Fine Grained Classification of Named Entities In Wikipedia

Maksim Tkachenko, Alexander Ulanov, Andrey Simanovsky

HP Laboratories
HPL-2010-166

Keyword(s):

named entity recognition; Wikipedia; classification

Abstract:

This report describes the study on classifying Wikipedia articles into an extended set of named entity classes. We employed semi-automatic method to extend Wikipedia class annotation and created a training set for 15 named entity classes. We implemented two classifiers. A binary named-entity classifier decides between articles about named entities and other articles. A support vector machine (SVM) classifier trained on a variety of Wikipedia features determines the class of a named entity. Combination of the two classifiers helped us to boost classification quality and obtain classification quality that is better than state of the art.

External Posting Date: October 21, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: October 21, 2010 [Fulltext]

Fine Grained Classification of Named Entities In Wikipedia

Maksim Tkachenko and Alexander Ulanov and Andrey Simanovsky

Abstract

This report describes the study on classifying Wikipedia articles into an extended set of named entity classes. We employed semi-automatic method to extend Wikipedia class annotation and created a training set for 15 named entity classes. We implemented two classifiers. A binary named-entity classifier decides between articles about named entities and other articles. A support vector machine (SVM) classifier trained on a variety of Wikipedia features determines the class of a named entity. Combination of the two classifiers helped us to boost classification quality and obtain classification quality that is better than state of the art.

1 Introduction

Wikipedia is a web-based collaborative multilingual encyclopedia. Since its launch it has grown to the largest knowledge base on the Internet. Volunteers over the whole world constantly contribute content to Wikipedia. Wikipedia contains over 16 million articles in 270 languages including over 3.3 million articles in English.

Being comprised of blocks of structured, semi-structured, and unstructured information, Wikipedia is an attractive resource for researchers in the areas of information extraction and natural language processing. Nothman at al. describe how Wikipedia can be used for named entity recognition in unstructured text [12]. Bunescu applies it in a disambiguation system [4] . Adrian discovers relationships between entities with the help of Wikipedia [9].

We test our named entity classifier on Wikipedia. Named entity classification is an important part of named entity extraction, which in turn is the first stage of various text analytics techniques.

2 Related Work

Recently, several attempts were made to classify Wikipedia into named entity classes. Binary classification approaches that decide only between instance (named entity) and concept and do not distinguish named-entity classes were implemented as well.

Bunescu and Pasca [4] employed dictionary of named entities to construct entity detection and disambiguation system. Their algorithm

Class	F-score
PERSON	74.2
TIMEX/NUMEX	99.4
FACILITY	85.4
PRODUCT	75.4
LOCATION	72.4
NATURAL OBJECTS	34.2
ORGANIZATION	71.6
VOCATION	90.5
EVENT	23.4
TITLE	8.9
NAME OTHER	0
UNIT	12.5
ALL	78.6
ALL (no articles)	55.0

Table 1: Watanabe's results with best ALL F-score.

predicted named entity class based on heuristics that tested capitalization of an article title and occurrences of it within the article content. Toral at al. [18] extended WordNet with named entities automatically extracted from Wikipedia. They counted title occurrences not only in an English version of an article but also took into consideration localized versions in 10 languages. The algorithm achieved F = 82.6%. Zirn at al. [20] presented more sophisticated methods for binary classification such as applying named entity recognizer to Wikipedia category titles or filtering titles in plural form. The combination of the methods was evaluated against ResearchCyc and accuracy of 84.5% was obtained.

Toral at al. [17] proposed a method of automatically building gazetteers for named entity recognition by using Wikipedia. The approach predicted analyzed Wikipedia definition sentence using WordNet noun hierarchy to predict named entity class.

Watanabe at al. [19] considered the problem of classifying internal Wikipedia links into 12 named entity classes. They introduced a graph structure with nodes representing anchor texts and edges introduced according to the HTML markup. Conditional Random Fields classifier was trained on the model. Table 1 shows results achieved by the system. Watanabe used a dataset constructed out of a random selection of 2300 articles from the Japanese version of Wikipedia.

Bhole et al. [1] combined heuristics with linear support vector machine to classify articles. Both methods were evaluated on manually labeled sample of 1000 articles. Heuristic used parts of Wikipedia markup such as infobox tables, coordinate templates, and matching keywords. It also matched dates to identify people. Evaluation showed precision of 100%, 96%, 100%, and recall of 62%, 49%, 10% for people, places, and organizations respectively. In order to apply machine learning, a standard bag-of-words representation was used for tokens comprising article titles. A preliminary step stemmed words and filtered them according to a stop-word list.

Dakka and Cucerzan [8] explored the use of Naive Bayes (NB) and SVM classifiers for classifying Wikipedia. They experimented with bag-of-words representation of different feature groups such as article text terms, structured data on article pages, first paragraphs of Wikipedia articles, Wikipedia abstracts, disambiguated surface forms, and contexts surrounding anchor texts of internal links. Articles were annotated according to tagset similar to CoNLL. An additional common pagearticle tag (OTHER_PAGE) was introduced. SVM achieved better results than NB.

The use of structured features, such as Wikipedia templates, infoboxes, and taxoboxes was explored in [15]. SVM and NB classifiers were applied to bag-of-words representations of articles. Tokens of titles, first sentences, first paragraphs, template names, and contents of “Infobox”, “Taxobox” and “Sidebar” templates were extracted. Some features were added to a feature vector with distinct prefixes. Authors compared their system with [8] and [12] (see Table 2). SVM got better scores.

Classifier	F-score
Dakka	90
Nothman	91
Tardif	95

Table 2: Tardiff’s F-score comparison.

Iman et al. [13] used SVM approach with beta-gamma threshold adjustment. Their motivation was to test a classification system that makes use of cross-language links and features. The features that were used are stemmed content words, tokenized attribute names from infoboxes, whitespace delimited words of categories, and persondata template attributes. They demonstrated that multilingual features raise the quality of classification as compared to using features from English articles only. The threshold adjustment technique improves classification quality as well.

Nothman at al. [12] explored if Wikipedia can be tagged with a bootstrap process. Their semi-supervised algorithm classified categories and definition nouns of articles and based on that predicted the article class. They evaluated their approach on on a set of 1300 hand-labeled articles and achieved F of 89%.

Bohn and Norvag [2] measured the quality of named entities extraction if categories are used. They selected a random subset of 585 articles which categories matched patterns like "* companies", "Organization based in *", etc. The selected set was compared with manual annotation results. F-score was 98%, 97% and 99.8% for companies, organizations, and people respectively.

Table 3 gives a summary of all reviewed results. The values in the table cannot be compared directly and are intended to give a rough outline of the state of the art. There is no publicity available benchmark on which all systems can be evaluated.

Classifier	PER	ORG	LOC	MISC	COMM	DAB
Bhole [1]	72.7	41.6	70.5	-	-	-
Dakka [8]	95.1	93.4	95.4	92.4	87.8	-
Tardif (Exp-1) [15]	96	95	99	94	-	-
Tardif (Exp-2) [15]	95	93	99	89	93	98
Iman [13]	95.7	82.2	92.4	-	-	-
Nothman [12]	97	85	95	80	79	96
Toral [17]	78.3	22.2	68.1	-	-	-

Table 3: Named entity classification F-scores.

3 Named Entity Classes

The first step in classification of the articles is the choice of named entities classes. Named entities can be classified into different sets of classes depending on the application. PERSON, LOCATION, and ORGANIZATION are constantly used since MUC-6 competition. Other classes vary. MISC class is used at CONLL conferences to denote entities that fall outside MUC classes. BNN guideline for the Question Answering task [3] described a hierarchy of 29 classes of named entities. Sekine at al. [14] defined a named entity hierarchy, which includes many fine grained subclasses such as museum or river and adds a wide range of classes such as product and event. We chose 15 regular named entity classes and three auxiliary classes (DISAMBIGUATION_PAGE, LIST_OF_PAGE, OTHER_PAGE). Our classes are selected from either BNN or Sekine's hierarchies. Table 4 lists entity classes and gives examples of each.

4 Binary Classification

It is difficult to divide articles about named entities from articles about common terms. There is an issue with distinguishing named entities of unknown class. Common terms

Class	Example
PERSON	Alexander Pushkin
GPE	Paris
GEOLOGICAL_REGION	Caribbean Sea
ASTRAL_BODY	Zeta Ursae Minoris
FACILITY	National Museum of Denmark
ORGANIZATION	Waste Management Inc
AUTO	Ford Mustang
WORK_OF_ART	Toy Story 2
GAME	Warcraft II: Tides and Darkness
ANIMAL	Beaded Lizard
PLANT	Honey locust
INSECT	Army worm
PERIOD	Jurassic
LANGUAGE	Chinese language
COMPUTER	Dell Inspiron
LIST_OF_PAGE	List of monarchs of Korea
DISAMBIGUATION_PAGE	Mummy (disambiguation)
OTHER_PAGE	Walking in the United Kingdom

and named entities of unknown class are typically presented by few training examples compared to the proportion of such articles in Wikipedia. As a result, relevant features exhibit bias. Table 3 shows summary of our experiments on assessing comparable classification quality for different classes. As we can see from the table, the F-score of classification of common articles is usually lower than F-score obtained for other classes. Other experiments support the same trend.

To distinguish articles about named entities from other articles, we use a combination of heuristics that had prior evidence that they perform well. First, we distinguish disambiguation and 'list of' articles. We rely on observation that 'list of' article can be recognized by “List of” prefix in the title and that most disambiguation articles are tagged by a category from “Category:Disambiguation pages” hierarchy, have 'disambiguation' word in title, or are constructed by using 'Disambig' or 'Surname' template.

Next we analyze titles and find titles with clarifying terms in parenthesis. The clarifying terms help to distingiush named entity class. For example, titles 'Mummy (film)', 'Red Heat (1985 film)' refer to entities of the WORK_OF_ART class. Our analysis of the number of different head nouns in parenthesized text shows that 400 terms cover about 470000 Wikipedia articles (Figure 1). We manually created a list of such designators of named entities (album, song, footballer, etc) and common terms (general) and utilized the list in binary classification.

Bohn and Norvag [2] showed that categories are useful in generation of named entities of a certain class. They used category patterns such as 'Companies established in *' to recognize and classify named entities of PERSON, LOCATION and ORGANIZATION classes. We produce additional patterns corresponding to our classes. We also extend the idea and introduce patterns based on Wikipedia templates ('Infobox Museum', 'Infobox Book', 'Birth date and age').

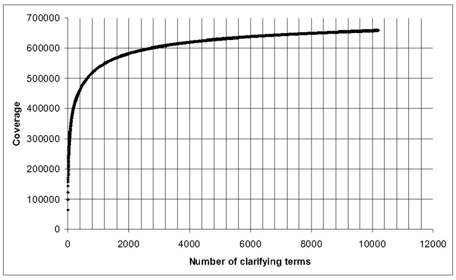

Figure 1: Number of clarifying terms plotted against their coverage

Links to foreign language articles covering the same topic as a given English one often preserve spelling. This is an indication in favor of the given article being an article about named entity. We count foreign articles with the same title and in case a 50% threshold is met we classify the article as an article about named entity.

In order to predict other classes we implement modified Bohn's variant of heuristic introduced by Bunescu [2, 4]. Following the Wikipedia naming conventions, article titles are capitalized if they are proper nouns. Consequently, a capitalized multi-word title may be considered a title of a named entity. However, the first word of an article title is always capitalized and this introduces prevents the heuristic to work for single-word titles. In order to resolve the issue with single words, we use a list of frequent lowercase words built from Wikipedia texts. We regard single-worded titles to be common terms if they are found in the list. If the title is not listed, we count occurrences of it within the article text. If at least α percents of the occurrences are capitalized, we decide that the article describes a named entity. The α threshold was chosen to be equal to 65% in accordance with Bohn.

5 Features for Text Classification

In order to make comprehensive class distinction between named entity classes, we implemented SVM-based classifiers. Our baseline method is similar to Tardif's classifier [15]. It uses a bag-of-words of the first paragraph of a Wikipedia article, title tokens, stemmed and tokenized category, and template names. Content tokens of 'Infobox', 'Sidebar' and 'Taxobox' templates are extracted and are also added to the representation. Category, template, and infobox features are extended with different prefixes. Feature set for unstructured text is reduced by removing stop-words and words with frequency less than 10 within trains set.

We experimented with different sets of Wikipedia features. We tested tables and tables headers as a generalization of infobox feature. We explored template argument tokens, section headers, clarifying terms, definition nouns of the first sentence in an article. None of the new features improved the results.

Our system is an extension of a baseline. We added 'list of' feature to the bag-of-words representation. A 'List of' article is analogous to a Wikipedia category but there is no restriction on how many 'List of' articles a given article may appear. 'List of X' article usually contains list of internal links to articles that are instances of X. For instance, "List of philanthropists" contains links to articles about people. For each Wikipedia article we construct the set of 'list of' articles that contain it. Tokenized and stemmed 'list of' article titles with the removed prefix 'List of ' are added to representation together with text tokens in order to retain relevant terms.

To increase separability between articles about named entities and articles about common terms, we add a boolean feature which is the result of binary classification heuristic (see the section Distinguishing Entities from Other Pages) to the model.

6 Experiments

6.1 Binary Classification

Binary classification was evaluated on a separate benchmark. We annotated 845 Wikipedia articles which were present on disambiguation articles of the most ambiguous terms presented in [11]. Named entities in that benchmark can fall outside the standard classes that we use. For example "AK-47" article refers to PRODUCT → WEAPON class in Sekine's hierarchy. We do not distinguish product subclasses. We took articles with the same titles but describing different notions (e.g. consider "Party" which can be legal, political, or birthday one). For these notions, an answer to the question if the article describes a named entity rather than a common word is difficult. Our evaluation shows that binary classifier demonstrates 95% precision and 95% recall.

6.2 Grained Classification: Data Set

We carried out experiments on different data sets. The first data set, namely CONLL_SET, consists of articles of four classes used at CoNLL evaluation [16] (PER, LOC, ORG, MISC). The second data set, to which we refer to as MAIN_SET, includes two additional classes, DISAMBIGUATION_PAGE and OHTER_PAGE. The third data set, the GRAINED_SET, is annotated to our guideline.

We generated training sets in semi-automatic manner. The main idea of our generation is to construct large gazetteers of named entities of a specified class and intersect the gazetteers with list of all Wikipedia titles. E.g., provided that we have gazetteer of automobiles, Wikipedia article is annotated with AUTO class if its title or a title of a redirect to it is listed in this gazetteer. To prevent ambiguity in annotation, we tagged articles that did not have corresponding disambiguation articles. The approach can lead to errors. E.g., Wikipedia has an article with a name X that should be classified as PERSON. However, X is listed in the GPE gazetteer. Thus, we needed an estimation of the train set quality.

```
...
<ul>
  <il>The <i>Sea of Japan</i> is a
    <i>marginal sea</i> of the
    western <i>Pacific Ocean</i>
  </il>
  <il><i>Dead Sea</i> ...</il>
  <il><i>Black Sea</i> ...</il>
  ...
</ul>
...
```

Figure 2: Example of HTML code to be wrapped.

Nadeau at al. [5] showed that it is possible to create accurate list of entities with small supervision. We follow similar procedure. Gazetteers are generated from the Web. First, we retrieve Web pages that satisfy a query of 4 (threshold was chosen according to [5]) manually classified entities ('Sea of Japan' AND 'Barents Sea' AND 'White Sea' AND 'Dead Sea'). The retrieved documents often contain other named entities of the same class as named entities in the query. If named entities are organized on a page in a regular HTML structure they can be extracted with the use of a web page wrapper.

Nadeau at al. referred to Cohen and Fan's wrapper [7] which manages to extract necessary information. We use a wrapper algorithm which processes only HTML tables and lists. If a candidate list contains seed example (perhaps, with small amount of additional text) surrounded by an HTML tag we extract sequence of tags that contain the seed example from the root of the HTML list. We named entity candidates have the same tags nesting. On figure 2, named entities 'Sea of Japan', 'Dead Sea', 'Black Sea' have the same tag nesting. We also see a noisy term 'marginal sea'. We include a candidate into the result list if it is capitalized and at list one token of the candidate is not contained in the list of frequent lowercase words of Wikipedia.

In order to create a training set for the class OTHER_PAGE, we used subset of frequent lowercase words constructed from Reuters corpus [10]. Disambiguation and 'list of' articles are annotated by heuristic methods proposed in section Distinguishing Named Entities from Other Pages.

Finally, we managed to annotate 5294 articles. Figure 3 shows classes distribution.

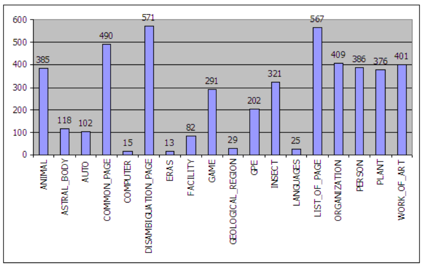

Figure 3: Classes distribution in the training set.

Class	Baseline P	 R	 F	System P	 R	 F
LOCATION	1	0.916	0.956	1	0.941	0.97
ORGANIZATION	0.952	1	0.976	1	1	1
PERSON	1	1	1	1	1	1
MISC	0.94	1	0.969	0.947	1	0.973
ALL (weighted avg.)	0.971	0.969	0.969	0.98	0.978	0.978

Table 5: Results for CONLL_SET.

To estimate the quality of the training set, we chose 40 (or less if the class had few examples in the training set) articles from each class and manually reviewed them. Figure 4 shows results of the review. Some part of manually reviewed and randomly selected from Wikipedia articles were selected into the test set. Test set included 550 Wikipedia articles.

6.3 Grained Classification: Evaluation

All experiments were run with LibSVM [6] using a linear kernel with C-value = 2.

Both systems evaluated by our test set and trained on semi-automatically generated test set. The comparison results of classifying Wikipedia articles for CONLL_SET, MAIN_SET and GRAINED_SET are reported in Tables 5, 6 and 7 respectively.

All tests show that our system outperformed baseline by averaged F-score. In CONLL_SET and MAIN_SET this is illustrated more clearly. For GRAINED_SET we got precision reduction small in contrast with baseline method.

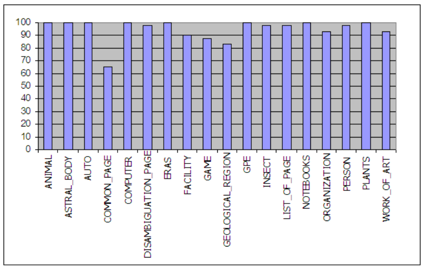

Figure 4: Train set quality estimation.

We also checked usefulness of two stage classifier, at the first, system perform heuristic to divide entities and common words. Secondly, we perform our SVM classifier trained to distinguish between entity classes. So, the results of this experiments are presented in Tables 6 and 7 in third column. Evaluation of this hierarchical classifier has not sense to CONLL_SET because only entities were considered. In two cases classification of OTHER_PAGE were increased, but with GRAINED_SET this did not lead to overall improvement. It make sense to MAIN_SET where MISC class collected entities of different grained classes and most misclassified articles were between MISC and OTHER_PAGE. When we performed this system to GRAINED_SET misclassification of the similar classes were increased and as result last column in Table 7 shows precision decrease. Thus hierarchical classifier with small portion of tags on its nodes may be useful for fine-grained named entity classification of Wikipedia.

	Baseline			System			Heuristic + System		
Class	P	R	F	P	R	F	P	R	F
LOCATION	1	0.874	0.933	1	0.916	0.956	1	0.949	0.974
ORGANIZATION	0.976	1	0.988	1	1	1	1	1	1
PERSON	1	1	1	1	1	1	1	1	1
MISC	0.924	0.77	0.84	0.927	0.913	0.92	0.933	1	0.966
DISAMBIGUATION_PAGE	1	1	1	1	1	1	1	1	1
OTHER_PAGE	0.361	0.846	0.506	0.605	0.885	0.719	0.88	0.92	0.90
ALL (weighted avg.)	0.93	0.877	0.894	0.95	0.938	0.942	0.97	0.98	0.974

Table 6: Results for MAIN_SET.

	Baseline			System			Heuristic + System		
Class	P	R	F	P	R	F	P	R	F
GPE	1	1	1	1	1	1	1	1	1
FACILITY	0.95	0.95	0.95	1	0.95	0.974	0.884	0.95	0.916
ASTRAL_BODY	1	0.975	0.987	1	0.975	0.987	1	0.975	0.987
AUTO	1	0.949	0.974	1	0.949	0.974	1	0.949	0.974
PLANT	0.976	1	0.988	0.976	1	0.988	0.976	1	0.988
COMPUTER	1	0.556	0.714	1	0.667	0.8	1	0.667	0.8
ANIMAL	1	0.95	0.974	1	0.95	0.974	1	0.95	0.974
PERSON	1	1	1	1	1	1	1	1	1
ORGANIZATION	0.976	1	0.988	0.952	1	0.976	0.952	1	0.976
WORK_OF_ART	0.864	0.95	0.905	0.709	0.975	0.821	0.727	1	0.842
INSECT	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975
GEOLOGICAL_REGION	1	0.436	0.607	0.962	0.641	0.769	0.963	0.667	0.788
LANGUAGES	1	0.947	0.973	1	0.947	0.973	1	0.947	0.973
GAME	0.927	0.974	0.95	0.927	0.974	0.95	0.884	0.974	0.927
ERAS	1	1	1	1	1	1	1	1	1
LIST_OF_PAGE	1	0.975	0.987	1	0.95	0.974	1	0.95	0.974
DISAMBIGUATION_PAGE	1	1	1	1	1	1	0.951	1	0.975
OTHER_PAGE	0.481	0.962	0.641	0.697	0.885	0.78	0.88	0.88	0.88
ALL (weighted avg.)	0.958	0.935	0.935	0.956	0.946	0.946	0.951	0.949	0.946

Table 7: Results for GRAINED_SET.

7 Conclusion

This work presented an approach to Wikipedia articles classification into classes of named entities. Our system uses two classifiers to produce final prediction. Binary classifier is used to distinguish between articles about common words and articles about named entities. SVM-based classifier is used to make a grained prediction. On our benchmarks we outperformed previous approaches. The high-quality classification is useful for a number of NLP tasks.

We also proposed a method of Wikipedia articles annotation with small supervision. In that way we construct train set containing over 5000 articles. We proposed way to generate Wikipedia annotation with small supervision. It make easy to select additional named entity class to our annotation.

References

[1] Abhijit Bhole, Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. Extracting named entities and relating them over time based on wikipedia. *Informatica*, 2007.

[2] Christian Bohn and Kjetil Norvag. Extracting named entities and synonyms from wikipedia. *Advanced Information Networking and Applications, International Conference on*, 0:1300–1307, 2010.

[3] Ada Brunstein. Annotation guidelines for answer types, 2002.

[4] Razvan C. Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*. The Association for Computer Linguistics, 2006.

[5] Council Canada, David Nadeau, Peter D. Turney, and Stan Matwin. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity, 2006.

[6] Chih chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines, 2001.

[7] William W. Cohen and Wei Fan. Learning page-independent heuristics for extracting data from web pages. In *In AAAI Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[8] W. Dakka and S. Cucerzan. Augmenting wikipedia with named entity tags. In *Proceedings of IJCNLP 2008*, 2008.

[9] Adrian Iftene and Alexandra Balahur-Dobrescu. Named entity relation mining using wikipedia. In *LREC*. European Language Resources Association, 2008.

[10] David D. Lewis, Yiming Yang, Tony G. Rose, Fan Li, G. Dietterich, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[11] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[12] Joel Nothman, Tara Murphy, and James R. Curran. Analysing wikipedia and gold-standard corpora for ner training. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 612–620, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[13] Iman Saleh, Kareem Darwish, and Aly Fahmy. Classifying wikipedia articles into ne's using svm's with threshold adjustment. In *Proceedings of the 2010 Named Entities Workshop*, pages 85–92, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[14] S. Sekine, K. Sudo, and C. Nobata. Extended named entity hierarchy. In M. Gonzáles Rodríguez and C. Paz Suárez Araujo, editors, *Proceedings of* 3^{rd} *International Conference on Language Resources and Evaluation (LREC'02)*, pages 1818–1824, Canary Islands, Spain, May 2002.

[15] Sam Tardif, James R Curran, and Tara Murphy. Improved text categorisation for wikipedia named entities. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 104–108, Sydney, Australia, December 2009.

[16] Erik F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan, 2002.

[17] Antonio Toral. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *In EACL 2006*, 2006.

[18] Antonio Toral, Rafael Munoz, and Monica Monachini. Named entity wordnet. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. ISBN: 2-9517408-4-0.

[19] Yotaro Watanabe, Masayuki Asahara, and Yuji Matsumoto. A graph-based approach to named entity categorization in wikipedia using conditional random fields. In *EMNLP-CoNLL*, pages 649–657. ACL, 2007.

[20] Cäcilia Zirn, Vivi Nastase, and Michael Strube. Distinguishing between instances and classes in the wikipedia taxonomy. In *ESWC*, pages 376–387, 2008.