Multilingual Named Entity Recognition using Parallel Data and Metadata from Wikipedia

Sungchul Kim*
POSTECH
Pohang, South Korea
subright@postech.ac.kr

Kristina Toutanova
Microsoft Research
Redmond, WA 98502
kristout@microsoft.com

Hwanjo Yu
POSTECH
Pohang, South Korea
hwanjoyu@postech.ac.kr

Abstract

In this paper we propose a method to automatically label multi-lingual data with named entity tags. We build on prior work utilizing Wikipedia metadata and show how to effectively combine the weak annotations stemming from Wikipedia metadata with information obtained through English-foreign language parallel Wikipedia sentences. The combination is achieved using a novel semi-CRF model for foreign sentence tagging in the context of a parallel English sentence. The model outperforms both standard annotation projection methods and methods based solely on Wikipedia metadata.

1 Introduction

Named Entity Recognition (NER) is a frequently needed technology in NLP applications. State-of-the-art statistical models for NER typically require a large amount of training data and linguistic expertise to be sufficiently accurate, which makes it nearly impossible to build high-accuracy models for a large number of languages.

Recently, there have been two lines of work which have offered hope for creating NER analyzers in many languages. The first has been to devise an algorithm to tag foreign language entities using metadata from the semi-structured Wikipedia repository: inter-wiki links, article categories, and cross-language links (Richman and Schone, 2008). The second has been to use parallel English-foreign language data, a high-quality NER tagger for English, and projected annotations for the foreign language (Yarowsky et al., 2001; Das and Petrov, 2011). Parallel data has also been used to improve existing monolingual taggers or other analyzers in two languages (Burkett et al., 2010a; Burkett et al., 2010b).

The goal of this work is to create high-accuracy NER annotated data for foreign languages. Here we combine elements of both Wikipedia metadata-based approaches and projection-based approaches, making use of parallel sentences extracted from Wikipedia. We propose a statistical model which can combine the two types of information. Similarly to the joint model of Burkett et al. (2010a), our model can incorporate both monolingual and bilingual features in a log-linear framework. The advantage of our model is that it is much more efficient as it does not require summing over matchings of source and target entities. It is a conditional model for target sentence annotation given an aligned English source sentence, where the English sentence is used only as a source of features. Exact inference is performed using standard semi-markov CRF model inference techniques (Sarawagi and Cohen, 2004).

Our results show that the semi-CRF model improves on the performance of projection models by more than 10 points in F-measure, and that we can achieve tagging F-measure of over 91 using a very small number of annotated sentence pairs.

The paper is organized as follows: We first describe the datasets and task setting in Section 2. Next, we present our two baseline methods: A Wikipedia metadata-based tagger and a cross-lingual projection tagger in Sections 3 and 4, respectively. We present our direct semi-CRF tagging model in Section 5.

2 Data and task

As a case study, we focus on two very different foreign languages: Korean and Bulgarian. The English and foreign language sentences that comprise our training and test data are extracted from Wikipedia (http://www.wikipedia.org). Currently there are more than 3.8 million articles in the English Wikipedia, 125,000 in the Bulgarian Wikipedia, and 131,000 in the Korean Wikipedia.

*This research was conducted during the author's internship at Microsoft Research

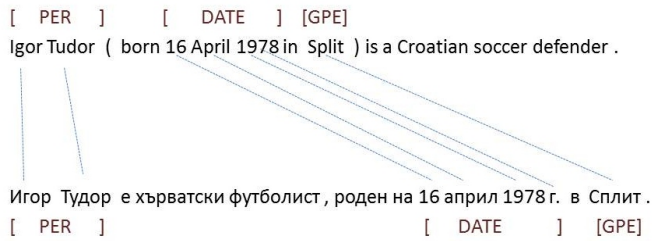

Figure 1: A parallel sentence-pair showing gold-standard NE labels and word alignments.

To create our dataset, we followed Smith et al. (2010) to find parallel-foreign sentences using comparable documents linked by inter-wiki links. The approach uses a small amount of manually annotated article-pairs to train a document-level CRF model for parallel sentence extraction. A total of 13,410 English-Bulgarian and 8,832 English-Korean sentence pairs were extracted.

Of these, we manually annotated 91 English-Bulgarian and 79 English-Korean sentence pairs with source and target named entities as well as word-alignment links among named entities in the two languages. Figure 1 illustrates a Bulgarian-English sentence pair with alignment.

The named entity annotation scheme followed has the labels GPE (Geopolitical entity), PER (Person), ORG (Organization), and DATE. It is based on the MUC-7 annotation guidelines, and GPE is synonymous with Location. The annotation process was not as rigorous as one might hope, due to lack of resources. The English-Bulgarian and English-Korean datasets were labeled by one annotator each and then annotations on the English sentences were double-checked by the other annotator. Disagreements were rare and were resolved after discussion.

The task we evaluate on is tagging of foreign language sentences. We measure performance by labeled precision, recall, and F-measure. We give partial credit if entities partially overlap on their span of words and match on their labels.

Table 1 shows the total number of English, Bulgarian and Korean entities and the percentage of entities that were manually aligned to an entity of the same type in the other language. The data sizes are fairly small as the data is used only to train models with very few coarse-grained features and for evaluation. These datasets are available at http://research.microsoft.com/en-us/people/kristout/nerwikidownload.aspx.

Language	Entities	Aligned %
English	342	93.9%
Bulgarian	344	93.3%
English	414	88.4%
Korean	423	86.5%

Table 1: English-Bulgarian and English-Korean data characteristics.

As we can see, less than 100% of entities have parallels in the other language. This is due to two phenomena: one is that the parallel sentences sometimes contain different amounts of information and one language might use more detail than the other. The other is that the same information might be expressed using a named entity in one language, and using a non-entity phrase in the other language (e.g. “He is from Bulgaria” versus “He is Bulgarian”). Both of these causes of divergence are much more common in the English-Korean dataset than in the English-Bulgarian one.

3 Wiki-based tagger: annotating sentences based on Wikipedia metadata

We followed the approach of Richman and Schone (2008) to derive named entity annotations of both English and foreign phrases in Wikipedia, using Wikipedia metadata. The following sources of information were used from Wikipedia: *category* annotations on English documents, *article links* which link from phrases in an article to another article in the same language, and *interwiki links* which link

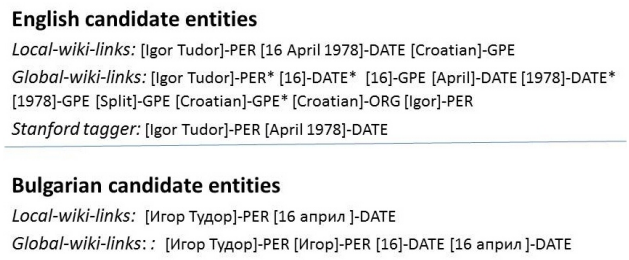

Figure 2: Candidate NEs for the English and Bulgarian sentences according to baseline taggers.

from articles in one language to comparable (semantically equivalent) articles in the other language. In addition to the Wikipedia-derived resources, the approach requires a manually specified map from English category key-phrases to NE tags, but does not require expert knowledge for any non-English language. We implemented the main ideas of the approach but some implementation details may differ.

To tag English language phrases, we first derived named entity categorizations of English article titles, by assigning a tag based on the article's category information. The category-to-NE map used for the assignment is a small manually specified map from phrases appearing in category titles to NE tags. For example, if an article has categories "People by", "People from", "Surnames" etc., it is classified as PER. Looking at the example in Figure 1, the article with title "Igor Tudor" is classified as PER because one of its categories is "Living people". The full map we use is taken from the paper (Richman and Schone, 2008).

Using the article-level annotations and *article links* we define a local English wiki-based tagger and a global English wiki-based tagger, which will be described in detail next.

Local English Wiki-based tagger. This Wiki-based tagger tags phrases in an English article based on the *article links* from these phrases to NE-tagged articles. For example, suppose that the phrase "Split" in the article with title "Igor Tudor" is linked to the article with title "Split", which is classified as GPE. Thus the local English Wiki-based tagger can tag this phrase as GPE. If, within the same article, the phrase "Split" occurs again, it can be tagged again even if it is not linked to a tagged article (this is the one sense per document assumption). Additionally, the tagger tags English phrases as DATE if they match a set of manually specified regular expressions. As a filter, phrases that do not contain a capitalized word or a number are not tagged with NE tags.

Global English Wiki-based tagger. This tagger tags phrases with NE tags if these phrases have ever been linked to a categorized article (the most frequent label is used). For example, if "Split" does not have a link anywhere in the current article, but has been linked to the GPE-labeled article with title "Split" in another article, it will still be tagged as GPE. We also apply a local+global Wiki-tagger, which tags entities according to the local Wiki-tagger and additionally tags any non-conflicting entities according to the global tagger.

Local foreign Wiki-based tagger. The idea is the same as for the local English tagger, with the difference that we first assign NE tags to foreign language articles by using the NE tags assigned to English articles to which they are connected with inter-wiki links. Because we do not have maps from category phrases to NE tags for foreign languages, using inter-wiki links is a way to transfer this knowledge to the foreign languages. After we have categorized foreign language articles we follow the same algorithm as for the local English Wiki-based tagger. For Bulgarian we also filtered out entities based on capitalization and numbers, but did not do that for Korean as it has no concept of capitalization.

Global foreign Wiki-based tagger The global and local+global taggers are analogous, using the categorization of foreign articles as above.

Figure 2 shows the tags assigned to English and Bulgarian strings according to the local and global Wiki-based taggers. The global Wiki-based tagger could assign multiple labels to the same string (corresponding to different senses in different occurrences). In case of multiple possible labels, the most frequent one is denoted by * in the Figure. The Figure also shows the results of the Stanford NER tagger for English (Finkel et al., 2005) (we used the MUC-7 classifier).

Table 2 reports the performance of the local (L Wiki-tagger), local+global (LG Wiki tagger) and the Stanford tagger. We can see that the local Wiki taggers have higher precision but lower recall than the local+global Wiki taggers. The local+global taggers

Language	L Wiki-tagger			LG Wiki-tagger			Stanford Tagger		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
English	92.8	75.1	83.0	79.7	89.5	**84.3**	86.5	77.5	81.7
Bulgarian	94.1	48.7	64.2	86.8	79.9	**83.2**			
English	92.6	75.6	83.2	84.1	86.7	**85.4**	82.2	71.9	76.7
Korean	89.5	57.3	**69.9**	43.2	78.0	55.6			

Table 2: English-Bulgarian and English-Korean Wiki-based tagger performance.

are overall best for English and Bulgarian. The local tagger is best for Korean, as the precision suffers too much due to the global tagger. This is perhaps due in part to the absence of the capitalization filter for Korean which improved precision for Bulgarian and English. The Stanford tagger is worse than the Wiki-based tagger, but it is different enough that it contributes useful information to the task.

4 Projection Model

From Table 2 we can see that the English Wiki-based taggers are better than the Bulgarian and Korean ones, which is due to the abundance and completeness of English data in Wikipedia. In such circumstances, previous research has shown that one can project annotations from English to the more resource-poor language (Yarowsky et al., 2001). Here we follow the approach of Feng et al. (2004) to train a log-linear model for projection.

Note that the Wiki-based taggers do not require training data and can be applied to any sentences from Wikipedia articles. The projection model described in this section and the Semi-CRF model described in Section 5 are trained using annotated data. They can be applied to tag foreign sentences in English-foreign sentence pairs extracted from Wikipedia.

The task of projection is re-cast as a ranking task, where for each source entity S_i, we rank all possible candidate target entity spans T_j and select the best span as corresponding to this source entity. Each target span is labeled with the NE label of the corresponding source entity. The probability distribution over target spans T_j for a given source entity S_i is defined as follows:

$$p(S_i|T_j) = \frac{\exp(\lambda f(S_i, T_j))}{\sum_{j'} \exp(\lambda \mathbf{f}(S_i, T'_j))}$$

where λ is a parameter vector, and $\mathbf{f}(S_i, T_j)$ is a feature vector for the candidate entity pair.

From this formulation we can see that a fixed set of English source entities S_i is required as input. The model projects these entities to corresponding foreign entities. We train and evaluate the projection model using 10-fold cross-validation on the dataset from Table 1. For training, we use the human-annotated gold English entities and the manually-specified entity alignments to derive corresponding target entities. At test time we use the local+global Wiki-based tagger to define the English entities and we don't use the manually annotated alignments.

4.1 Features

We present the features for this model in a lot of detail since analogous feature types are also used in our final direct semi-CRF model. The features are grouped into four categories.

Word alignment features
We exploit a feature set based on HMM word alignments in both directions (Och and Ney, 2000). To define the features we make use of the posterior alignment link probabilities as well as the most likely (Viterbi) alignments. The posterior probabilities are the probabilities of links in both directions given the source and target sentences: $P(a_i = j|s, t)$ and $P(a_j = i|s, t)$.

If a source entity consists of positions $i_1, \ldots, i_m$ and a potential corresponding target entity consists of positions $j_1, \ldots, j_n$, the word-alignment derived features are:

- Probability that each word from one of the entities is aligned to a word from the other entity, estimated as:

  $\prod_{i \in i_1 \ldots i_m} \sum_{j \in j_1 \ldots j_n} P(a_i = j|s, t)$ We use an analogous estimate for the probability in the other direction.

- Sum of posterior probabilities of links from words inside one entity to words outside another entity $\sum_{i \in i_1 \ldots i_m}(1 - \sum_{j \in j_1 \ldots j_n} P(a_i = j|s,t))$. Probabilities from the other HMM direction are estimated analogously.

- Indicator feature for whether the source and target entity can be extracted as a phrase pair according to the combined Viterbi alignments (grow-diag-final) and the standard phrase extraction heuristic (Koehn et al., 2003).

Phonetic similarity features
These features measure the similarity between a source and target entity based on pronunciation. We utilize a transliteration model (Cherry and Suzuki, 2009), trained from pairs of English person names and corresponding foreign language names, extracted from Wikipedia. The transliteration model can return an n-best list of transliterations of a foreign string, together with scores. For example the top 3 transliterations in English of the Bulgarian equivalent of "Igor Tudor" from Figure 1 are *Igor Twoodor*, *Igor Twoodore*, and *Igore Twoodore*.

We estimate phonetic similarity between a source and target entity by computing Levenshtein and other distance metrics between the source entity and the closest transliteration of the target (out of a 10-best list of transliterations). We use normalized and un-normalized Levenshtein distance. We also use a BLEU-type measure which estimates character n-gram overlap.

Position/Length features
These report relative length and position of the English and foreign entity following (Feng et al., 2004).

Wiki-based tagger features
These features look at the degree of match between the source and target entities based on the tags assigned to them by the local and global Wiki-taggers for English and the foreign language, and by the Stanford tagger for English. These are indicator features separate for the different source-target tagger combinations, looking at whether the taggers agree in their assignments to the candidate entities.

4.2 Model Evaluation

We evaluate the tagging F-measure for projection models on the English-Bulgarian and English-Korean datasets. 10-fold cross-validation was used to estimate model performance. The foreign language NE F-measure is reported in Table 3. The best Wiki-based tagger performance is shown on the last line as a baseline (repeated from Table 2).

We present a detailed evaluation of the model to gain understanding of the strengths and limitations of the projection approach and to motivate our direct semi-CRF model. To give an estimate of the upper bound on performance for the projection model, we first present two oracles. The goal of the oracles it to estimate the impact of two sources of error for the projection model: the first is the error in detecting English entities, and the second is the error in determining the corresponding foreign entity for a given English entity.

The first oracle ORACLE1 has access to the gold-standard English entities and gold-standard word alignments among English and foreign words. For each source entity, ORACLE1 selects the longest foreign language sequence of words that could be extracted in a phrase pair coupled with the source entity word sequence (according the standard phrase extraction heuristic (Koehn et al., 2003)), and labels it with the label of the source entity. Note that the word alignments do not uniquely identify the corresponding foreign phrase for each English phrase and some error is possible due to this. The performance of this oracle is closely related to the percentage of linked source-target entities reported in Table 1. The second oracle ORACLE2 provides the performance of the projection model when gold-standard source entities are known, but the corresponding target entities still have to be determined by the projection model (gold-standard alignments are not known). In other words, ORACLE2 is the projection model with all features, where in the test set we provide the gold standard English entities as input. The performance of ORACLE2 is determined by the error in automatic word alignment and in determining phonetic correspondence. As we can see the drop due to this error is very large, especially on Korean, where performance drops from 90.0 to 81.9 F-measure.

The next section in the Table presents the perfor-

Method	English-Bulgarian			English-Korean		
	Prec	Rec	F1	Prec	Rec	F1
ORACLE1	98.3	92.9	95.5	95.5	85.1	90.0
ORACLE2	96.7	86.3	91.2	90.5	74.7	81.9
PM-WF	71.7	80.0	75.7	85.1	72.2	78.1
PM+WF	73.6	81.3	77.2	87.6	74.9	**80.8**
Wiki-tagger	86.8	79.9	**83.2**	89.5	57.3	69.9

Table 3: English-Bulgarian and English-Korean Projection tagger performance.

mance of non-oracle projection models, which do not have access to any manually labeled information. The local+global Wiki-based tagger is used to define English entities, and only automatically derived alignment information is used. PM+WF is the projection model using all features. The line above, PM-WF represents the projection model without the Wiki-tagger derived features, and is included to show that the gain from using these features is substantial. The difference in accuracy between the projection model and ORACLE2 is very large, and is due to the error of the Wiki-based English taggers. The drop for Bulgarian is so large that the best projection model PM+WF does not reach the performance of 83.2 achieved by the baseline Wiki-based tagger. When source entities are assigned with error for this language pair, projecting entity annotations from the source is not better than using the target Wiki-based annotations directly. For Korean while the trend in model performance is similar as oracle information is removed, the projection model achieves substantially better performance (80.8 vs 69.9) due to the much larger difference in performance between the English and Korean Wiki-based taggers.

The drawback of the projection model is that it determines target entities only by assigning the best candidate for each source entity. It cannot create target entities that do not correspond to source entities, it is not able to take into account multiple conflicting source NE taggers as sources of information, and it does not make use of target sentence context and entity consistency constraints. To address these shortcomings we propose a direct semi-CRF model, described in the next section.

5 Semi-CRF Model

Semi-Markov conditional random fields (semi-CRFs) are a generalization of CRFs. They assign labels to segments of an input sequence $\mathbf{x}$, rather than to individual elements x_i and features can be defined on complete segments. We apply Semi-CRFs to learn a NE tagger for labeling foreign sentences in the context of corresponding source sentences with existing NE annotations.

The semi-CRF defines a distribution over foreign sentence labeled segmentations (where the segments are named entities with their labels, or segments of length one with label "NONE"). To formally define the distribution, we introduce some notation following Sarawagi and Cohen (2005):

Let $\mathbf{s} = \langle s_1, \ldots, s_p \rangle$ denote a segmentation of the foreign sentence $\mathbf{x}$, where a segment $s_j = \langle t_j, u_j, y_j \rangle$ is determined by its start position t_j, end position u_j, and label y_j. Features are defined on segments and adjacent segment labels. In our application, we only use features on segments. The features on segments can also use information from the corresponding English sentence $\mathbf{e}$ along with external annotations on the sentence pair $\mathbf{A}$.

The feature vector for each segment can be denoted by $F(j, \mathbf{s}, \mathbf{x}, \mathbf{e}, \mathbf{A})$ and the weight vector for features by $\mathbf{w}$. The probability of a segmentation is then defined as:

$$P(\mathbf{s}|\mathbf{x}, \mathbf{e}, \mathbf{A}) = \frac{\sum_j \exp \mathbf{w}' F(j, \mathbf{s}, \mathbf{x}, \mathbf{e}, \mathbf{A})}{Z(\mathbf{x}, \mathbf{e}, \mathbf{A})}$$

In the equation above Z represents a normalizer summing over valid segmentations.

5.1 Features

We use both boolean and real-valued features in the semi-CRF model. Example features and their values are given in Table 4. The features are the ones that fire on the segment of length 1 containing the Bulgarian equivalent of the word "Split" and labeled with label GPE (t_j=13,u_j=13,y_j=GPE), from the English-Bulgarian sentence pair in Figure 1.

The features look at the English and foreign sentence as well as external annotations **A**. Note that the semi-CRF model formulation does not require a fixed labeling of the English sentence. Different and possibly conflicting NE tags for candidate English and foreign sentence substrings according to the Wiki-based taggers and the Stanford tagger are specified as one type of external annotations (see Figure 2). Another annotation type is derived from HMM-based word alignments and the transliteration model described in Section 4. They provide two kinds of alignment links between English and foreign tokens: one based on the HMM-word alignments (posterior probability of the link in both directions), and another based on different character-based distance metrics between transliterations of foreign words and English words. The transliteration model and distance metrics were described in Section 4 as well. For the example Bulgarian correspondent of "Split" in the figure, the English "Split" is linked to it according to both the forward and backward HMM, and according to two out of the three transliteration distance measures. A third annotation type is automatically derived links between foreign candidate entity strings (sequences of tokens) and best corresponding English candidate entities. The candidate English entities are defined by the union of entities proposed by the Wiki-based taggers and the Stanford tagger. Note that these English candidate entities can be overlapping and inconsistent without harming the model. We link foreign candidate segments with English candidate entities based on the projection model described in Section 4 and trained on the same data. The projection model scores every source-target entity pair and selects the best source for each target candidate entity. For our example target segment, the corresponding source candidate entity is "Split", labeled GPE by the local+global Wiki-tagger and by the global Wiki-tagger.

The features are grouped into three categories:

Group 1. Foreign Wiki-based tagger features. These features look at target segments and extract indicators of whether the label of the segment agrees with the label assigned by the local, global, and/or local+global wiki tagger. For the example segment from the sentence in Figure 1, since neither the local nor global tagger have assigned a label GPE, the first three features have value zero. In addition to tags on the whole segment, we look at tag combinations for individual words within the segment as well as two words to the left and right outside the segment. In the first section in Table 4 we can see several feature types and and their values for our example.

Group 2. Foreign surface-based features. These features look at orthographic properties of the words and distinguish several word types. The types are based on capitalization and also distinguish numbers and punctuation. In addition, we make use of word-clusters generated by JCluster. [1]

We look at properties of the individual words as well as the concatenation for all words in the segment. In addition, there are features for words two words to the left and two words to the right outside the segment. The second section in the Table shows several features of this type with their values.

Group 3. Label match between English and aligned foreign entities. These features look at the linked English segment for the candidate target segment and compare the tags assigned to the English segment by the different English taggers to the candidate target label. In addition to segment-level comparisons, they also look at tag assignments for individual source tokens linked to the individual target tokens (by word alignment and transliteration links). The last section in the Table contains sample features with their values. The feature SOURCE-E-WIKI-TAG-MATCH looks at whether the corresponding source entity has the same local+global Wiki-tagger assigned tag as the candidate target entity. The next two features look at the Stanford tagger and the global Wiki-tagger. The real-valued features like SCORE-SOURCE-E-WIKI-TAG-MATCH return the score of the matching between the source and target candidate entities (according to the projection model), if the labels match. In this way, more confident matchings can impact the target tags more than less confident ones.

5.2 Experimental results

Our main results are listed in Table 5. We perform 10-fold cross-validation as in the projection experiments. The best Wiki-based and projection models are listed as baselines at the bottom of the table.

[1] Software distributed by Joshua Goodman http://research.microsoft.com/en-us/downloads/0183a49d-c86c-4d80-aa0d-53c97ba7350a/default.aspx.

Method	English-Bulgarian			English-Korean		
	Prec	Rec	F1	Prec	Rec	F1
MONO	86.7	79.4	82.9	89.1	57.1	69.6
BI	90.1	83.3	86.6	88.6	79.8	84.0
MONO-ALL	94.7	86.2	90.3	90.2	84.3	87.2
BI-ALL-WT	95.7	87.6	91.5	92.4	87.6	89.9
BI-ALL	96.4	89.4	**92.8**	94.7	87.9	**91.2**
Wiki-tagger	86.8	79.9	**83.2**	89.5	57.3	69.9
PM+WF	73.6	81.3	77.2	87.6	74.9	**80.8**

Table 5: English-Bulgarian and English-Korean semi-CRF tagger performance.

Feature Description	Example Value
WIKI-TAG-MATCH	0
WIKI-GLOBAL-TAG-MATCH	0
WIKIGLOBAL-POSSIBLE-TAG	0
WIKI-TAG&LABEL	NONE&GPE
WIKI-GLOBAL-TAG&LABEL	NONE&GPE
FIRST-WORD-CAP	1
CONTAINS-NUMBER	0
PREV-WORD-CAP	0
WORD-TYPE&LABEL	Xxxx&GPE
WORD-CLUSTER& LABEL	101&GPE
SEGMENT-WORD-TYPE&LABEL	Xxxx&GPE
SEGMENT-WORD-CLUSTER&LABEL	Xxxx&GPE
SOURCE-E-WIKI-TAG-MATCH	1
SOURCE-E-STANFORD-TAG-MATCH	0
SOURCE-E-WIKI-GLOBAL-TAG-MATCH	1
SOURCE-E-POSSIBLE-GLOBAL	1
SOURCE-E-ALL-TAG-MATCH	0
SOURCE-W-FWA-TAG & LABEL	GPE & GPE
SOURCE-W-BWA-TAG & LABEL	GPE & GPE
SCORE-SOURCE-E-WIKI-TAG-MATCH	-0.009
SCORE-SOURCE-E-GLOBAL-TAG-MATCH	-0.009
SCORE-SOURCE-E-STANFORD-TAG-MATCH	-1

Table 4: Features with example values.

We look at performance using four sets of features: (*i*) Monolingual Wiki-tagger based, using only the features in Group 1 (MONO); (*ii*) Bilingual label match and Wiki-tagger based, using features in Groups 1 and 3 (BI); (*iii*) Monolingual all, using features in Groups 1 and 2 (MONO-ALL), and (*iv*) Bilingual all, using all features (BI-ALL). Additionally, we report performance of the full bilingual model with all features, but when English candidate entities are generated only according to the local+global Wiki-taggger (BI-ALL-WT).

The main results show that the full semi-CRF model greatly outperforms the baseline projection and Wiki-taggers. For Bulgarian, the F-measure of the full model is 92.8 compared to the best baseline result of 83.2. For Korean, the F-measure of the semi-CRF is 91.2, more than 10 points higher than the performance of the projection model.

Within the semi-CRF model, the contribution of English sentence context was substantial, leading to 2.5 point increase in F-measure for Bulgarian (92.8 versus 90.3 F-measure), and 4.0 point increase for Korean (91.2 versus 87.2).

The additional gain due to considering candidate source entities generated from all English taggers was 1.3 F-measure points for both language pairs (comparing models BI-ALL and BI-ALL-WT).

If we restrict the semi-CRF to use only features similar to the ones used by the projection model, we still obtain performance much better than that of the projection model: comparing BI to the projection model, we see gains of 9.4 points for Bulgarian, and 4 points for Korean. This is due to the fact that the semi-CRF is able to relax the assumption of one-to-one correspondence between source and target entities, and can effectively combine information from multiple source and target taggers.

We should note that the proposed method can only tag foreign sentences in English-foreign sentence pairs. The next step for this work is to train monolingual NE taggers for the foreign languages, which can work on text within or outside of Wikipedia. Preliminary results show performance of over 80 F-measure for such monolingual models.

6 Related Work

As discussed throughout the paper, our model builds upon prior work on Wikipedia metadata-based NE tagging (Richman and Schone, 2008) and cross-lingual projection for named entities (Feng et al., 2004). Other interesting work on aligning named entities in two languages is reported in (Huang and Vogel, 2002; Moore, 2003).

Our direct semi-CRF tagging approach is related to bilingual labeling models presented in previous

work (Burkett et al., 2010a; Smith and Smith, 2004; Snyder and Barzilay, 2008). All of these models jointly label aligned source and target sentences. In contrast, our model is not concerned with tagging English sentences but only tags foreign sentences in the context of English sentences. Compared to the joint log-linear model of Burkett et al. (2010a), our semi-CRF approach does not require enumeration of n-best candidates for the English sentence and is not limited to n-best candidates for the foreign sentence. It enables the use of multiple unweighted and overlapping entity annotations on the English sentence.

7 Conclusions

In this paper we showed that using resources from Wikipedia, it is possible to combine metadata-based approaches and projection-based approaches for inducing named entity annotations for foreign languages. We presented a direct semi-CRF tagging model for labeling foreign sentences in parallel sentence pairs, which outperformed projection by more than 10 F-measure points for Bulgarian and Korean.

References

David Burkett, John Blitzer, and Dan Klein. 2010a. Joint parsing and alignment with weakly synchronized grammars. In *Proceedings of NAACL*.

David Burkett, Slav Petrov, John Blitzer, and Dan Klein. 2010b. Learning better monolingual models with unannotated bilingual text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 46–54, Uppsala, Sweden, July. Association for Computational Linguistics.

Colin Cherry and Hisami Suzuki. 2009. Discriminative substring decoding for transliteration. In *EMNLP*, pages 1066–1075.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA, June. Association for Computational Linguistics.

Donghui Feng, Yajuan Lv, and Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP*, pages 372–379.

Jenny Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*.

Fei Huang and Stephan Vogel. 2002. Improved named entity translation and bilingual named entity extraction. In *ICMI*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*, pages 127–133.

Robert C. Moore. 2003. Learning translations of named-entity phrases from parallel corpora. In *EACL*.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Alexander E. Richman and Patrick Schone. 2008. Mining wiki resources for multilingual named entity recognition. In *ACL*.

Sunita Sarawagi and William W. Cohen. 2004. Semi-markov conditional random fields for information extraction. In *In Advances in Neural Information Processing Systems 17*, pages 1185–1192.

Sunita Sarawagi and William W. Cohen. 2005. Semi-markov conditional random fields for information extraction. In *In Advances in Neural Information Processing Systems 17 (NIPS 2004)*.

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: using English to parse Korean. In *EMNLP*.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *HLT*, pages 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Benjamin Snyder and Regina Barzilay. 2008. Crosslingual propagation for morphological analysis. In *AAAI*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT*.