Classifying Wikipedia Articles into NE's using SVM's with Threshold Adjustment

Iman Saleh
Faculty of Computers and Information, Cairo University
Cairo, Egypt
iman.saleh@fci-cu.edu.eg

Kareem Darwish
Cairo Microsoft Innovation Center
Cairo, Egypt
kareemd@microsoft.com

Aly Fahmy
Faculty of Computers and Information, Cairo University
Cairo, Egypt
a.fahmy@fci-cu.edu.eg

Abstract

In this paper, a method is presented to recognize multilingual Wikipedia named entity articles. This method classifies multilingual Wikipedia articles using a variety of structured and unstructured features and is aided by cross-language links and features in Wikipedia. Adding multilingual features helps boost classification accuracy and is shown to effectively classify multilingual pages in a language independent way. Classification is done using Support Vectors Machine (SVM) classifier at first, and then the threshold of SVM is adjusted in order to improve the recall scores of classification. Threshold adjustment is performed using beta-gamma threshold adjustment algorithm which is a post learning step that shifts the hyperplane of SVM. This approach boosted recall with minimal effect on precision.

1 Introduction

Since its launch in 2001, Wikipedia has grown to be the largest and most popular knowledge base on the web. The collaboratively authored content of Wikipedia has grown to include more than 13 million articles in 240 languages.[1] Of these, there are more than 3 million English articles covering a wide range of subjects, supported by 15 million discussion, disambiguation, and redirect pages.[2] Wikipedia provides a variety of structured, semi-structured and unstructured resources that can be valuable in areas such information retrieval, information extraction, and natural language processing. As shown in Figure 1, these resources include page redirects, disambiguation pages, informational summaries (infoboxes), cross-language links between articles covering the same topic, and a hierarchical tree of categories and their mappings to articles.

Many of the Wikipedia pages provide information about concepts and named entities (NE). Identifying pages that provide information about different NE's can be of great help in a variety of NLP applications such as named entity recognition, question answering, information extraction, and machine translation (Babych and Hartley, 2003; Dakka and Cucerzan, 2008). This paper attempts to identify multilingual Wikipedia pages that provide information about different types of NE, namely persons, locations, and organizations. The identification is done using a Support Vector Machines (SVM) classifier that is trained on a variety of Wikipedia features such as infobox attributes, tokens in text, and category links for different languages aided by cross-language links in pages. Using features from different languages helps in two ways, namely: clues such infobox attributes may exist in one language, but not in the other, and this allows for tagging pages in multiple languages simultaneously. To improve SVM classification beta-gamma threshold adjustment was used to improve recall of different NE classes and consequently overall F measure.

The separating hyperplane suggested by the SVM typically favors precision at the cost of recall and needs to be translated (via threshold adjustment) to tune for the desired evaluation metric.

Beta-gamma threshold adjustment was generally used when certain classes do not have a sufficient number of training examples, which may lead to poor SVM recall scores (Shanahan and Roma, 2003). It was used by Shanahan and Roma (2003) to binary classify a set of articles and proved to improve recall with little effect on precision.

[1] http://en.wikipedia.org/wiki/Wikipedia
[2] http://en.wikipedia.org/wiki/Special:Statistics

Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 85–92,
Uppsala, Sweden, 16 July 2010.

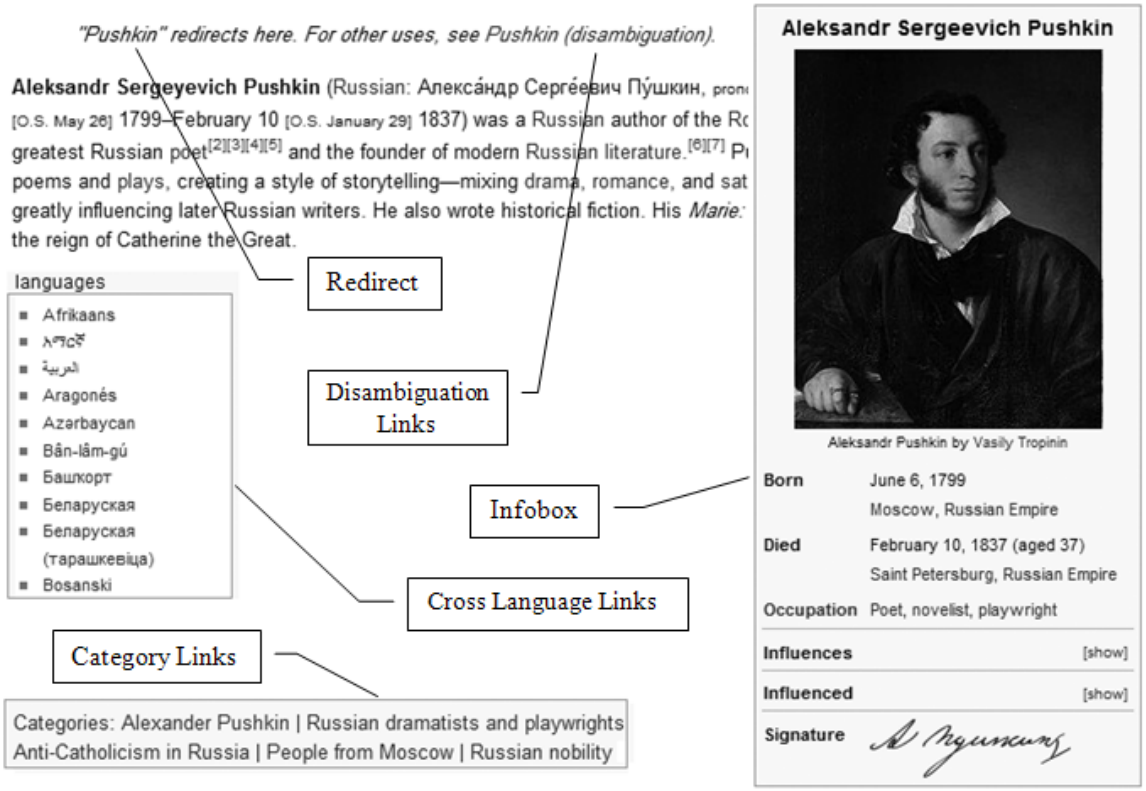

Figure 1. Sample Wikipedia article

However, the technique seems to generalize beyond cases where very few training examples are present, and it is shown in this paper to yield improvements in recall and overall F-measure in the presence of hundreds of training examples, performing better than threshold adjustment using cross validation for the specific task at hand.

The contribution of this paper lies in: introducing a language independent system that utilizes multilingual features from Wikipedia articles in different languages and can be used to effectively classify Wikipedia articles written in any language to the NE classes of types person, location, and organization; and modifying beta-gamma threshold adjustment to improve overall classification quality even when many training examples are available. The features and techniques proposed in this paper are compared to previous work in the literature.

The rest of the paper is organized as follows: Section 2 provides information about the structure and feature of Wikipedia; Section 3 surveys prior work on the problem; Section 4 describes the classification approach including features and threshold adjustment algorithm; Section 5 describes the datasets used for evaluation; Section 6 presents the results of the experiments; and Section 7 concludes the paper.

2 Wikipedia Pages

Wikipedia pages have a variety of types including:

- Content pages which constitute entries in Wikipedia (as in Figure 1). Content pages typically begin with an abstract containing a brief description of the article. They may contain semi-structured data such as infoboxes and persondata, which provide factoids about concepts or entities in pages using attribute-value pairs. Persondata structures are found only in people pages. Most of the articles in Wikipedia belong to one or more category, and the categories a page belongs to are listed in the footer of the page. As in Figure 1, the entry for Alexander Pushkin belongs to categories such as "Russian Poets" and "1799 births". Content pages provide information about common concepts or named entities of type person, location, or organization (Dakka and Cucerzan, 2008). A page in Wikipedia is linked to its translations in other languages through cross language links. These links redirects user to the same Wikipedia article written in different language.
- Category pages which lists content pages that belong to a certain category. Since

categories are hierarchical, a category page lists its parent category and sub-categories below it.

- Disambiguation pages which help disambiguate content pages with the same titles. For example, a disambiguation page for "jaguar" provides links to jaguar the cat, the car, the guitar, etc.
- Redirect pages redirect users to the correct article if the name of the article entered was not exactly the same. For example, "President Obama" is redirected to "Barak Obama".

3 Related Work

This section presents some of the effort pertaining to identifying NE pages in Wikipedia and some background on SVM threshold adjustment.

3.1 Classifying Wikipedia Articles

Toral and Munoz (2006) proposed an approach to build and maintain gazetteers for NER using Wikipedia. The approach makes use of a noun hierarchy obtained from WordNet in addition to the first sentence in an article to recognize articles about NE's. A POS tagger can be used in order to improve the effectiveness of the algorithm. They reported F-measure scores of 78% and 68% for location and person classes respectively. The work in this paper relies on using the content of Wikipedia pages only.

Watanabe et al. (2007) considered the problem of tagging NE's in Wikipedia as the problem of categorizing anchor texts in articles. The novelty of their approach is in exploiting dependencies between these anchor texts, which are induced from the HTML structure of pages. They used Conditional Random Fields (CRF) for classification and achieved F-measure scores of 79.8 for persons, 72.7 for locations, and 71.6 for organizations. This approach tags only NE's referenced inside HTML anchors in articles and not Wikipedia articles themselves.

Bhole et al. (2007) and Dakka and Cucerzan (2008) used SVM classifiers to classify Wikipedia articles. Both used a bag of words approach to construct feature vectors. In Bhole et al. (2007), the feature vector was constructed over the whole text of an article. They used a linear SVM and achieved 72.6, 70.5, and 41.6 F-measure for tagging persons, locations, and organizations respectively. For a Wikipedia article, Dakka and Cucerzan (2008) used feature vectors constructed using words in the full text of the article, the first paragraph, the abstract, the values in infoboxes, and the hypertext of incoming links with surrounding words. They reported 95% and 93% F-measure for person and location respectively. Using a strictly bag of words approach does not make use of the structure of Wikipedia articles and is compared against in the evaluation.

Richman and Schone (2008) and Nothman et al. (2008) annotated Wikipedia text with NE tags to build multilingual training data for NE taggers. The approach of Richman and Schone (2008) is based on using Wikipedia category structure to classify Wikipedia titles. Identifying NE's in other languages is done using cross language links of articles or categories of articles. Nothman et al. (2008) used a bootstrapping approach with heuristics based on the head nouns of categories and the opening sentence of an article. Evaluating the system is done by training a NE tagger using the generated training data. They reported an average 92% F-measure for all NE's.

Silberer et al. (2008) presented work on the translation of English NE to 15 different languages based on Wikipedia cross-language links with a reported precision of 95%. The resulting NE's were not classified. This paper extends the work on cross language links and uses features from multilingual pages to aid classification and to enable simultaneous tagging of entities across languages.

3.2 SVM Threshold Adjustment

Support Vector Machines (SVM) is a popular classification technique that was introduced by Vapnik (Vapnik, 1995). The technique is used in text classification and proved to provide excellent performance compared to other classification techniques such as k-nearest neighbor and naïve Bayesian classifiers. As in Figure 2, SVM attempts to find a maximum margin hyperplane that separates positive and negative examples. The separating hyperplane can be described as follows: <W, X> + b = 0 or $\sum_{i=1}^{n} wi \cdot xi + b$, Where W is the normal to the hyperplane, X is an input feature vector, and b is the bias (the perpendicular distance from the origin to the hyperplane). When the number of examples for each class is not equivalent, the SVM may overfit the class that has fewer training examples. Further, the SVM training is not informed by the evaluation metric. Thus, SVM training may lead to a sub-optimal

separating hyperplane. Several techniques were proposed to rectify the problem by translating the hyperplane by only adjusting bias b, which is henceforth referred to as threshold adjustment.

Some of these techniques adjust SVM threshold during learning (Vapnik 1998; Lewis 2001), while others consider threshold adjustment as a post learning step (Shanahan and Roma, 2003). One type of the later is beta-gamma threshold adjustment algorithm (Shanahan and Roma, 2003; Zhai et al., 1998), which is a post learning algorithm that has been shown to provide significant improvements for classification tasks in which very few training examples are present such as in adaptive text filtering. Such threshold adjustment allows for the tuning of an SVM to the desired measure of goodness (ex. F1 measure). A full discussion of beta-gamma threshold adjustment is provided in the experimental setup section. In the presence of many training examples, some of the training examples are set aside as a validation set to help pick an SVM threshold. Further, multi-fold cross validation is often employed.

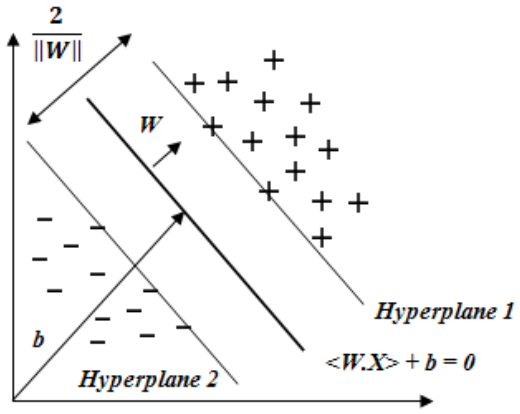

Figure 2. SVMs try to maximize the margin of separation between positive and negative examples

4 Classification Approach

Features: The classification features included content-based features such as words in page abstracts and structure-based features such category links. All the features are binary. The features are:

- Stemmed content words extracted from abstracts: an abstract for a NE may include keywords that may tell of the entity type. For example, an abstract for an NE of type person would typically include words such as "born", "pronounced", and more specific words that point to profession, role, or job (ex. president, poet, etc.).
- White space delimited attribute names from infoboxes: in the presence of infoboxes structures, the attribute names provide hints of the entity type. For example, an infobox of location may include attribute names such as "latitude", "longitude", "area", and "population".
- White space delimited words in category links for a page: category names may include keywords that would help disambiguate a NE type. For example, categories of NE of type person may include the words "births", "deaths", "people", occupation such as "poet" or "president", nationality such "American" or "Russian", etc.
- Persondata structure attributes: persondata only exist if the entity refers to a person.

The features used herein combine structural as well as content-based features from multiple languages unlike features used in the literature which were monolingual. Using multilingual features enables language independent classification of any Wikipedia article written in any language. Moreover, using primarily structural features in classification instead of the whole content of the articles allows for the effective use of multilingual pages without the need for language specific stemmers and stopword lists, the absence of which may adversely affect content based features.

Classification: Classifying Wikipedia pages was done in two steps: First training an SVM classifier; and then adjusting SVM thresholds based on beta-gamma adjustment to improve recall. Beta-gamma threshold adjustment was compared to cross-fold validation threshold adjustment. All Wikipedia articles were classified using a linear SVM. Classification was done using the Liblinear SVM package which is optimized for SVM classification problems with thousands of features (Fan et al., 2008). A variant of the beta-gamma threshold adjustment algorithm as described by (Shanahan and Roma, 2003; Zhai et al., 1998) is used to adjust the threshold of SVM. The basic steps of the algorithm are as follows:

- Divide the validation set into n folds such that each fold contains the same number of positive examples
- For each fold i,

- Classify examples in a fold and sort them in descending order based on SVM scores, where the SVM score of SVM is the perpendicular distance between an example and the separating hyperplane.
- Calculate F-measure, which is the goodness measure used in the paper, at each example.
- Determine the point of maximum F-measure and set θ_{N_i} to the SVM score at this point.
- Repeat previous steps for the set consisting of all folds other than i and set $\theta_{Max} = \theta_{N_i}$ and $\theta_{Min} = \theta_{M_i}$, where θ_{M_i} is the SVM score at the point of minimum F-measure.
- Compute $\beta_i = \frac{\theta_{N_i} - \theta_{Max}}{\theta_{Min} - \theta_{Max}}$

- $\beta = \frac{\sum \beta_i}{n}$
- The optimal threshold is obtained by interpolating between θ_{Max} and θ_{Min} obtained from the whole validation set as follows:
  $\theta_{Opt} = \alpha\, \theta_{Min} + (1 - \alpha)\, \theta_{Max}$,
  where $\alpha = \beta + (1 - \beta)e^{-\gamma M}$, M is the number of documents in the validation set, and γ is the inverse of the estimated number of documents at the point of the optimal threshold (Zhai et al., 1998). In this work, it is assigned a value that is equivalent to the inverse of the number of examples at θ_{Max}.

Since the number of training examples in Shanahan and Roma (2003) were small, n-fold cross-validation was done using the training set. In this work, the validation and training sets were non-overlapping. Further, in the work of Shanahan and Roma (2003), θ_{Min} was set to the point that yields utility = 0 as they used a filtering utility measure that can produce a utility of 0. Since no F-measure was found to equal zero in this work, minimum F-measure point was used instead.

For comparison, n-fold cross validation was used to obtain θ_{N_i} for each of the folds and then θ_{opt} as the average of all θ_{N_i}. Further, using a bag-of-words approach is used for comparison, where a feature vector in constructed based on the full text of an article.

5 Data Set

To train and test the tagging of Wikipedia pages with NE tags, a dataset of 4,936 English Wikipedia pages was developed by the authors and with split using a 60/20/20 training, validation, and testing split. The characteristics of the dataset, which is henceforth referred to as MAIN, are presented in Table 1. The English articles had links to 128 different languages, with: 16,912 articles having cross-language links; 93.3 pages on average per language; 97 languages with fewer than 100 links; with a minimum of 1 page per language (for 14 languages); and a maximum of 918 pages for French. To compare the inclusion of multilingual pages in training and testing, two variants of MAIN were used, namely: MAIN-E which has only English pages, and MAIN-EM which has English and multilingual pages from 13 languages with the most pages – Spanish, French, Finnish, Dutch, Polish, Portuguese, Italian, Norwegian, German, Danish, Hungarian, Russian, and Swedish. Other languages had too few pages. To stem text, Porter stemmer was used for English and snowball stemmers[3] were used for the other 13 languages. For all the languages, stopwords were removed. For completeness, another set was constructed to include all 128 languages to which the English pages had cross language links. This set is referred to as the MAIN-EM+ set. The authors did not have access to stemmers and stopword lists in all these languages, so simple tokenization was performed by breaking text on whitespaces and punctuation. Since many English pages don't have cross language links and most languages have too few pages, a new dataset was constructed as a subset of the aforementioned dataset such that each document in the collection has an English page with at least one cross language link to one of the 13 languages with the most pages in the bigger dataset. Table 2 details the properties of the smaller dataset, which is henceforth referred to as SUB. SUB had five variants, namely:

- SUB-E with English pages only
- SUB-EM with English and multilingual pages from the 13 languages in MAIN-EM
- SUB-M which the same as SUM-EM excluding English.
- SUB-EM+ with English pages and multilingual pages in 128 languages.
- SUB-M+ which is the same as SUB-EM+ excluding English.

The articles used in the experiments were randomly selected out of all the content articles in Wikipedia, about 3 million articles. Articles were randomly assigned to training and test sets

[3] http://snowball.tartarus.org/

and manually annotated in accordance to the CONLL – 2003 annotation guidelines[4] which are based on (Chinchor et al., 1999). Annotation was based on reading the contents of the article and then labeling it with the appropriate class. All the data, including first sentence in an article, infobox attributes, persondata attributes, and category links, were parsed from a 2010 Wikipedia XML dump.

6 Evaluation and Results

The results of classifying Wikipedia articles using SVM and threshold adjustment for MAIN-E, MAIN-EM, and MAIN-M are reported in Tables 3, 4, and 5 respectively. Tables 6, 7, 8, 9, and 10 report results for SUB-E, SUB-EM, SUB-M, SUB-EM+, and SUB-M+ respectively. In all, n is the number of cross folds used to calculate β, with n ranging between 3 and 10. The first row is the baseline scores of SVM classification without threshold adjustment. The remaining rows are the scores of SVM classification after adjusting threshold. The adjustment is performed by adding θ_{Opt} to the bias value b learned by the SVM. A t-test with 95% confidence (p-value $\leq$ 0.05) is used to determine statistical significance.

For the MAIN-E dataset, SVM threshold relaxation yielded statistically significant improvement over the baseline of using an SVM directly for location named entity. For other types of named entities improvements were not statistically significant.

Threshold adjustment led to statistically significant improvement for: all NE types for SUB-EM and SUB-EM+; for organizations for SUB-E and SUB-M+; and for locations and organization for SUB-EM. The improvements were most pronounced when recall was very low. For example, F1 measure for organization in the SUB-M dataset improved by 18 points due to a 26 point improvement in recall – though at the expense of precision.

It seems that threshold adjustment tends to benefit classification more when: using smaller training sets – as is observed when comparing the results for MAIN and SUB datasets, and when classification leads to very low recall – as indicated by organization NE for SUB datasets.

Tables 11 and 12 compare the results for the different variations of the MAIN and SUB datasets respectively. As indicated in the Tables 11 and 12, the inclusion of more and more language pages with English led to improved classification with consistent improvements in precision and recall for MAIN and consistent improvements in precision for SUB. For the SUB-M and SUB-M+ datasets, the exclusion of English led to degradation on F1 measure, with the degradation being particularly pronounced for organizations. The drop can be attributed to the loss of much valuable training examples, because there are more English pages compared to other languages. Despite the loss, proper identification of persons and locations remained high enough for many practical applications. Further, the results suggest that given more training data in the other languages, the features suggested in the paper would likely yield good classification results. Unlike the MAIN datasets, the inclusion of more languages for training and testing (from SUB-M to SUB-M+ & from SUB-EM to SUB-EM+) did not yield any improvements except for location and organization types from SUB-EM to SUB-EM+. This requires more investigation.

Tables 13 and 14 report the results of using term frequency representation of the entire page as features – a bag of words (BOWs)– as in Bhole et al. (2007). Using semi-structured data as classification features is better than using BOW representation. This could be due to the smaller number of features of higher value. In the BOW results with multilingual page inclusion, except for location NE type only in the SUB dataset, the use of term frequencies of multilingual words hurt F1-measure for the SUB and MAIN datasets. This can be attributed to the increased sparseness of the training and test data.

7 Conclusions

This paper presented a language independent method for identifying multilingual Wikipedia articles referring to named entities. An SVM was trained using multilingual features that make use of unstructured and semi-structured portions of Wikipedia articles. It was shown that using multilingual features was better than using features obtained from English articles only. Multilingual features can be used in classifying multilingual articles and is particularly useful for languages other than English, where fewer useful features are present. The number of Infobox properties and category links in English MAIN was 32,262 and 9,221 respectively, while in German there are 4,618 properties and 1,657 category links. These numbers are even lower in all other languages.

[4] http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt

	Training	Validation	Test
Person	822	300	251
Locations	676	221	266
Organizations	313	113	110
Non	1085	366	414
Total	2896	1000	1040

Table 1. Characteristics of MAIN dataset: the number of Wikipedia pages in the dataset

	Training	Validation	Test
Person	332	128	95
Locations	360	115	144
Organizations	102	30	42
Non	435	150	184
Total	1229	423	465

Table 2. Characteristics of SUB dataset: the number of Wikipedia pages in the dataset

cross folds	**Person**			**Location**			**Organization**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	98.7	90.4	94.4	94.6	85.7	89.9	**90**	**73.6**	**81.0**
n = 3	97.9	92.0	94.9	94.4	89.0	91.6	87.2	74.5	80.4
n = 4	96.7	92.0	94.3	94.4	89.0	91.6	87.2	74.5	80.4
n = 5	96.6	92.0	94.3	94.4	89.0	91.6	80.0	76.4	78.0
n =6	96.7	92.4	94.5	94.4	89.4	91.9	85.6	75.4	80.2
n =7	**96.7**	**92.8**	**94.7**	94.4	89.4	91.9	85.6	75.4	80.2
n =8	96.7	92.8	94.7	***94.0***	***90.6***	***92.3***	80.0	76.4	78.0
n =9	95.2	94.0	94.6	94.0	89.8	91.9	80.8	76.4	78.5
n = 10	94.8	94.0	94.4	94.0	90.6	92.3	77.9	80.0	78.9

Table 3. Results for MAIN-E: Best F1 bolded and italicized if significantly better than baseline.

cross folds	**Person**			**Location**			**Organization**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	99.1	91.6	95.2	94.7	87.2	90.8	91.0	73.6	81.4
n = 3	99.7	91.2	95.2	94.7	87.2	90.8	90.1	74.5	81.6
n = 4	99.1	91.6	95.2	94.7	87.9	91.2	**90.2**	**75.4**	**82.2**
n = 5	**99.1**	**92.4**	**95.7**	94.4	89	91.6	86.4	75.4	80.6
n =6	98.3	92.4	95.3	94.7	87.9	91.2	87.4	75.4	81.0
n =7	98.3	92.4	95.3	93.7	90.2	91.9	82.3	76.4	79.2
n =8	98.3	92.8	95.5	93.7	90.2	91.9	85.7	76.4	80.8
n =9	98.3	92.8	95.5	**92.4**	**92.4**	**92.4**	82.3	76.4	79.2
n = 10	97.9	92.8	95.3	92.8	92.1	92.4	82.3	76.4	79.2

Table 4. Results for MAIN-EM: Best F1 bolded and italicized if significantly better than baseline.

cross folds	**Person**			**Location**			**Organization**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	**99.6**	**92.0**	**95.6**	95.0	87.2	90.9	91.0	73.6	81.4
n = 3	98.3	92.4	95.3	94.3	88.3	91.2	**91.0**	**74.5**	**82.0**
n = 4	98.3	92.8	95.5	93.7	90.2	91.9	91.0	74.5	82.0
n = 5	98.3	92.8	95.5	**93.8**	**90.9**	**92.3**	89.2	75.4	81.8
n =6	97.9	93.2	95.5	93.0	91.3	92.2	88.3	75.4	81.4
n =7	95.5	93.6	94.6	93.4	90.9	92.2	87.4	75.4	81.0
n =8	95.5	93.6	94.6	91.8	92.8	92.3	85.7	76.4	80.8
n =9	95.9	93.2	94.5	92.0	92.0	92.0	84.0	76.4	80.0
n = 10	95.2	94.8	95.0	91.7	92.0	91.9	85.7	76.4	80.8

Table 5. Results for MAIN-EM+: Best F1 bolded and italicized if significantly better than baseline.

The effect of using SVM and beta-gamma threshold adjustment algorithm to improve recognizing NE's in Wikipedia was also demonstrated. The algorithm was shown to improve scores of location NE's particularly. The appropriate number of folds was found to be 8 using our dataset. Finally, the results suggest that the use of semi-structured data as classification features is significantly better than the using unstructured data only or BOWs. The paper also showed that the use of multilingual features with BOWs was not very useful.

For future work, the proposed technique can be used to create large sets of tagged Wikipedia pages in a variety of languages to aid in building parallel lists of named entities that can be used to improve MT and in training transliterator engines. Further, this work can help in building resources such gazetteers and tagged NE data in many languages for the rapid development of NE taggers in general text. Wikipedia has the advantage of covering many topics beyond those that are typically covered in news articles.

cross folds	**Person**			**Location**			**Organization**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	100	92.6	96.2	**98.5**	**91.7**	**95**	87.0	49.0	62.5
n = 3	100	92.6	96.2	97.8	91.7	94.6	84	51.2	63.6
n = 4	100	92.6	96.2	97.8	91.7	94.6	85.2	56	67.7
n = 5	100	93.7	96.7	96.4	92.4	94.3	***87.0***	***65.8***	***75.0***
n =6	100	93.7	96.7	95.7	93.7	94.7	85.7	58.5	69.6
n =7	100	93.7	96.7	95.7	93.7	94.7	87.0	65.8	75.0
n =8	100	93.7	96.7	95.7	93.7	94.7	87.0	65.8	75.0
n =9	**100**	**94.7**	**97.3**	95.0	94.4	94.8	87.0	65.8	75.0
n = 10	100	94.7	97.3	95.0	94.4	94.8	87.0	65.8	75.0

Table 6. Results for SUB-E: Best F1 bolded and italicized if significantly better than baseline.

cross folds	**Person**			**Location**			**Organization**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	100	91.6	95.6	99.2	88.9	93.8	100	46.3	63.3
n = 3	98.9	92.6	95.6	99.2	88.2	93.3	***100***	***53.6***	***69.8***
n = 4	98.9	92.6	95.6	98.5	91.7	95.0	92.0	56.0	69.7
n = 5	98.9	92.6	95.6	99.2	88.9	93.8	92.0	56.0	69.7
n =6	98.9	92.6	95.6	***98.5***	***93.7***	***96.0***	92.0	56.0	69.7
n =7	99.0	92.6	95.6	98.5	93.7	96.0	92.0	56.0	69.7
n =8	99.0	92.6	95.6	97.8	93.7	95.7	92.0	56.0	69.7
n =9	***98.9***	***95.7***	***97.3***	95.2	95.8	95.5	92.0	56.0	69.7
n = 10	98.9	95.7	97.3	93.2	96.5	94.9	92.0	56.0	69.7

Table 7. Results for SUB-EM: Best F1 bolded and italicized if significantly better than baseline.

References

Babych, Bogdan, and Hartley, Anthony (2003). *Improving Machine Translation quality with automatic Named Entity recognition*. 7th Int. EAMT workshop on MT and other lang. tech. tools -- EACL'03, Budapest, Hungary.

Bhole, Abhijit, Fortuna, Blaz, Grobelnik, Marko, and Mladenic, Dunja. (2007). *Extracting Named Entities and Relating Them over Time Based on Wikipedia*. Informatica (Slovenia), 31, 463-468.

Chinchor, Nancy, Brown, Erica, Ferro, Lisa, and Robinson, Patty. (1999). 1999 Named Entity Recognition Task Definition: MITRE.

Dakka, Wisam., and Cucerzan, Silviu. (2008). *Augmenting Wikipedia with Named Entity Tags.* 3rd IJCNLP, Hyderabad, India.

Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen. (2008). *LIBLINEAR: A Library for Large Linear Classication.* Journal of Machine Learning Research 9, 1871-1874.

Nothman, Joel, Curran, James R., and Murphy, Tara. (2008). *Transforming Wikipedia into Named Entity Training Data.* Australian Lang. Tech. Workshop.

Richman, Alexander E., and Schone, Patrick. (2008, June). *Mining Wiki Resources for Multilingual Named Entity Recognition.* ACL-08: HLT, Columbus, Ohio.

Shanahan, James G., and Roma, Norbert. (2003). *Boosting support vector machines for text classification through parameter-free threshold relaxation*. CIKM'03. New Orleans, LA, US

Silberer, Carina, Wentland, Wolodja, Knopp, Johannes, and Hartung, Matthias. (2008). *Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration.* LREC'08, Marrakech, Morocco.

Toral, Antonio, and Mu˜noz, Rafael (2006). *A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia*, EACL-2008. Italy.

Vapnik, Vladimir N. (1995). *The nature of statistical learning theory*: Springer-Verlag New York, Inc.

Watanabe, Yotaro, Asahara, Masayuki, and Matsumoto, Yuji. (2007). *A Graph-Based Approach to Named Entity Categorization in Wikipedia using Conditional Random Fields.* EMNLP-CoNLL, Prague, Czech Republic

Zhai, Chengxiang, Jansen, Peter, Stoica, Emilia, Grot, Norbert, and Evans, David A. (1998). *Threshold Calibration in CLARIT Adaptive Filtering.* TREC-7, Gaithersburg, Maryland, US.

cross folds	**Person**			**Location**			**Organization**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	100	90.5	95	99.2	90.3	94.5	100	47.6	64.5
n = 3	98.9	92.6	95.6	98.5	91.7	95	100	47.6	64.5
n = 4	98.9	92.6	95.6	98.5	91	94.6	96	57	71.6
n = 5	98.9	92.6	95.6	**98.5**	**92.4**	**95.3**	96	57	71.6
n =6	98.9	92.6	95.6	95.8	94.4	95	96	57	71.6
n =7	98.9	92.6	95.6	97	93	95	100	54.8	70.8
n =8	98.9	92.6	95.6	95.8	94.4	95	92.6	59.5	72.5
n =9	98.8	93.7	96.2	95	95	95	***96***	***59.5***	***73.5***
n = 10	***98.9***	***94.7***	***96.8***	94.5	95.8	95.2	92.6	59.5	72.5

Table 8. Results for SUB-EM+: Best F1 bolded and italicized if significantly better than baseline.

cross folds	**Person**			**Location**			**Organization**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	97.4	77.9	86.5	97.4	78.5	86.9	100	21.9	36.0
n = 3	97.4	78.9	87.2	96.7	80.5	87.9	100	24.4	39.2
n = 4	97.5	82.0	89.0	95.3	84.0	89.3	71.4	36.6	48.4
n = 5	97.5	82.0	89.0	94.6	84.7	89.4	100	24.4	39.2
n =6	**96.3**	**83.0**	**89.3**	94.6	84.7	89.4	100	24.4	39.2
n =7	95.2	83.0	88.8	***94.6***	***86.0***	***90.2***	77.8	34	47.4
n =8	97.5	83.0	89.8	91.8	86.0	88.9	70.8	41.5	52.3
n =9	95.2	84.2	89.4	94.6	86.0	90.0	61.3	46.3	52.8
n = 10	91.2	87.4	89.2	64.9	96.5	77.6	***60.6***	***48.8***	***54.0***

Table 9. Results for SUB-M: Best F1 bolded and italicized if significantly better than baseline.

cross folds	**Person**			**Location**			**Organization**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	97.3	76.8	85.9	97.4	77	86	100	19	32
n = 3	97.4	77.9	86.5	95	81.2	87.6	100	23.8	38.5
n = 4	97.4	77.9	86.5	95.8	78.5	86.2	91.7	26.2	40.7
n = 5	**97.4**	**80**	**87.9**	95.9	80.5	87.5	86.7	30.9	45.6
n =6	96.2	80	87.3	91	84.7	87.8	91.7	26.2	40.7
n =7	96.2	80	87.3	**92.4**	**84.7**	**88.4**	91.7	26.2	40.7
n =8	95	80	86.8	75	93.7	83.3	***79.2***	***45.2***	***57.6***
n =9	92.8	82	87	89.3	86.8	88	79.2	45.2	57.6
n = 10	90.9	84.2	87.4	65.9	97.9	78.8	79.2	45.2	57.6

Table 10. Results for SUB-M+: Best F1 bolded and italicized if significantly better than baseline.

MAIN	F1-measure		
	E	EM	EM+
Person	94.4	95.2	***95.6***
Location	89.9	90.8	**90.9**
Organization	81.0	81.4	**81.4**

Table 11. Comparing results for MAIN-{E, EM, and EM+}: Best F1 bolded and italicized if significantly better than MAIN-E

SUB	F1-Measure				
	E	EM	M	EM+	M+
Person	**96.2**	95.6	86.5	95	85.9
Location	**95**	93.8	86.9	94.5	86
Organization	62.5	63.3	36.0	**64.5**	32

Table 12. Comparing results for SUB-{E, EM, M, EM+, and M+}: Best F1 bolded

MAIN	F1-measure		
	E	EM	EM+
Person	**86.8**	85.0	84.5
Location	**87.4**	85.8	85.5
Organization	**58.0**	51.8	53.4

Table 13. Comparing results of BOWs for MAIN-{E, EM, and EM+}: Best F1 bolded

SUB	F-Measure				
	E	EM	M	EM+	M+
Person	**82.0**	80.6	68.0	79.3	61.9
Location	88.5	**90.7**	83.8	90.0	82.3
Organization	**35.6**	22.6	21.4	33.3	22.6

Table 14. Comparing results of BOWs for SUB-{E, EM, M, EM+, and M+}: Best F1 bolded