONLLTRAIN data, two Wikipedia derived models, and both together. Results in bold exceed the respective CONLL on CONLL performance.

when tested on CoNLL, and as well as a CoNLL model tested on BBN.[24] A key result of our work is that the performance of non-corresponding hand-annotated corpora is often exceeded by Wikipedia-trained models.[25]

This is similarly apparent when evaluating on Wikipedia text (WIKIGOLD), where our Wikipedia-trained model significantly outperforms gold-standard training by 10–12% Exact F-score.[26] Although this is much smaller than the 23% difference when testing on CoNLL, it emphasises that automatically-derived training data can produce top results given an appropriately-matched evaluation corpus.

To account for the overall low performance on this corpus, Balasuriya et al. (2009) suggest that Wikipedia is a difficult evaluation target for NER, containing a wider variety of entity types, with longer names and less cues for their identification than traditional newswire corpora. Our result may also be compared to Mika et al.'s (2008) training data which under-performed a CoNLL-trained model on Wikipedia text. Overall, these results demonstrate that, ignoring idiosyncratic annotation variations, our English Wikipedia-trained models perform very well.

Train \ Language	Exact				MucEval			
	EN	DE	ES	NL	EN	DE	ES	NL
CONLL	*85.2*	*66.5*	*79.6*	*78.6*	*89.9*	*72.8*	*87.7*	*85.9*
WIKI-2	61.3	55.8	61.0	64.0	73.0	66.9	75.8	76.8

Table 18: Results on CONLL 2002–3 TEST corpora when training on WIKI-2.

7.4. Multilingual evaluation and joint training

Table 17 shows DEV results on all CoNLL 2002–3 corpora. We generalise to show results with WIKI-2, despite WIKI-3 performing significantly better (1.7%) on ESDEV. While CoNLL-trained and WIKI-2-trained models differ by 22% F-score in English, the equivalent margins in DE, ES and NL are markedly smaller (3–17%). This may be partly reflecting the lower performance of CoNLL-trained systems on these languages, suggesting their annotations are less predictable; it is also a reflection of the C&C NER system being primarily tuned for English performance. Nonetheless, WIKI-2 results are not far from CoNLL-trained results, and are significant improvements over WIKI-BASE,[27] though none as much as English, which received the greatest attention when tuning the automatic annotation process to the evaluation corpus.

We also experiment with training on a combined corpus consisting of CoNLLTRAIN and WIKI-2 in each language (Table 17). Apart from German, where performance increases 4.1%, this extra data degrades the CoNLL model's performance. Final CoNLLTEST results in Table 18 show large drops in performance from EN and DEDEV results when training on WIKI-2, with smaller increases in ES and NL.

Tables 19 and 20 show German DEV and TEST results on CoNLL and EUROPARL. In the DEV results, we find a CoNLL-trained model performs almost as well on EUROPARL as it does on CoNLL, and Wikipedia therefore does not outperform CoNLL when evaluating on EUROPARL. However, CoNLL per-

[24]Our WIKI-2 result on BBNTEST is almost significantly better ($0.005 < p \leq 0.01$) than CONLL. Our DEVWIKI-0 results differ from these inter-corpus results only by chance, while WIKI-2 is significantly better.

[25]Since we only have multiple gold-standard corpora in English and German, we cannot yet validate this claim for other languages.

[26]BBN and CONLL F-scores on WIKIGOLD differ by chance.

[27]WIKI-0 performs significantly better than WIKI-BASE, and link inference improves significantly on this result except in Dutch (NL) where link inference results differ by chance.

Train \ Test	Exact		MucEval	
	CONLL	EUROPARL	CONLL	EUROPARL
CONLL	*63.6*	**61.2**	*71.0*	**66.0**
WIKI-BASE	53.5	40.0	64.1	45.1
WIKI-0	56.7	49.0	67.2	56.1
WIKI-1	59.0	53.4	69.6	61.0
WIKI-2	60.9	**55.2**	70.9	**61.8**
WIKI-3	**61.6**	51.7	**71.8**	59.0

Table 19: German DEV results with Wikipedia and gold-standard training corpora. The best inter-corpus and Wikipedia-trained results are marked in bold.

Train \ Test	Exact		MucEval	
	CONLL	EUROPARL	CONLL	EUROPARL
CONLL	*66.5*	**49**	*72.8*	56
WIKI-3	56.6	48	67.8	**59**

Table 20: German TEST results with our best Wikipedia model.

forms 12% worse on the EUROPARLTEST data than DEV; this result differs from WIKI-3 performance only by chance, and WIKI-3 outperforms CONLL when using the MucEval metric. Since the EUROPARL data consists of only two parliamentary transcripts, one for DEV and one TEST, it is unsurprising that their differing subject matter may cause vastly different results. Parliamentary transcripts are also a more formalised genre than news, and contain a high frequency of honorific terms like Herr that are infrequent in Wikipedia. Nonetheless, the similar TEST performance of CONLL and WIKI-3 models on EUROPARL again illustrates Wikipedia's effective use as a versatile and cheap source of NER training data.

Our Russian results (DEV Table 21; TEST Table 22) are unusual in that our baseline system performance differs insignificantly from WIKI-0 (which attempts to identify adjectival entities and nationalities), which is significantly better than models with link inference. This may reflect the fact that we spent very little time adapting Russian Wikipedia to the ABH annotation schema, and that we do not evaluate on the challenging MISC entity type. It may also stem from the difficulty of applying our simple string-based matching approaches in link inference to the complex morphology of Russian. However, our Wikipedia models again perform at a similar margin from the same-corpus result (13–14% Exact) to what we find in other languages.

Train \ Test	EXACT ABH	MUCEVAL ABH
ABH	*78.7*	*83.8*
WIKI-BASE	65.3	73.9
WIKI-0	**65.8**	**74.1**
WIKI-1	64.9	73.3
WIKI-2	64.8	73.2
WIKI-3	64.9	73.2

Table 21: Russian DEV results with Wikipedia and gold-standard training.

Train \ Test	EXACT ABH	MUCEVAL ABH
ABH	*79.8*	*85.0*
WIKI-BASE	65.5	74.1

Table 22: Russian TEST results with our best Wikipedia model.

Train/test	EXACT					MUCEVAL				
	EN	DE	ES	NL	RU	EN	DE	ES	NL	RU
gold standard	***85.2***	*66.5*	*79.6*	*78.6*	*79.8*	***89.9***	*72.8*	*87.7*	*85.9*	*85.0*
WIKI-2	*82.4*	***90.5***	***83.5***	***89.7***	***82.4***	*89.2*	***93.3***	***90.1***	***93.8***	***88.4***

Table 23: Performance given training and evaluation corpora produced by the same process. Row 1: gold-standard CONLL and RUABH corpora. Row 2: silver-standard WIKI-2. The highest result in each column is marked in bold.

7.5. Self-similar evaluation

We also assess the reliability of our WIKI-2 corpora by evaluating them on automatic annotations of other Wikipedia articles produced using the same process. Table 23 compares these results to the self-similar results we have already presented on gold-standard TEST corpora (i.e. CONLL on CONLL, or ABH on ABH). Although our corpus selection process may automatically remove many difficult cases, we see that the resulting annotations are predictable (or learnable) to an extent roughly equivalent to those in manually-annotated corpora.

The main replicable evaluation in Richman and Schone (2008) also uses

NE type	Spanish		Russian	
	R&S 2008	This work	R&S 2008	This work
All	84.6	90.1	80.2	87.5
ORG	70.1	73.4	71.2	81.3
PER	82.1	92.8	75.1	92.7

Table 24: Comparing our self-similar MUCEVAL results to Richman and Schone (2008).

NE type	CONLL			WIKIGOLD			WIKI-2-similar		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
LOC	63.8	70.6	67.0	70.8	80.9	75.5	*80.8*	*87.5*	*84.0*
ORG	65.0	36.2	46.5	63.0	48.0	54.0	*78.2*	*71.0*	*74.4*
PER	87.6	77.9	82.5	80.0	84.0	82.0	*90.7*	*90.7*	*90.7*
MISC	29.0	54.0	38.0	43.0	58.0	49.0	*73.6*	*72.6*	*73.1*
All	62.1	60.5	61.3	64.6	68.7	66.6	*82.0*	*82.7*	*82.4*

Table 25: ENWIKI-2 performance (EXACT metric) broken down by named entity type.

self-similar Wikipedia testing,[28] so we present comparative results in Table 24.[29] Our overall self-similar results are 5–8% higher, but consider different entity types. Considering only ORG and PER, for which Richman and Schone gave results, our self-similar evaluation consistently outperforms Richman and Schone (2008), by up to 17.6% on RUPER.

7.6. Entity type performance

We present our English results on recognising each entity type in Table 25. Regardless of test corpus, our best performance is on PER, followed by LOC, with much lower performance on the diverse ORG and MISC types, corresponding with our article classification results (see Table 9). The written form of ORG and MISC entity names is generally much less regular than PER and LOC; using a finer-grained type scheme might provide lower entropy over forms. However, we find low MISC precision and low ORG recall, suggesting that many organisations are incorrectly identified as MISC, which is the second-highest form of per-token error on CONLL, behind marking non-entity tokens as MISC. This ORG-MISC confusion may also relate to a nested NE approach describing MISC entities within ORGs (e.g. [ORG [MISC Australian] Mutual Provident Society]) or ORG entities within MISC (e.g. [MISC [ORG Apple] iPod]), as well as metonymy in entities like New York Times as an organisation or publication, or a band and its self-titled album.

[28] The gold-standard corpora used by Richman and Schone are not publicly available.

[29] This comparison is very rough, since every component differs between our experiments, including: entity types (Richman and Schone use ACE types: PERSON, GPE, ORGANIZATION, VEHICLE, WEAPON, LOCATION, FACILITY, DATE, TIME, MONEY & PERCENT); articles used in training and portion selection; maturity of Wikipedia; and machine learner (they use a modified version of BBN's IdentiFinder (Bikel et al., 1999)).

8. Discussion and future work

Our results clearly demonstrate the use of Wikipedia to derive high-performance NE-annotated data in many languages, and while we only present evaluations on languages with existing NER corpora, our results suggest their application to the many resource-scarce languages covered by Wikipedia.

Our initial approach (Nothman et al., 2008) focused on English Wikipedia and was optimised through extensive analysis and comparison between our Wikipedia-derived corpora and the target gold standards (Nothman et al., 2009). We have since presented state-of-the-art approaches to labelling and classifying Wikipedia's articles (Tardif et al., 2009), transferring this knowledge into German (Ringland et al., 2009), and evaluating our English Wikipedia-derived corpus on manually-annotated Wikipedia data (Balasuriya et al., 2009), as reviewed in this paper. However, we had not yet taken a more broadly multilingual approach to article classification or the derivation of training data to test the robustness of our approach across languages and gold standards.

The present work succeeds in overcoming differences in capitalisation conventions between languages such as English and German, and also identifies that non-English Wikipedias have sufficient structural and textual information to create usable training data. However, the sorts of extensive analysis we used in English to match our Wikipedia corpora more closely to gold-standard targets (Nothman et al., 2009) are outside the scope of this paper, and hence we by no means consider our non-English performance as the method's upper-bound.

It is also apparent that the distribution of entity mentions in Wikipedia (see Table 14) does not match newswire corpora or general-domain text, and we plan to investigate more robust text selection techniques to reduce the discrepancy between Wikipedia and target domains' entity distributions.

Similarly, an ideal general training corpus should be widely varied in topic and language, but our current process only considers 0.2% of English Wikipedia articles in creating WIKI-3, suggesting topical coverage is low. We are also concerned about the utility of including almost-identical sentences from automatically-generated pages in Wikipedia (usually derived from location gazetteers), which may make up a large proportion of Wikipedia languages with few contributors. In future work, we intend to explore methods for redefining a sentence or article's utility, measuring how much information it would add to an existing corpus, and utilising measures of Wikipedia article

quality (e.g. Hu et al., 2007).

While we harnesses Wikipedia's breadth of language, tagging only the four CoNLL NE types ignores Wikipedia's diverse coverage of technical and popular domains; MISC performance remains low, even when testing on gold Wikipedia annotations. We are yet to evaluate corpora produced with our medium or fine-grained classifications, or to take advantage of our ability to re-target these fine-grained classifications by mapping them to another schema. Further, by using domain-oriented article classifications and sentence selection, we foresee this method being used for rapid construction of entity-annotated corpora in particular domains.

Our work also highlights the brittleness of NER evaluation. CoNLL does not provide annotation guidelines, and various inconsistencies appear both within a corpus and between the various CoNLL corpora in different languages. For example, the adjectival forms of entities such as nationalities and religions are usually annotated in the English CoNLL data, mostly in German and Dutch, and vary rarely in Spanish. This makes achieving good results without substantial corpus-specific tuning impossible. CoNLL-style annotations are directed at token-tagging approaches suited to machine learning, and hence label each token with at most a single entity. Nested entity mentions, such as [ORG [LOC New York] Stock Exchange], cannot be described in CoNLL and add evaluation ambiguity.

This points us towards the benefits of an extrinsic evaluation, using an NER application such as Question Answering, where the severity of errors can be more meaningfully evaluated, especially since our work shows that a Wikipedia-derived model is likely to excel over a news-trained model when extracting entity-related information from diverse sources.

9. Conclusion

We have demonstrated a method of automatically producing named entity-annotated text in a number of languages from Wikipedia, based on labelling each outgoing link with the entity type of the target article. Our results demonstrate this approach will be highly effective and efficient for creating NER models in resource-scarce languages. It even performs comparably to existing gold-standard corpora when idiosyncratic annotation scheme variations are ignored.

Our method initially requires classification of all Wikipedia articles into NE types. We present a multilingual state-of-the-art supervised classification

approach—achieving up to 94.9% on coarse and 89.9% on fine-grained entity types—and compare it to other approaches from the literature. In order to model and evaluate classification, we have labelled 4,800 English, 870 German and 1,500 other-language Wikipedia articles with fine-grained NE types. We demonstrate the combination of popular and randomly selected articles as ideal for training such a classification approach.

Using publicly available CoNLL 2002-3 shared task test data and other corpora, we have evaluated the performance of NER models trained with 3.5 million tokens of Wikipedia-derived annotations in each of English, German, Spanish, Dutch and Russian. Our Wikipedia models do not perform as well on traditional NER evaluation data as models trained on corresponding traditional training data, which is unsurprising given the domain mismatch.

However, we have found that in English and German, Wikipedia-derived NER models perform as well or better than gold models on inter-corpus evaluations, such that Wikipedia is better training data for CoNLL text than the BBN corpus, and is as good as CoNLL for BBN. Further, our silver-standard annotations outperform traditional training on a manually-annotated collection of Wikipedia articles (Balasuriya et al., 2009) by 10-12% F-score. Together these suggest that a Wikipedia model may be better for NER in some domains than existing gold standards, but also generally applicable where training data is not available to match a particular target.

In other languages, our results are generally consistent with these conclusions, with Wikipedia models closer in performance to gold models (12-19%) than in English (24%) when comparing Wikipedia results on gold-standard test corpora to models built from corresponding training data.

We also evaluate performance on annotated copora produced by the same automated method as our training data, with strong results across all languages (see section 7.5) suggesting that the automatic annotations are learnable to a similar extent to gold-standard data. We have shown better performance on Wikipedia text than Mika et al. (2008) (see section 7.3), and arguably better performance on automatically-annotated test data than Richman and Schone (2008).

Within the Wikipedia processing literature, this task of generating NE-annotated corpora is arguably the most intensive use of Wikipedia's structured features together with its sentential text. We use Wikipedia's category graph, infoboxes and bag-of-words content in article classification; article body text and outgoing links in deriving training data; incoming link texts, redirects and information from disambiguation pages as aliases for inferring

additional outgoing links; and inter-language links to transfer knowledge between languages. Nonetheless, there are other Wikipedia features we do not utilise: citations, revision history, extra-sentential structure, text styling, etc.

Our work illustrates the wealth of linguistic and world knowledge freely available in Wikipedia's structured and unstructured content. We exploit this knowledge to derive enormous and accurate NE-annotated corpora for a variety of domains and languages.

Acknowledgements

We would like to thank members of Schwa Lab and the anonymous reviewers for their helpful feedback on all of the research described here. This work has been supported under the Computable News project at the Capital Markets CRC. Nothman was supported by a University of Sydney Honours Scholarship and a Vice-Chancellor's Research Scholarship. Ringland was supported by a University of Sydney Postgraduate Award. Radford was supported by an Australian Postgraduate Award. Nothman and Radford were also supported by a Capital Markets CRC PhD top-up scholarship. Curran and Murphy were funded under Australian Research Council Discovery grants DP0665973 and DP1097291.

Appendix A. Key phrases used in category keyword classification approach

The following is the complete set of 141 case-sensitive keywords/phrases matched against Wikipedia category titles for the classification approach described in section 4.2.1.

LOC	asteroids; Asteroids; asteroid stubs; cities; Cities; Counties; Countries; geography stubs; infobox lake; Infobox Settlement; lakes; Lakes of; mountains; Mountains; municipalities; Municipalities; Populated places; Regions; Republics; Rivers of; Settlements; States; Suburbs of; Territories; towns; Towns; Unincorporated communities ; villages; Villages; Water bodies
ORG	Advocacy groups; Agencies; booksellers; bookstores; Businesses; Clubs; Club stub; Colleges; companies; Companies; Company stub; Corporations; Legislatures; Media by; musical groups; music groups; Newspapers; Organizations; Political parties; record labels; Record labels; software companies; Teams; Team stub; Unions; Universities; University stub
PER	academics; Actors needing; actors; alumni; Alumni; Biography stub; births; by occupation; Characters; composers; deaths; Fellows of; football defedners; footballers; football forwards; Free software programmers; Given names ; guitarists; human names; living people; Living people; musicians; painters; Participants; People by; People from; People in; personnel; pitchers; players; poets; producers; singers; Surname; Year of birth; Year of death
MISC	albums; album stubs; books; bowl games; discographies; facilities; films; film stubs; games; hAudio; Houses on; journals; magazines; Magazines; novels; Operas; plot summary; racehorse; Racehorse; Ship infoboxes; Ships; Singlechart; singles; Sonnets; Stations of; television series; Television series
NON	about singers; Centuries; -Class; Days; features; Gaelic games; History of; Incidents; List of; Lists; Months; navigational boxes; Screenshots; sports and games; Wars; Wikipedia; WikiProject; Years
DAB	disambiguation; Disambiguation; List

References

Adafre, S.F., de Rijke, M., 2006. Finding similar sentences across multiple languages in Wikipedia, in: Proceedings of the Workshop on NEW TEXT, 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 62–69.

Adar, E., Skinner, M., Weld, D.S., 2009. Information arbitrage across multi-lingual Wikipedia, in: Proceedings of Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain. pp. 94–103.

An, J., Lee, S., Lee, G.G., 2003. Automatic acquisition of named entity tagged corpus from world wide web, in: The Companion Volume to the Proceedings

of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan. pp. 165–168.

Ando, R.K., Zhang, T., 2005. A high-performance semi-supervised learning method for text chunking, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan. pp. 1–9.

Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., Curran, J.R., 2009. Named entity recognition in Wikipedia, in: Proceedings of the Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Singapore. pp. 10–18.

Bhole, A., Fortuna, B., Grobelnik, M., Mladenić, D., 2007. Mining Wikipedia and relating named entities over time, in: Proceedings of the 10th International Multiconference on Information Society, Jožef Stefan Institute, Ljubljana. pp. 117–180.

Biadsy, F., Hirschberg, J., Filatova, E., 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, Columbus, Ohio. pp. 807–815.

Bikel, D.M., Schwartz, R., Weischedel, R.M., 1999. An algorithm that learns what's in a name. Machine Learning 34, 211–231.

BNC, 2007. The British National Corpus. Version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. `http://www.natcorp.ox.ac.uk/`.

Bouma, G., Duarte, S., Islam, Z., 2009. Cross-lingual alignment and completion of Wikipedia templates, in: Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, pp. 21–29.

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

Bunescu, R., Paşca, M., 2006. Using encyclopedic knowledge for named entity disambiguation, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy. pp. 9–16.

Chinchor, N., 1998a. MUC-7 test scores introduction (Appendix B), in: Proceedings of the 7th Message Understanding Conference.

Chinchor, N., 1998b. Overview of MUC-7, in: Proceedings of the 7th Message Understanding Conference.

Chinchor, N., Robinson, P., 1998. MUC-7 named entity task definition (version 3.5), in: Proceedings of the 7th Message Understanding Conference.

Ciaramita, M., Altun, Y., 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, Whistler, British Columbia.

Cucerzan, S., 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, Prague, Czech Republic. pp. 708–716.

Curran, J.R., Clark, S., 2003. Language independent NER using a maximum entropy tagger, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, Canada. pp. 164–167.

Dakka, W., Cucerzan, S., 2008. Augmenting Wikipedia with named entity tags, in: Proceedings of the 3rd International Joint Conference on Natural Language Processing, Hyderabad, India. pp. 545–552.

Dickinson, M., Meurers, W.D., 2003. Detecting errors in part-of-speech annotation, in: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary. pp. 107–114.

Ehrmann, M., Turchi, M., Steinberger, R., 2011. Building a multilingual named entity-annotated corpus using annotation projection, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria. pp. 118–124.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A., 2005. Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence 165, 91–134.

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J., 2008. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Reseach 9, 1871–1874.

Faruqui, M., Padó, S., 2010. Training and evaluating a German named entity recognizer with semantic generalization, in: Proceedings of Die Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS) 2010, Saarbrücken, Germany.

Fernandes, E., Brefeld, U., 2011. Learning from partially annotated sequences, in: Machine Learning and Knowledge Discovery in Databases. volume 6911 of *Lecture Notes in Computer Science*, pp. 407–422.

Ferrández, S., Toral, A., Óscar Ferrández, Ferrández, A., Muñoz, R., 2007. Applying Wikipedia's multilingual knowledge to cross–lingual question answering, in: Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems, Paris, France. pp. 352–363.

Filatova, E., 2009. Multilingual Wikipedia, summarization, and information trustworthiness, in: Proceedings of the SIGIR Workshop on Information Access in a Multilingual World, Boston, Massachusetts. pp. 19–24.

Fu, R., Qin, B., Liu, T., 2011. Generating chinese named entity data from a parallel corpus, in: Proceedings of 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand. pp. 264–272.

Gabrilovich, E., Markovitch, S., 2006. Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge, in: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, Massachusetts. pp. 1301–1306.

Hu, M., Lim, E.P., Sun, A., Lauw, H.W., Vuong, B.Q., 2007. Measuring article quality in Wikipedia: Models and evaluation, in: Proceedings of the 16th ACM Conference on Information and Knowledge Management, Lisbon, Portugal. pp. 243–252.

Kazama, J., Torisawa, K., 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, Prague, Czech Republic. pp. 698–707.

Kazama, J., Torisawa, K., 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, Columbus, Ohio. pp. 407–415.

Kiss, T., Strunk, J., 2006. Unsupervised multilingual sentence boundary detection. Computational Linguistics 32, 485–525.

Laws, F., Scheible, C., Schütze, H., 2011. Active learning with amazon mechanical turk, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK.. pp. 1546–1556.

Lawson, N., Eustice, K., Perkowitz, M., Yetisgen-Yildiz, M., 2010. Annotating large email datasets for named entity recognition with mechanical turk, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles. pp. 71–79.

LDC, 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*. Linguistic Data Consortium. Version 5.6.1 2005.05.23, `http://www.ldc.upenn.edu/Projects/ACE/docs/English-Entities-Guidelines_v5.6.1.pdf`.

Liao, W., Veeramachaneni, S., 2009. A simple semi-supervised algorithm for named entity recognition, in: Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing, Boulder, Colorado. pp. 58–65.

Loper, E., Bird, S., 2002. NLTK: The natural language toolkit, in: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia, Pennsylvania. pp. 63–70.

Ma, X., 2010. Toward a name entity aligned bilingual corpus, in: Proceedings of the Workshop on Methods for the Automatic Acquisition of Language Resources and Their Evaluation Methods, Valletta, Malta.

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

Marcus, M., Santorini, B., Marcinkiewicz, M., 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19, 313–330.

Medelyan, O., Legg, C., 2008. Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense, in: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy, Chicago, Illinois.

de Melo, G., Weikum, G., 2010. MENTA: Inducing multilingual taxonomies from Wikipedia, in: Proceedings of the Nineteenth ACM Conference on Information and Knowledge Management, Toronto, Canada. pp. 1099–1108.

Merchant, R., Okurowski, M.E., Chinchor, N., 1996. The multilingual entity task (MET) overview, in: TIPSTER Text Program Phase II: Proceedings of a Workshop held at Vienna, Virginia, pp. 445–448.

Mihalcea, R., Csomai, A., 2007. Wikify!: Linking documents to encyclopedic knowledge, in: Proceedings of the 16th ACM Conference on Information and Knowledge Management, pp. 233–242.

Mika, P., Ciaramita, M., Zaragoza, H., Atserias, J., 2008. Learning to tag and tagging to learn: A case study on Wikipedia. IEEE Intelligent Systems 23, 26–33.

Mikheev, A., Moens, M., Grover, C., 1999. Named entity recognition without gazetteers, in: Proceedings of the ninth conference of European chapter of the Association for Computational Linguistics, Bergen, Norway. pp. 1–8.

Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Lingvisticae Investigationes 30, 3–26.

Nadeau, D., Turney, P.D., Matwin, S., 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity, in: Proceedings of the 19th Canadian Conference on Artificial Intelligence, pp. 266–277.

Nastase, V., Strube, M., Börschinger, B., Zirn, C., Elghafari, A., 2010. WikiNet: A very large scale multi-lingual concept network, in: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta. pp. 19–21.

Navigli, R., Ponzetto, S.P., 2010. BabelNet: Building a very large multilingual semantic network, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 216–225.

Nobata, C., Collier, N., Tsuji, J., 2000. Comparison between tagged corpora for the named entity task, in: Proceedings of the Workshop on Comparing Corpora, Seattle, Washington. pp. 20–27.

Noreen, E.W., 1989. Computer Intensive Methods for Testing Hypotheses. John Wiley & Sons.

Nothman, J., 2008. Learning named entity recognition from Wikipedia. Honours Thesis. School of IT, University of Sydney.

Nothman, J., Curran, J.R., Murphy, T., 2008. Transforming Wikipedia into named entity training data, in: Proceedings of the Australian Language Technology Workshop, Hobart. pp. 124–132.

Nothman, J., Murphy, T., Curran, J.R., 2009. Analysing Wikipedia and gold-standard corpora for NER training, in: Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece. pp. 612–620.

Ponzetto, S.P., Navigli, R., 2009. Large-scale taxonomy mapping for restructuring and integrating Wikipedia, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California. pp. 2083–2088.

Potthast, M., Stein, B., Anderka, M., 2008. A Wikipedia-based multilingual retrieval model, in: Proceedings of Advances in Information Retrieval: 30th European Conference on IR Research, Glasgow, UK. pp. 522–530.

Ratinov, L., Roth, D., 2009. Design challenges and misconceptions in named entity recognition, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, Boulder, Colorado. pp. 147–155.

Richman, A.E., 2010. (personal communication).

Richman, A.E., Schone, P., 2008. Mining wiki resources for multilingual named entity recognition, in: 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, Ohio. pp. 1–9.

Ringland, N., Nothman, J., Murphy, T., Curran, J.R., 2009. Classifying articles in English and German Wikipedia, in: Proceedings of the Australasian Language Technology Association Workshop, Sydney, Australia. pp. 20–28.

Ruiz-Casado, M., Alfonseca, E., Castells, P., 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets, in: Advances in Web Intelligence. volume 3528 of *Lecture Notes in Computer Science*, pp. 380–386.

Samy, D., Moreno, A., Guirao, J.M., 2005. A proposal for an Arabic named entity tagger leveraging a parallel corpus, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, Borovets, Bulgaria. pp. 459–465.

Schmid, H., 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Technical Report. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Schönhofen, P., 2006. Identifying document topics using the Wikipedia category network, in: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong. pp. 456–462.

Schönhofen, P., Benczúr, A., Bíró, I., Csalogány, K., 2008. Cross-language retrieval with Wikipedia, in: Advances in Multilingual and Multimodal Information Retrieval. volume 5152 of *Lecture Notes in Computer Science*, pp. 72–79.

Sekine, S., Sudo, K., Nobata, C., 2002. Extended named entity hierarchy, in: Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands. pp. 1818–1824.

Shah, R., Lin, B., Gershman, A., Frederking, R., 2010. SYNERGY: A named entity recognition system for resource-scarce languages such as Swahili using online machine translation, in: Proceedings of the Second Workshop on African Language Technology, Valletta, Malta. pp. 21–26.

Sorg, P., Cimiano, P., 2008. Enriching the crosslingual link structure of Wikipedia: A classification-based approach, in: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence, Chicago, Illinois.

Strube, M., Ponzetto, S.P., 2006. WikiRelate! Computing semantic relatedness using Wikipedia, in: Proceedings of the 21st national conference on Artificial intelligence, Boston, Massachusetts. pp. 1419–1424.

Suchanek, F.M., Kasneci, G., Weikum, G., 2007. YAGO: A core of semantic knowledge — unifying WordNet and Wikipedia, in: Proceedings of the 16th international conference on World Wide Web, Banff, Alberta. pp. 697–706.

Suchanek, F.M., Kasneci, G., Weikum, G., 2008. YAGO: A large ontology from Wikipedia and WordNet. Web Semantics: Science, Services and Agents on the World Wide Web 6, 203–217.

Suzuki, J., Isozaki, H., 2008. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, Ohio. pp. 665–673.

Tardif, S., Curran, J.R., Murphy, T., 2009. Improved text categorisation for Wikipedia named entities, in: Proceedings of the Australian Language Technology Workshop, Sydney, Australia. pp. 104–108.

Tjong Kim Sang, E.F., 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition, in: Proceedings of the 6th Conference on Natural Language Learning, Taipei, Taiwan. pp. 1–4.

Tjong Kim Sang, E.F., De Meulder, F., 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the 7th Conference on Natural Language Learning, Edmonton, Canada. pp. 142–147.

Tkatchenko, M., Ulanov, A., Simanovsky, A., 2011. Classifying Wikipedia entities into fine-grained classes, in: Proceedings of the IEEE 27th International Conference on Data Engineering Workshops, pp. 212–217.

Toral, A., Ferrández, S., Monachini, M., Muñoz, R., 2011. Web 2.0, language resources and standards to automatically build a multilingual named entity lexicon. Language Resources and Evaluation , 1–37.

Toral, A., Muñoz, R., 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, Trento, Italy.

Urbansky, D., Thom, J., Schuster, D., Schill, A., 2011. Training a named entity recognizer on the web, in: Web Information System Engineering – WISE 2011. volume 6997 of *Lecture Notes in Computer Science*, pp. 87–100.

Voyer, R., Nygaard, V., Fitzgerald, W., Copperman, H., 2010. A hybrid model for annotating named entity training corpora, in: Proceedings of the Fourth Linguistic Annotation Workshop, pp. 243–246.

Watanabe, Y., Asahara, M., Matsumoto, Y., 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, Prague, Czech Republic. pp. 649–657.

Weischedel, R., Brunstein, A., 2005. BBN pronoun coreference and entity type corpus. LDC2005T33, Linguistic Data Consortium, Philadelphia.

Wentland, W., Knopp, J., Silberer, C., Hartung, M., 2008. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration, in: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco. pp. 28–30.

Wong, Y., Ng, H.T., 2007. One class per named entity: Exploiting unlabeled text for named entity recognition, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India. pp. 1763–1768.

Wu, D., Lee, W.S., Ye, N., Chieu, H.L., 2009. Domain adaptive bootstrapping for named entity recognition, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore. pp. 1523–1532.

Wu, F., Weld, D.S., 2007. Autonomously semantifying Wikipedia, in: Proceedings of the 16th ACM Conference on Information and Knowledge Management, Lisbon, Portugal. pp. 41–50.

Yarowsky, D., Ngai, G., Wicentowski, R., 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora, in: Proceedings of the First International Conference on Human Language Technology Research, San Francisco. pp. 109–116.