Augmenting Wikipedia with Named Entity Tags

Wisam Dakka
Columbia University
1214 Amsterdam Avenue
New York, NY 10027
wisam@cs.columbia.edu

Silviu Cucerzan
Microsoft Research
1 Microsoft Way
Redmond, WA 98052
silviu@microsoft.com

Abstract

Wikipedia is the largest organized knowledge repository on the Web, increasingly employed by natural language processing and search tools. In this paper, we investigate the task of labeling Wikipedia pages with standard named entity tags, which can be used further by a range of information extraction and language processing tools. To train the classifiers, we manually annotated a small set of Wikipedia pages and then extrapolated the annotations using the Wikipedia category information to a much larger training set. We employed several distinct features for each page: bag-of-words, page structure, abstract, titles, and entity mentions. We report high accuracies for several of the classifiers built. As a result of this work, a Web service that classifies any Wikipedia page has been made available to the academic community.

1 Introduction

Wikipedia, one of the most frequently visited web sites nowadays, contains the largest amount of knowledge ever gathered in one place by volunteer contributors around the world (Poe, 2006). Each Wikipedia article contains information about one entity or concept, gathers information about entities of one particular type of entities (the so-called *list pages*), or provides information about homonyms (*disambiguation pages*). As of July 2007, Wikipedia contains close to two million articles in English. In addition to the English-language version, there are 200 versions in other languages. Wikipedia has about 5 million registered contributors, averaging more than 10 edits per contributor.

Natural language processing and search tools can greatly benefit from Wikipedia by using it as an authoritative source of common knowledge and by exploiting its interlinked structure and disambiguation pages, or by extracting concept co-occurrence information. This paper presents a successful study on enriching the Wikipedia data with named entity tags. Such tags could be employed by disambiguation systems such as Bunescu and Paşca (2006) and Cucerzan (2007), in mining relationships between named entities, or in extracting useful facet terms from news articles (e.g., Dakka and Ipeirotis, 2008).

In this work, we classify the Wikipedia pages into categories similar to those used in the CoNLL shared tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) and ACE (Doddington et al., 2004). To the best of our knowledge, this is the first attempt to perform such classification on the English language version of the collection. [1] Although the task settings are different, the results we obtained are comparable with those previously reported in document classification tasks.

We examined the Wikipedia pages to extract several feature groups for our classification task. We also observed that each entity/concept has at least two pseudo-independent views (page-based features and link-based features), which allow the use a co-training method to boost the performance of classifiers trained separately on each view.

The classifier that achieved the best accuracy on out test set was applied then to all Wikipedia pages and its classifications are provided to the academic community for use in future studies through a Web service.[2]

[1] Watanabe et al. (2007) have reported recently experiments on categorizing named entities in the Japanese version of Wikipedia using a graph-based approach.

[2] The Web service is available at *wikinet.stern.nyu.edu*.

2 Related Work

This study is related to the area of named entity recognition, which has supported extensive evaluations (CoNLL and ACE). Since the introduction of this task in MUC-6 (Grishman and Sundheim, 1996), numerous systems using various ways of exploiting entity-specific and local context features were proposed, from relatively simple character-based models such as Cucerzan and Yarowsky (2002) and Klein et al. (2003) to complex models making use of various lexical, syntactic, morphological, and orthographical information, such as Wacholder et al. (1997), Fleischman and Hovy (2002), and Florian et al. (2003). While the task we address is not the conventional named entity recognition but rather document classification, our classes are a derived from the labels traditionally employed in named entity recognition, following the CoNLL and ACE guidelines, as described in Section 3.

The areas of text categorization and document classification have also been extensively researched over time. These task have the goal of assigning to each document in a collection one or several labels from a given set, such as Newsgroups (Lang, 1995), Reuters (Reuters, 1997), Yahoo! (Mladenic, 1998), Open Directory Project (Chakrabarti et al., 2002), and Hoover's Online (Yang et al., 2002). Various supervised machine learning algorithms have been applied successfully to the document classification problem (e.g., Joachims, 1999; Quinlan, 1993; Cohen, 1995). Dumais et al. (1998) and Yang and Liu (1999) reported that support vector machines (SVM) and K-Nearest Neighbor performed the best in text categorization. We adopted SVM as our algorithm of choice because of these findings and also because SVMs have been shown robust to noise in the feature set in several studies. While Joachims (1998) and Rogati and Yang (2002) reported no improvement in SVM performance after applying a feature selection step, Gabrilovich and Markovitch (2004) showed that for collection with numerous redundant features, aggressive feature selection allowed SVMs to actually improve their performance. However, performing an extensive investigation of classification performance across various machine learning algorithms has been beyond the purpose of this work, in which we ran classification experiments using SVMs and compared them only with the results of similar systems employing Naïve Bayes.

In addition to the traditional bag-of-words, which has been extensively used for the document classification task (e.g. Sebastiani, 2002), we employed various other Wikipedia-specific feature sets. Some of these have been previously employed for various tasks by Gabrilovich and Markovitch, (2006); Overell and Ruger (2006), Cucerzan (2007), and Suchanek et al. (2007).

3 Classifying Wikipedia Pages

The Wikipedia pages that we analyzed in this study can be divided into three types:

Disambiguation Page (DIS): is a special kind of page that usually contains the word "disambiguation" in its title, and that contains several possible disambiguations of a term.

Common Page (COMM): refers to a common object rather than a named entity. Generally, if the name of an object or concept appears non-capitalized in text then it is very likely that the object or the concept is of common nature (heuristic previously employed by Bunescu and Paşca, 2006). For example, the Wikipedia page "Guitar" refers to a common object rather than a named entity.

Named Entity Page: refers to a specific object or set of objects in the world, which is/are commonly referred to using a certain proper noun phrase. For example, any particular person is a named entity, though the concept of "people" is not a named entity. Note that most names are ambiguous. "Apollo" can refer to more than 30 different entities of different types, for example, the Finnish rock band of the late 1960s/early 1970s , the Greek god of light, healing, and poetry, and the series of space missions run by NASA.

To classify the named entities in Wikipedia, we adopted a restricted version of the ACE guidelines (ACE), using four main entity classes (also similar to the classes employed in the CoNLL evaluations):

Animated Entities (PER): An animate entity can be either of type human or non-human. **Human** entities are either humans that are known to have lived (e.g., "Leonardo da Vinci", "Britney Spears", "Gotthard of Hildesheim", "Saint Godehard") or humanoid individuals in fictional works, such as books, movies, TV shows, and comics (e.g., "Harry Potter", "Batman", "Sonny" the robot from the movie “I, Robot”). Fictional characters also include mythological figures and deities (e.g. “Zeus”, “Apollo”, “Jupiter”). The fictional nature of a character must be explicitly indicated. **Non-human** entities are any particular animal or alien that has lived or that is described in a fictional work and can be singled out using a name.

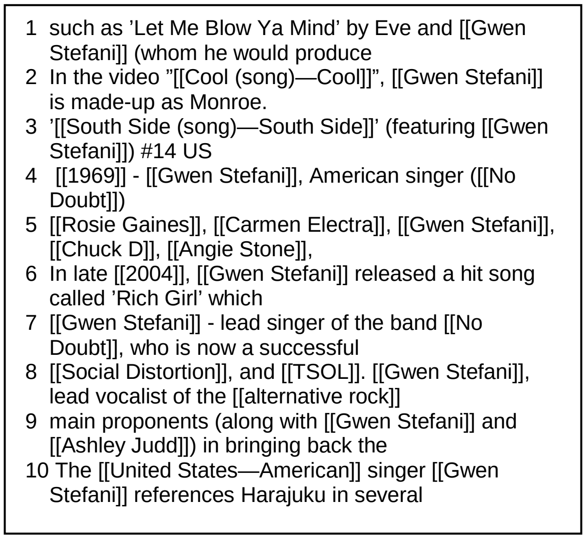
Organization Entities (ORG): An organization entity must have some formally established association. Typical examples are businesses (e.g., “Microsoft”, “Ford”), governmental bodies (e.g., “United States Congress”), non-governmental organizations (e.g., “Republican Party”, “American Bar Association”), science and health units (e.g., “Massachusetts General Hospital”), sports organizations and teams (e.g., “Angolan Football Federation”, “San Francisco 49ers”), religious organizations (e.g., “Church of Christ”), and entertainment organizations, including formally organized music groups (e.g., “San Francisco Mime Troupe”, the rock band “The Police”). Industrial sectors and industries (e.g., “Petroleum industry”) are also treated as organization entities, as well as all media and publications.

Location Entities (LOC): These are physical locations (regions in space) defined by geographical, astronomical, or political criteria. They are of three types: **Geo-Political** entities are composite entities comprised of a physical location, a population, a government, and a nation (or province, state, county, city, etc.). A Wikipedia page that mentions all these components should be labeled as Geo-Political Entity (e.g., “Hawaii”, “European Union”, “Australia”, and “Washington, D.C.”). **Locations** are places defined on a geographical or astronomical basis and do not constitute a political entity. These include mountains, rivers, seas, islands, continents (e.g., “the Solar system”, “Mars”, “Hudson River”, and “Mount Rainier”). **Facilities** are artifacts in the domain of architecture and civil engineering, such as buildings and other permanent man-made structures and real estate improvements: airports, highways, streets, etc.

Miscellaneous Entities (MISC): About 25% of the named entities in Wikipedia are not of the types listed above. By examining several hundred examples, we concluded that the majority of these named entities can be classified in one of the following classes: **Events** refer to historical events or actions with some certain duration, such as wars, sport events, and trials (e.g., “Gulf War”, “2006 FIFA World Cup”, “Olympic Games”, “O.J. Simpson trial”). **Works of art** refer to named works that are imaginative in nature. Examples include books, movies, TV programs, etc. (e.g., the “Batman” movie, “The Tonight Show”, the “Harry Potter” books). **Artifacts** refer to man-made objects or products that have a name and cannot generally be labeled as art. This includes mass-produced merchandise and lines of products (e.g. the camera “Canon PowerShot Pro1”, the series “Canon PowerShot”, the type of car “Ford Mustang”, the software “Windows XP”). Finally **Processes** include all named physical and chemical processes (e.g., “Ettinghausen effect”). Abstract formulas or algorithms that have a name are also labeled as processes (e.g., “Naive Bayes classifier”).

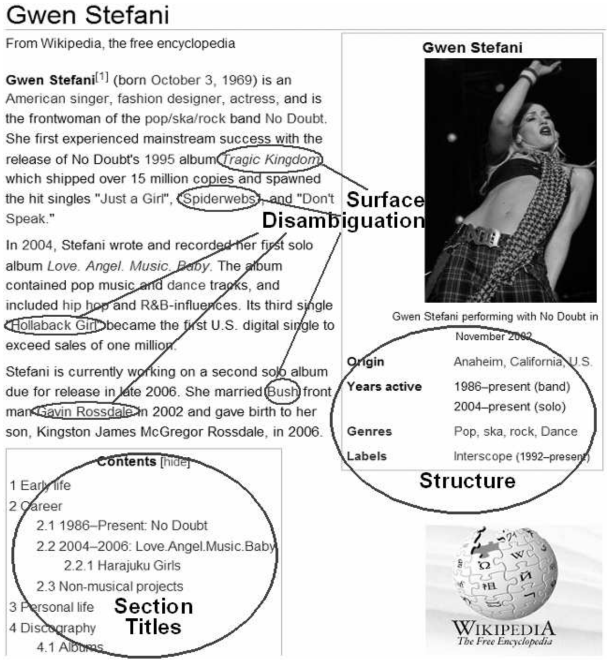
4 Features Used. Independent Views

When creating a Wikipedia page and introducing a new entity, contributors can refer to other related Wikipedia entities, which may or may not have corresponding Wikipedia pages. This way of generating content creates an internal web graph and, interesting, results in the presence of two different and pseudo-independent views for each entity. We can represent an entity using the content written on the entity page, or alternatively, using the context from a reference on the related page. For example, Figures 1 and 2 show the two independent views of the entity “Gwen Stefani”.



1 such as ’Let Me Blow Ya Mind’ by Eve and [[Gwen Stefani]] (whom he would produce
2 In the video ”[[Cool (song)—Cool]]”, [[Gwen Stefani]] is made-up as Monroe.
3 ’[[South Side (song)—South Side]]’ (featuring [[Gwen Stefani]]) #14 US
4 [[1969]] - [[Gwen Stefani]], American singer ([[No Doubt]])
5 [[Rosie Gaines]], [[Carmen Electra]], [[Gwen Stefani]], [[Chuck D]], [[Angie Stone]],
6 In late [[2004]], [[Gwen Stefani]] released a hit song called ’Rich Girl’ which
7 [[Gwen Stefani]] - lead singer of the band [[No Doubt]], who is now a successful
8 [[Social Distortion]], and [[TSOL]]. [[Gwen Stefani]], lead vocalist of the [[alternative rock]]
9 main proponents (along with [[Gwen Stefani]] and [[Ashley Judd]]) in bringing back the
10 The [[United States—American]] singer [[Gwen Stefani]] references Harajuku in several

Figure 1. A partial list of contextual references taken from Wikipedia for the named entity “Gwen Stefani”. (There are over 600 such references.)



Gwen Stefani

From Wikipedia, the free encyclopedia

Gwen Stefani[1] (born October 3, 1969) is an American singer, fashion designer, actress, and is the frontwoman of the pop/ska/rock band No Doubt. She first experienced mainstream success with the release of No Doubt's 1995 album *Tragic Kingdom*, which shipped over 15 million copies and spawned the hit singles "Just a Girl", "Spiderwebs", and "Don't Speak."

In 2004, Stefani wrote and recorded her first solo album *Love. Angel. Music. Baby.* The album contained pop music and dance tracks, and included hip hop and R&B-influences. Its third single "Hollaback Girl" became the first U.S. digital single to exceed sales of one million.

Stefani is currently working on a second solo album due for release in late 2006. She married Bush front man Gavin Rossdale in 2002 and gave birth to her son, Kingston James McGregor Rossdale, in 2006.

Gwen Stefani

Gwen Stefani performing with No Doubt in November 2002

Origin	Anaheim, California, U.S.
Years active	1986–present (band) 2004–present (solo)
Genres	Pop, ska, rock, Dance
Labels	Interscope (1992–present)

Contents [hide]

1 Early life
2 Career
2.1 1986–Present: No Doubt
2.2 2004–2006: Love.Angel.Music.Baby
2.2.1 Harajuku Girls
2.3 Non-musical projects
3 Personal life
4 Discography
4.1 Albums

Figure 2. Wikipedia page for the named entity “Gwen Stefani”. Other than the regular text, information such as surface and disambiguated entities, structure properties, and section titles can be easily extracted.

We utilize this important observation to extract our features based on these two independent views: page-based features and context features. We discuss these in greater detail next.

4.1 Page-Based Features

A typical Wikipedia page is usually written and edited by several contributors. Each page includes a rich set of information including the following elements: titles, section titles, paragraphs, multimedia objects, hyperlinks, structure data, surface entities and their disambiguations. Figure 2 shows some of these elements in the page dedicated to singer “Gwen Stefani”. We use the Wikipedia page XML syntax to draw a set of different page-based feature vectors, including the following:

Bag of Words (BOW): This vector is the term frequency representation of the entire page.

Structured Data (STRUCT): Many Wikipedia pages contain useful data organized in tables and other structural representations. In Figure 2, we see that contributors have used a table representation to list different properties about Gwen Stefani. We extract for each page, using the Wikipedia syntax, the bag-of-words feature vector that corresponds to this structured data only.

```
<abstract>
   Gwen Rene StefaniSome sources give Stefani's first name
   as Gwendolyn, but her first name is simply Gwen. Her list-
   ing on the California Birth Index from the Center for Health
   Statistics gives a birth name of Gwen Rene Stefani.
</abstract>
```

Figure 3. The abstract provided by Wikipedia for “Gwen Stefani”. Note the concatenation of “Stefani” and “Some”, which results in a new word, and is a relevant example of noise encountered in Wikipedia text.

First Paragraph (FPAR): We examined several hundred pages, and observed that a human could label most of the pages by reading only the first paragraph. Therefore, we built the feature vector that contains the bag-of-word representation of the page’s first paragraph.

Abstract (ABS): For each page, Wikipedia provides a summary of several lines about the entity described on the page. We use this summary to draw another bag-of-word feature vector based on the provided abstracts only. For example, Figure 3 shows the abstract for the entity “Gwen Stefani”.

Surface Forms and Disambiguations (SFD): Contributors use the Wikipedia syntax to link from one entity page to another. In the page of Figure 2, for example, we have references to several other Wikipedia entities, such as “hip hop”, “R&B”, and “Bush”. Wikipedia page syntax lets us extract the disambiguated meaning of each of these references, which are “Hip hop music,” “Rhythm and blues,” and “Bush band”, respectively. For each page, we extract all the surface forms used by contributors in text (such as “hip hop”) and their disambiguated meanings (such as “Hip hop music”), and build feature vectors to represent them.

4.2 Context Features

Figure 1 shows some of the ways contributors to Wikipedia refer to the entity “Gwen Stefani”. The Wikipedia version that we analyzed contains about 35 million references to entities in the collection. On average, each page has five references to other entities.

We decided to make use of the text surrounding these references to draw contextual features, which can capture both syntactic and semantic properties of the referenced entity. For each entity reference, we compute the feature vectors by using a text window of three words to the left and to the right of the reference.

BOW	1,821,966	ABS	372,909
SFD	847,857	BCON	35,178,120
STRUCT	159,645	FPAR	781,938

Table 1. Number of features in each group, as obtained by examining all the Wikipedia pages.

We derived a unigram context model and a bigram context model, following the findings of previous work that such models benefit from employing information about the position of words relative to the targeted term:

Unigram Context (UCON): The feature vector is constructed in a way that preserves the positional information of words in the context. Each feature f^t_i in the vector represents the total number of times a term *t* appears in position *i* around the entity.

Bigram Context (BCON): The bigram-based context model was built in a similar way to UCON, so that relative positional information is preserved.

5 Challenges

For our classification task, we faced several challenges. First, many Wikipedia entities have only a partial list of the feature groups discussed above. For example, contributors may refer to entities that do not exist in Wikipedia but might be added in the future. Also, not all the page-based features groups are available for every entity page. For instance, abstracts and structure features are only available for 68% and 79% of the pages, respectively. Second, we only had available several hundred labeled examples (as described in Section 6.1). Third, the feature space is very large compared to the typical text classification problem (see Table 1), and a substantial amount of noise plagues the data. A further investigation revealed that the difference in the dimensionality compared to text classification stems from the way Wikipedia pages are created: contributors make spelling errors, introduce new words, and frequently use slang, acronyms, and other languages than English.

We utilize all the features groups described in Section 4 and various combinations of them. This provides us with greater flexibility to use classifiers trained on different feature groups when Wikipedia entities miss certain types of features.

In addition, we try to take advantage of the independent views of each entity by employing a co-training procedure (Blum and Mitchell, 1998; Nigam and Ghani, 2000). In previous work, this has been shown to boost the performance of the weak classifiers on certain feature groups. For example, it is interesting to determine whether we can use the STRUCT view of a Wikipedia pages to boost the performance of the classifiers based on context. Alternatively, we can employ co-training on the STRUCT and SFD features, hypothesized as two independent views of the data.

6 Experiments and Findings

6.1 Training Data

We experimented with two data sets: **Human Judged Data** (HJD): This set was obtained in an annotation effort that followed the guidelines presented in Section 3. Due to the cost of the labeling procedure, this set was limited to a small random set of 800 Wikipedia pages. **Human Judged Data Extended** (HJDE): The initial classification results obtained using a small subset of HJD hinted to the need for more training data. Therefore, we devised a procedure that takes advantage of the fact that Wikipedia contributors have assigned many of the pages to one or more lists. For example, the page "List of novelists" contains a reference to "Orhan Pamuk", which is part of the HJD and is labeled as PER. Our extension procedure first uses the pages in the training set from HJD to extract the lists in Wikipedia that contain references to them and then projects the entity labels of the seeds to all elements in the lists. Unfortunately, not all the Wikipedia lists contain only references named entities of the same category. Furthermore, some lists are hierarchical and include sub-lists of different classes. To overcome these issues, we examined only leaf lists and manually filtered all the lists that by definition could have pages of different categories. Finally, we filtered out all list pages that contain entities in two or more entity classes (as described in Section 3).

Our partially manual extension procedure is as follows: 1) Pick a random sample of 400 entities from HJD along with their human judged labels; 2) Extract all the lists that contain any entity from this labeled sample; 3) Filter out the lists that contain entities from different entity classes (PER, ORG, LOC, MISC, and COM); 4) propagate the entity labels of the known entities in the lists to the other referenced entities; 5) Choose a random sample from all labeled pages with respect to the entity class distribution observed in HJD**.**

PER	MISC	ORG	LOC	COMM
41%	25.1%	11.2%	11.7%	11%

Table 2. The distribution of labels in the HJDE data set.

Our extension procedure resulted initially in 770 lists, which were then reduced to 501. In step (5), we chose a maximal random sample from all labeled pages in HJDE so that it matched the entity class distribution in the original HJD training set (shown in Table 2).

6.2 Classification

From the numerous machine learning algorithms available for our classification task (e.g., Joachims, 1999; Quinlan, 1993; Cohen, 1995), we chose to the SVMs (Vapnik, 1995), and the Naïve Bayes (John and Langley, 1995) algorithms because both can output probability estimates for their predictions, which are necessary for the co-training procedure. We use an implementation of SVM (Platt, 1999) with linear kernels and the Naïve Bayes implementation from the machine learning toolkit Weka3. Our implementation of co-training followed that of Nigam and Ghani (2000).

Using the HJDE data, we experimented with learning a classifier for each feature group discussed in Section 4. We report the results for two classification tasks: binary classification to identify all the Wikipedia pages of type PER, and 5-fold classification (PER, COM, ORG, LOC, and MISC).

To reduce the feature space, we built a term frequency dictionary taken from one year's worth of news data and restrict our feature space to contain only terms with frequency values higher than 10.

6.3 Results on Bag-of-words

This feature group is of particular interest, since it has been widely used for document classification and also, because every Wikipedia page has a BOW representation. We experimented with the two classification tasks for this feature group. For the binary classification task, both SVM and Naïve Bayes performed remarkably well, obtaining accuracies of 0.962 and 0.914, respectively. Table 3 shows detailed performance numbers for SVM and Naïve Bayes for the multi-class task. Unlike in the binary case, Naïve Bayes falls short of achieving results similar to those from SVM, which obtains an average F-measure of 0.928 and an average precision of 0.931.

	Precision		Recall		F-measure	
	SVM	NB	SVM	NB	SVM	NB
PER	0.944	0.918	0.959	0.771	0.951	0.838
MISC	0.927	0.824	0.920	0.687	0.924	0.750
ORG	0.940	0.709	0.928	0.701	0.934	0.705
LOC	0.958	0.459	0.949	0.863	0.954	0.599
COMM	0.887	0.680	0.869	0.714	0.878	0.697

Table 3. Precision, recall, and F1 measure for the multi-class classification task. Results are obtained using SVM and Naïve Bayes after a stratified cross-validation using HJDE data set and the bag-of-words features.

SFD	83.14%	ABS	68.96%
STRUCT	79.55%	BCON	83.57%

Table 4. Percentage of available examples HJDE for each feature group.

	Precision		Recall		F-measure	
	SVM	NB	SVM	NB	SVM	NB
BOW	**0.901**	0.858	0.894	0.880	**0.897**	0.869
SFD	0.851	0.775	0.830	0.882	0.840	0.825
STRUCT	**0.888**	0.840	0.875	0.856	**0.881**	0.848
FPAR	0.867	0.872	0.854	0.896	0.860	**0.884**
ABS	0.861	0.833	0.852	0.885	0.857	0.858
BCON	0.311	0.245	0.291	0.334	0.300	0.283

Table 5. Average precision, recall, and F1 measure values for the multi-class task. Results are obtained using SVM and Naïve Bayes across the different feature groups on the test set of HJDE.

6.4 Results on Other Feature Groups

We present now the results obtained using other groups of features. We omit the results on UCON due to their similarity with BCON. Recall that these features may not be present in all Wikipedia pages. Table 4 shows the availability of these features in the HJDE set. The lack of one feature group has a negative impact on the results of the corresponding classifier, as shown in Table 5. Noticeably, the results of the STRUCT features are very encouraging and confirm our hypothesis that such features are distinctive in identifying the type of the page. While results using STRUCT and FPAR are high, they are lower than the results obtained on BOW. In general, using SVM with BOW performed better than any other feature set, averaging 0.897 F-measure on test set. This could be because when using BOW, we have a larger training set than any other feature group. SVM with STRUCT and Naïve Bayes with FPAR performed second and third best, with average F1 measure values of 0.881 and 0.860, respectively. The results also show that it is difficult to learn if a page is COMM in all learning combination. This could be related to the membership complexity of that class. Finally, the results on the bigram contextual features, namely BCON, for both SVM and Naïve Bayes are not encouraging and surprisingly low.

6.5 Results for Co-training

Motivated by the fact that some feature groups can be seen as independent views of the data, we used a co-training procedure to boost the classification accuracy. One combination of views that we examined is BCON with BOW, hoping to boost the classification performance of the bigram context features, as this classifier could be used for entities in any new text, not only for Wikipedia pages . Unfortunately, the results were not encouraging in either of the cases (SVM and Naïve Bayes) and for none of the other feature groups used instead of BOW. This indicates that the context features extracted have limited power and that further investigation of extracting relevant context features from Wikipedia is necessary.

7 Conclusions and Future Work

In this paper, we presented a study on the classification of Wikipedia pages with named entity labels. We explored several alternatives for extracting useful page-based and context-based features such as the traditional bag-of-words, page structure, hyperlink text, abstracts, section titles, and *n*-gram contextual features. While the classification with page features resulted in high classification accuracy, context-based and structural features did not work similarly well, either alone or in a co-training setup. This motivates future work to extract better such features. We plan to examine employing more sophisticated ways both for extracting contextual features and for using the implicit Wikipedia graph structure in a co-training setup.

Recently, the Wikipedia foundation has been taken steps toward enforcing a more systematic way to add useful structured data on each page by suggesting templates to use when a new page gets added to the collection. This suggests that in a not-so-distant future, we may be able to utilize the structured data features as attribute-value pairs rather than as bags of words, which is prone to losing valuable semantic information.

Finally, we have applied our classifier to all Wikipedia pages to determine their labels and made these data available in the form of a Web service, which can positively contribute to future studies that employ the Wikipedia collection.

References

ACE Project. At http://www.nist.gov/speech/history/index.htm

Reuters-1997. 1997. Reuters-21578 text categorization test collection. At http://www.daviddlewis.com/resources/testcollections/reuters21578

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT'98*, pages 92–100.

A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. 1998. NYU: Description of the MENE named entity system as used in MUC. In *Proceedings of MUC-7*.

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL-2006*, pages 9–16.

S. Chakrabarti, M.M. Joshi, K. Punera, and D.M. Pennock. 2002. The structure of broad topics on the web. In: *Proceedings of WWW '02*, pages 251–262.

W.W. Cohen. 1995. Fast effective rule induction. In *Proceedings of ICML'95*.

S. Cucerzan. 2007. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of EMNLP-CoNLL 2007*, pages 708–716.

S. Cucerzan and D. Yarowsky. 2002. Language Independent NER using a Unified Model of Internal and Contextual Evidence, in *Proceedings of CoNLL 2002*, pages 171–174.

W. Dakka and P. G. Ipeirotis. 2008. Automatic Extraction of Useful Facet Terms from Text Documents. In *Proceedings of ICDE 2008 (to appear).*

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. ACE program – task definitions and performance measures. In *Proceedings of LREC*, pages 837–840.

S. Dumais, J. Platt, D. Heckerman, and M. Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of CIKM '98*, pages 148–155.

M. Fleischman and E. Hovy. 2002. Fine Grained Classification of Named Entities. In *Proceedings of COLING'02*, pages 267–273.

R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, Named Entity Recognition through Classifier Combination, in *Proceedings of CoNLL 2003*, pages 168–171.

E. Gabrilovich and S. Markovitch. 2004. Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with c4.5. In *Proceedings of ICML '04*, page 41.

E. Gabrilovich and S. Markovitch. 2006. Overcoming-the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of AAI 2006*.

R. Grishman and B. Sundheim. 1996. Message Understanding Conference - 6: A brief history. In *Proceedings of COLING*, 466-471.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML '98*, pages 137–142.

T. Joachims. 1999. Making large-scale support vector machine learning practical. *Advances in kernel methods: support vector learning*, pages 169–184.

G.H. John and P. Langley. 1995. Estimating continuous distributions in Bayesian classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

D. Klein, J. Smarr, H. Nguyen, and C. D. Manning. 2003. Named Entity Recognition with Character-Level Models, in *Proceedings of CoNLL 2003*.

K. Lang. 1995. NewsWeeder: Learning to filter netnews. In Proceedings of ICML'95, pages 331–339.

D. Mladenic. 1998. Feature subset selection in text learning. In *Proceedings of ECML '98*, pages 95–100.

K. Nigam and R. Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *Proceedings of CIKM'00*, pages 86–93.

S.E. Overell and S. Ruger. 2006. Identifying and grounding descriptions of places. In *Workshop on Geographic Information Retrieval, SIGIR 2006*.

J.C. Platt. 1999. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods: support vector learning*, pages 185–208.

M. Poe. 2006. The hive: Can thousands of wikipedians be wrong? How an attempt to build an online encyclopedia touched off history's biggest experiment in collaborative knowledge. The Atlantic Monthly, September 2006.

J.R. Quinlan. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc.

M. Rogati and Y. Yang. 2002. High-performing feature selection for text classification. In *Proceedings of CIKM '02*, pages 659–661.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing. Surveys*, 34(1):1–47.

F.M. Suchanek, G. Kasneci, and G. Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of WWW 2007*.

E.F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147.

E. F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158.

V.N. Vapnik. 1995. The nature of statistical learning theory. Springer-Verlag New York, Inc.

N. Wacholder., Y. Ravin, and M. Choi. 1997. Disambiguation of proper names in text. In *Proceedings of ANLP'97*, pages 202-208.

Y. Watanabe, M. Asahara, and Y. Matsumoto. 2007. A Graph-based Approach to Named Entity Categorization in Wikipedia using Conditional Random Fields. In *Proc. of EMNLP-CoNLL 2007*, pages 649-657.

Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR '99*, pages 42–49.

Y. Yang, S. Slattery, and R. Ghani. 2002. A study of approaches to hypertext categorization. *Journal of. Intelligent. Information. Systems.*, 18(2-3):219–241.