Joint Bilingual Name Tagging for Parallel Corpora

Qi Li, Haibo Li, Heng Ji
Computer Science Department and Linguistics Department
Queens College and Graduate Center, City University of New York
{liqiearth,lihaibo.c,hengjicuny}@gmail.com
Wen Wang, Jing Zheng
Speech Technology and Research Laboratory, SRI International
{wwang,zj}@speech.sri.com
Fei Huang
IBM T.J. Watson Research Center
{huangfe}@us.ibm.com

ABSTRACT

Traditional isolated monolingual name taggers tend to yield inconsistent results across two languages. In this paper, we propose two novel approaches to jointly and consistently extract names from parallel corpora. The first approach uses standard linear-chain Conditional Random Fields (CRFs) as the learning framework, incorporating cross-lingual features propagated between two languages. The second approach is based on a joint CRFs model to jointly decode sentence pairs, incorporating bilingual factors based on word alignment. Experiments on Chinese-English parallel corpora demonstrated that the proposed methods significantly outperformed monolingual name taggers, were robust to automatic alignment noise and achieved state-of-the-art performance. With only 20% of the training data, our proposed methods can already achieve better performance compared to the baseline learned from the whole training set.
[1]

1. INTRODUCTION

Effective extracting and aligning names from bilingual data is an important task to various natural language processing (NLP) and information access applications, such as name pair and translation template mining [7], statistical word alignment [6], machine translation (MT) [8], cross-lingual information extraction [17], cross-lingual information retrieval [4] and cross-lingual question answering [15]. This is a challenging task because it requires both name tagging from two languages and alignment to be handled correctly. However, traditional name tagging approaches for single languages (e.g. [10, 16]) cannot address this requirement because they were all built on data and resources which are specific to each language without using any cross-lingual features. In addition, due to separate decoding processes and word alignment errors, the results on parallel data may not be consistent across languages, even if monolingual name tagging systems achieved state-of-the-art performance on each language. Previous methods on bilingual lexicon acquisition cannot address this problem either because they relied on frequency of words appearing in bilingual corpora [19]. But many names are often domain specific and new names are created frequently so they cannot be found in existing bilingual gazetteers.

Fortunately, each language-specific tagger has its own advantages and disadvantages, and the features and resources from two languages are often complementary for parallel data. Since the parallel data sets for English and Chinese are easier for us to obtain and understand, we choose the pair of English and Chinese as an example language pair, with three types of names: persons (PER), organizations (ORG) and geo-political entities (GPE). On one side, English features can help Chinese name tagging in various ways. (1). Due to its efficient symbolic system [3], Chinese texts often include ambiguous name abbreviations which can also be interpreted as common words. However, they appear as full names in English translation which are much easier to detect, such as organization name: "亚行 *(*Asian Development Bank)" and GPE names: "津 *(Tianjin)*", "台 *(Taiwan)*". (2). English features can help fix Chinese word segmentation errors. For example, it's difficult to segment a sequence of local names "山东省平邑县九间棚村" because it includes some common words such as "九间 */ nine rooms*" and "棚 */ shed*". But its English translation is based on pronunciation: "*Jiujianpeng village of Pingyi county of Shandong province*", which clearly indicates three GPE names: "九间棚村*/Jiujianpeng village*", "平邑县*/Pingyi county*" and "山东省*/Shandong province*". (3) Chinese nested organizations are often translated into simpler abbreviations in English and thus easier to identify, e.g."香港上海汇丰银行 */ Hong Kong Shanghai Huifeng Bank*" is translated into "*HSBC*" in English.

On the other hand, Chinese features can help English name tagging: (1) name identification: It's easy to identify"贵州茅台酒厂（集团）公司 */Guizhou Mao - tai Distillery (Group) Company* " as an organization name in Chinese, but in English "Guizhou Mao" can be mistakenly tagged as a person name because of its capitalization feature and the incorrect tokenization between "Mao" and "tai". (2) name classification: In English it's often difficult to disambiguate organization from facility, while the Chinese contexts often use different prepositions or verbs to indicate the correct name types.

Based on the above motivations, we develop the first new model based on linear-chain CRFs by projecting features from one lan-

[1]All of the resources and open source programs developed in this paper are made freely available for research purpose at http://nlp.cs.qc.cuny.edu/cuny_jointtagger.tar.gz

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

guage to the other using word alignment. However, in this framework, the knowledge from two languages is implicitly transferred on feature-level instead of label-level. Therefore, the tagged name pairs are not guaranteed to be consistent. For example, in the following sentence, "新华社杭州九月二日电（记者慎海雄沈锡权*) / Xinhua News Agency, Hangzhou, September 2nd, by reporters Haixiong Shen and Xiquan Shen*" , the Chinese baseline tagger mistakenly labeled "慎海雄沈锡权 */Haixiong Shen Xiquan Shen*" as one single name. Cross-lingual features cannot fix this error because both of its name type and boundary match the two names in English "*Haixiong Shen*" and "*Xiquan Shen*". We propose the second model based on joint CRFs which not only incorporates both monolingual and cross-lingual features, but also conduct decoding for two languages simultaneously so that the labeling of two sides can mutually enhance each other. For the above example, the joint CRFs can split the Chinese name into two: "慎海雄 */Haixiong Shen*" and "沈锡权 */Xiquan Shen*".

2. RELATED WORK

Some recent work has explored name tagging for parallel data. [13] presented a sequence of cost models to learn name translation pairs. This approach greatly relies on language-specific information such as repeated strings from both languages and capitalization clues. [9] proposed an approach to extract bilingual name pairs. Their method extracted names from each language first, and then computed the cost scores based on name tagging, name transliteration and word translation to rank candidate name pairs. [5] extended their ranking method by incorporating bilingual alignment, bilingual type re-assignment and monolingual candidate certainty. [11] described a joint inference model to improve entity extraction and translation. All of these previous approaches can still be considered as adding a post-processing step after two isolated name taggers. In contrast, our proposed joint CRFs approach integrates name tagging and alignment into one single unified model.

To the best of our knowledge, there was no previous work that we can directly compare to because they all used different data sets or definitions of name types. The approach described in [5] is the closest to our method of linear-chain CRFs using cross-lingual features, but there are some fundamental differences between their work and ours in the following aspects. (1). [5] was a two-step method. They added a post-processing step after two isolated name taggers. Their method extracted names from each language with monolingual name tagging first, and then computed the linking score to align name pairs. Therefore, in their approach, Chinese-English name pair lists with translation or transliteration confidence values are required to compute the linking scores. In contrast our proposed joint CRFs approach naturally integrates name tagging and bilingual alignment into one decoding process and does not require any translation and transliteration components to discover the mappings betweens names from different languages. (2). The proposed learning methods are different. (3). They used manual alignment during testing, while we evaluate on both manual alignment and automatic alignment.

3. BASELINE APPROACH

The input of a bilingual name tagger is aligned (manually or automatically) parallel sentence pairs in two languages (Chinese and English in this paper). We apply the Stanford word segmenter [2] with Peking University standard to segment Chinese sentences. For example, for the parallel sentence pair demonstrated in Figure 1, our system should extract name pairs that appear in two sentences,

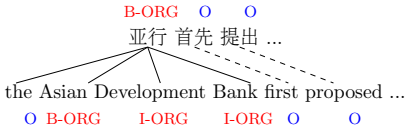

Figure 1: Example of a parallel sentence pair. Solid lines represent alignments between names, while dashed lines denote other alignments

such as the organization name pair of (亚行, Asian Development Bank).

A natural approach is to consider each side of a sentence pair in isolation, and solve the sequence labeling problem on each side using linear-chain CRFs. And during post-processing, we can remove all of those name pairs that are mis-aligned in boundaries or labeled with different types.

We adopt the linear-chain CRFs [12] as our learning method. In linear-chain CRFs, given an input sequence $\mathbf{x}$, the conditional distribution of the output label sequence $\mathbf{y}$ is defined as:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \cdot exp \sum_{j=1}^{L} \sum_{k=1}^{K} \theta_k \cdot f_k(y_j, y_{j-1}, \mathbf{x}, j) \quad (1)$$

where f_k is a feature function, θ_k is its weight, and $Z(\mathbf{x})$ is a normalization function factor.

To cast name tagging as a sequence labeling problem, the *BIO* tagging scheme [16] is applied as our label alphabet. BIO stands for: *B-X*, token at the beginning of a name; *I-X*, token within a name; and *O*, token out of any name, where *X* denotes the name type. Table 1 summarizes the features for the baseline, where we assume the i-th token is the token in the current step.

4. LINEAR-CHAIN CRFS WITH CROSS-LINGUAL FEATURES

The baseline approach described above neglects the dependency between aligned sentences. Given the hypothesis that the context of sentences pair can help disambiguation and reduce errors mutually, we present a new approach which still takes linear-chain CRFs as the learning framework, but exploits cross-lingual contexts based on alignments.

Let $\mathbf{x}_c = (x_{c,1}...x_{c,L})$ and $\mathbf{x}_e = (x_{e,1}...x_{e,M})$ be the input Chinese-English sentence pair; $\mathbf{y}_c = (y_{c,1}...y_{c,L})$ and $\mathbf{y}_e = (y_{e,1}...y_{e,M})$ be the corresponding output label sequences. The subscripts c and e denote Chinese and English respectively. In Chinese each $x_{c,i}$ is a word, while in English each $x_{e,j}$ represents a token. We use $\mathcal{A} = \{(i,j)\}$ to denote the set of Chinese-English alignments, an alignment (i,j) indicates a Chinese word $x_{c,i}$ is aligned to an English token $x_{e,j}$. For simplicity we take Chinese side as an example, the hidden variables $\mathbf{y}_c$ is not only conditioned on $\mathbf{x}_c$ but also conditioned on $\mathbf{x}_e$ and its alignment $\mathcal{A}$. The conditional probability of $\mathbf{y}_c$ can be extended as:

$$P(\mathbf{y}_c|\mathbf{x}_c, \mathbf{x}_e, \mathcal{A}) = \tfrac{1}{Z} \cdot exp \sum_{i=1}^{T} \sum_{k=1}^{K} (\theta_k \cdot f_k(y_{c,i}, y_{c,i-1}, \mathbf{x}_c, \mathbf{x}_e, i, \mathcal{A}[i])) \quad (2)$$

where $\mathcal{A}[i]$ represents the indices of English tokens which are aligned to the i-th Chinese word. This still follows the linear-chain structure in which we need to build one model for each language. The distinction from the baseline approach is that, with an English sequence $\mathbf{x}_e$ and its alignment $\mathcal{A}$, we can propagate the context from

Language	Feature	Description
Common Language-independent	n-gram	Unigram, bigram and trigram token sequences in the context window of the current token. For example, $w_{-2}w_{-1}$=", the"; w_3="first".
	Part-of-Speech	Part-of-Speech tags in the contexts are used. For example, POS_1=N .
	Dictionary	Various types of gazetteers, such as person names, organizations, countries and cities, titles and idioms are used. For example, a feature "B-Country" means the current token is the first token of an entry of our country name list.
	Conjunction	Conjunctions of various features. For example, POS_1POS_2=N&N
English-specific	Brown Word Cluster	To reduce sparsity, we use the Brown clusters learned from ACE English corpus as features [1]. We use the clusters with prefixes of length 4, 6, 10 and 20.
	Case and Shape	English capitalization and morphology analysis based features. For example, "*InitCap*" indicates whether the token's first character is capitalized.
	Chunking	Chunking tags are used as features. For example, $Chunk_1 = I\_NP$.
	Global feature	Sentence level and document level features. For example, T_{FIRST} means the token is in the first sentence of a document.
Chinese-specific	Rule-based feature	Some heuristic rules are designed to detect person names using first name and last name character lists. For example, for a sequence of words, if all characters appear in the first name character list, and the length of each word is less than 2, then the sequence is likely to be a person's first name.

Table 1: Monolingual Features of Baseline Systems, with Figure 1 as an example. The token "Asian" of the English sentence is used as current token. The subscripts represent the offsets from the current token.

English to Chinese according to its alignment, and vice versa. Therefore, not only is the output from two languages more accurate, but the entity pair detection performance is improved consequently as well. Ideally, we can generate arbitrary variants from the feature function $f_k(y_{c,i}, y_{c,i-1}, \mathbf{x}_c, \mathbf{x}_e, i, \mathcal{A}[i])$. In practice, we use the same feature set as in Section 3, but aggregate features of $x_{e,j}$ and its corresponding English tokens as observed features.

5. BILINGUAL JOINT CRFS MODEL

Although the approach in section 4 already takes into account of the dependencies between sentence pair, it requires separate models for two languages, and the prediction from one side cannot directly influence the assignment of the other, because the inference is on implicit feature level rather than the label level. In this section, we propose a bilingual CRFs framework which jointly models the bilingual sentence pair by utilizing their alignments.

We define the conditional probability of output $\mathbf{y_c}$ and $\mathbf{y_e}$ jointly as:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{i=1}^{L} \Psi_c(y_{c,i}, y_{c,i-1}, \mathbf{x}_c, i) \prod_{j=1}^{M} \Psi_e(y_{e,j}, y_{e,j-1}, \mathbf{x}_e, j) \prod_{(i,j)\in\mathcal{A}} \Psi_a(y_{c,i}, y_{e,j}, \mathbf{x}_c, \mathbf{x}_e, i, j) \quad (3)$$

This distribution is factorized by three cliques of factors: $\{\Psi_c\}$ are potentials of Chinese linear-chain factors, $\{\Psi_e\}$ are potentials of English linear-chain factors, and $\{\Psi_a\}$ are potentials of bilingual factors. Factors in each clique share the same feature set and weights. $Z(\mathbf{x})$ is the normalization factor which sums over potentials of all possible assignments of $\mathbf{y}_c$ and $\mathbf{y}_e$.

5.1 Monolingual Linear-chain Factors

Similar to monolingual name tagging, for any sentence in each language we define factors over all pairs of consecutive variables (y_{t-1}, y_t), which enables the model to capture the dependency between consecutive variables. The potential function of monolingual factors Ψ_c and Ψ_e is defined as

$$\Psi(y_t, y_{t-1}, \mathbf{x}, t) = \exp(\sum_k \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}, t)) \quad (4)$$

where f_k is a binary feature function and λ_k is the corresponding real-valued weight.

5.2 Bilingual Alignment Factors

The label of a Chinese word is often highly correlated with its aligned English token, and vice versa. For instance, in the example in Figure 1, the Chinese word "亚行" and its English counterpart "*Asian Development Bank*" should be both labeled as organizations. In order to model the correlation between the labels of aligned word-tokens, we introduce factors that link output variables in two languages based on alignments. For alignment (i, j) in which Chinese word $x_{c,i}$ is aligned to English token $x_{e,j}$, we define a bilingual factor over $y_{c,i}$ and $y_{e,j}$. This factor template bridges two monolingual linear chains, and makes it possible to propagate information across two sentences. The potential function of bilingual factors Ψ_a is defined as:

$$\Psi_a(y_{c,i}, y_{e,j}, \mathbf{x}_c, \mathbf{x}_e, i, j) = \exp(\sum_k \lambda_k f_k(y_{c,i}, y_{e,j}, \mathbf{x}_c, \mathbf{x}_e, i, j)) \quad (5)$$

This allows us to design arbitrary binary features based on both $\mathbf{x}_c$ and $\mathbf{x}_e$. A simple feature function for the above example is:

$$f(y_{c,i}, y_{e,j}, \mathbf{x}_c, \mathbf{x}_e, i, j) = \begin{cases} 1 & \text{if } y_{c,i} = \text{B-ORG}, y_{e,j} = \text{I-ORG and } x_{e,j} = \text{Bank} \\ 0 & \text{otherwise} \end{cases}$$

If this feature attains high weight, the aligned word-token pair is likely to represent an organization entity given the English token is "*Bank*".

Figure 2 illustrates the factor graph representation of the model for the example in Figure 1. In this figure, white circles represent hidden variables $\mathbf{y}_c$ and $\mathbf{y}_e$, gray circles represent observed sentence pair. Theoretically the factors can be linked to the whole observed sequences, for simplicity we only show the link to those at the same step.

5.3 Inference and Training

Since cycles are introduced by bilingual factors, typical inference algorithms for marginal probability and MAP such as Forward-backward and Viterbi algorithms cannot be exploited, and the exact inference is intractable in general. In this work we use an efficient loopy belief propagation method named Tree-Based Reparameterization (TRP) [20, 18] to perform approximate inference on the loopy graph.

Given a set of training data $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$, the feature weights $\Lambda = \{\lambda_k\}$ are estimated using maximum likelihood estimation

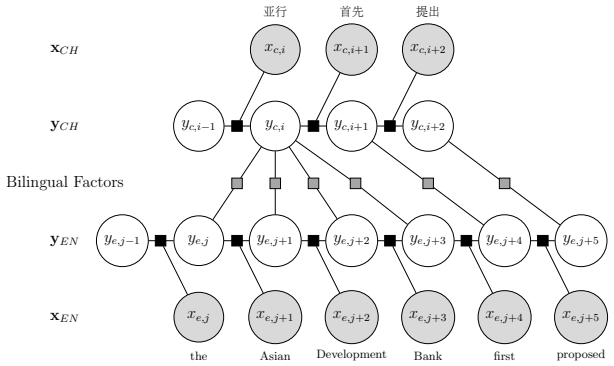

Figure 2: Graphical representation of bilingual CRFs model. Squares represent factors over input and output variables, for simplicity, the links between bilingual factors and input variables are not shown.

(MLE). The log-likelihood of the training set is calculated as:

$$\mathcal{L}_\Lambda = \sum_{i=1}^{N} \log p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)};\Lambda) \tag{6}$$

To avoid over-fitting we introduce Gaussian prior $\frac{|\Lambda|^2}{2\sigma^2}$ as regularization term to $\mathcal{L}_\Lambda$. Then the partial derivative of the log-likelihood with respect to parameter λ_k is:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \lambda_k} &= \sum_{i=1}^{N} F_k(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \\ &- \sum_{i=1}^{N} \sum_{\mathbf{y}'} p(\mathbf{y}'|\mathbf{x}^{(i)};\Lambda) F_k(\mathbf{x}^{(i)}, \mathbf{y}') - \frac{\lambda_k}{\sigma^2}\end{aligned} \tag{7}$$

where $F_k(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ denotes the count of feature f_k over the i-th instance. The first term is the empirical count of λ_k, and the second term is the expected count of λ_k under the model distribution. Given the gradient, optimization algorithms such as L-BFGS can be applied to maximize the log-likelihood.

5.4 Features

Given such a framework, the remaining challenge is to design features for both monolingual and bilingual factors. There are various possible ways to define cross-lingual features in this joint model. For instance, one possibility is to define them based on some conjunctions of the observed values from two languages, but such features require very large of training data and thus suffer from data sparsity. In our framework, each feature is defined as a conjunction of assignment and features from the input sequence; therefore we only need to design features of the input sequence. We use the features presented in Section 3 for monolingual factors. The features for the proposed bilingual factors are based on the combination of the monolingual features from the corresponding words/tokens. For instance, given a bilingual factor over $x_{c,i}$ and $x_{e,j}$ and alignment (i, j), the sets of monolingual features from $x_{c,i}$ and $x_{e,j}$ are merged as features to form the factor. In this way, both monolingual features and cross-lingual transferred features are incorporated in a uniformed manner.

6. EXPERIMENTS

6.1 Data Set and Evaluation Measures

Type	English	Chinese	Bilingual Pairs
GPE	4049	4077	4031
PER	1053	1048	1044
ORG	1547	1549	1541
ALL	6649	6674	6616

Table 2: The number of names in the bilingual data set.

We asked four bilingual speakers to manually annotate the Parallel Treebank, which contains 288 Chinese-English parallel documents aligned at token level manually. The manual annotations were reviewed and adjudicated/corrected with several additional passes to form the final ground-truth. 230 documents are randomly selected for training, and the remaining 58 documents are used for blind test. Some statistics about this bilingual data set are given in Table 2. The last column (Bilingual Pairs) of the table shows the number of name pairs detected with manual alignment. Since the translation is not exactly literal, some names in one language may have no correspondences in the other. As a result, the number of name pairs may be slightly smaller than the number of names in each language.

A name pair in a system output is considered as correct if and only if both names in two languages are correct and they have the same name type. The scores are computed using bilingual sentence pairs and name pairs, which are detected according to token-based alignment.

6.2 Overall Performance Comparison

Table 3 shows the proposed approaches (with both manual alignment and automatic alignment [14][2]) dramatically outperformed the baseline on all name types, at 99.9% confidence level according to Wilcoxon Matched-Pairs Signed-Ranks Test. The joint model achieved even better performance than single human annotator on person names with manual alignment; and the top F-score with automatic alignment for organization names. This indicates that our proposed models are robust to alignment noise so that they can be effectively applied to bilingual parallel data with automatic alignment, and avoid the necessity of costly manual alignment.

6.3 Learning Curves

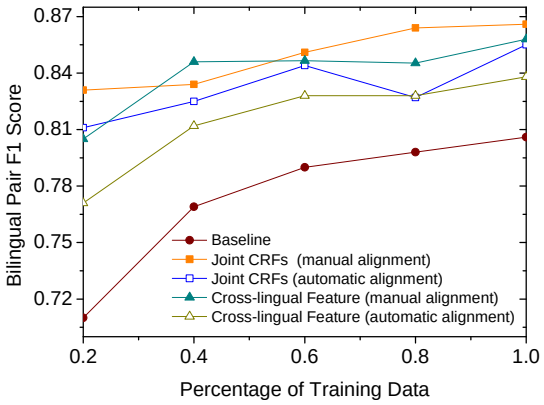

Figure 3: Performance on different size of training data.

[2]We applied GIZA++ 2.0 toolkit to produce automatic word alignment; default parameter setting for training, 5 iterations of IBM model 1, 3 ,4 and HMM alignment model were performed respectively; the alignment f-measure was 56.7%

Method \ Type		Bilingual Name Pair Tagging			
		GPE	PER	ORG	ALL
Baseline	P	89.2	91.3	68.9	85.9
	R	86.7	81.8	51.1	78.2
	F	87.9	86.3	58.7	81.9
Linear-Chain CRFs with Cross-lingual Features (Manual Alignment)	P	91.2	92.4	69.8	87.2
	R	91.4	89.1	61.5	84.5
	F	**90.7**	**91.2**	**65.4**	**85.8**
Joint CRFs (Manual Alignment)	P	90.8	94.0	68.6	86.9
	R	92.8	90.1	61.5	85.6
	F	***91.8****	***92.0****	**64.9**	***86.3****
Linear-Chain CRFs with Cross-lingual Features (Automatic Alignment)	P	90.6	97.0	70.7	87.8
	R	88.8	83.9	57.6	81.3
	F	**89.7**	**90.0**	**63.5**	**84.4**
Joint CRFs (Automatic Alignment)	P	89.9	92.6	71.2	86.6
	R	88.7	84.9	61.9	82.3
	F	**89.3**	**88.6**	***66.2****	**84.4**
Human Annotator	F	95.5	89.9	93.8	94.1

Table 3: Performance (%) on bilingual data set (the bold F-scores are significantly better than the baseline; while the scores marked with * are the best for each type).

Figure 3 shows the overall performance of our models when they are learned from different size of training data. In order to balance the small size of training data and test data, we randomly selected half of the test set (29 documents) for test. We can see that with only 20% of the training data, each of our proposed methods (with manual alignment or automatic alignment) can already achieve better performance compared to the baseline learned from 100% training data. In particular, when using 20% training data, the joint CRFs model obtained 12.1% higher F-score with manual alignment and 10.1% higher F-score with automatic alignment over the baseline. As the training size increases, the proposed approaches consistently outperformed the baseline.

7. CONCLUSIONS AND FUTURE WORK

In this paper we developed two novel bilingual name tagging methods incorporating cross-lingual features to jointly extract names from bilingual data which significantly outperformed high-quality single-language name taggers, and achieved state-of-the-art performance.

Although our experiments were conducted on the Chinese-English name tagging task, we believe our proposed models are generally applicable for other NLP tasks (e.g. POS tagging and chunking) in other language pairs which contain complementary linguistic features. Currently the labeled monolingual data is widely available while manual annotations for parallel data remain highly expensive. Therefore in the future we are interested in exploring semi-supervised learning algorithms for joint bilingual name tagging. Finally, the joint improvement of name tagging and alignment can be iteratively executed, therefore it would be interesting to investigate the convergence conditions.

8. ACKNOWLEDGEMENTS

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053, the U.S. NSF Grants IIS-0953149 and IIS-1144111 and the U.S. DARPA BOLT program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

9. REFERENCES

[1] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, pages 467–479, 1992.

[2] P.-C. Chang, M. Galley, and C. D. Manning. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, June 2008.

[3] Y. R. Chao. The efficiency of the chinese language. In *Proc. the General Conference of UNESCO*, 1946.

[4] H.-H. Chen, S.-J. Huang, Y.-W. Ding, and S.-C. Tsai. Proper Name Translation in Cross-Language Information Retrieval. In *Proc. ACL*, 1998.

[5] Y. Chen, C. Zong, and K.-Y. Su. On jointly recognizing and aligning bilingual named entities. In *ACL*, 2010.

[6] Y. Deng and Y. Gao. Guiding Statistical Word Alignment Models With Prior Knowledge. In *Proc. ACL*, 2007.

[7] D. Feng, Y. Lv, and M. Zhou. A new approach for english-chinese named entity alignment. In *Proc. PACLIC*, 2004.

[8] U. Hermjakob, K. Knight, and H. D. III. Name translation in statistical machine translation: Learning when to transliterate. In *Proc. ACL*, 2008.

[9] F. Huang and S. Vogel. Improved named entity translation and bilingual named entity extraction. In *Proc. 2002 International Conference on Multimodal Interfaces*, 2002.

[10] H. Ji and R. Grishman. Analysis and repair of name tagger errors. In *Proc. COLING-ACL*, 2006.

[11] H. Ji and R. Grishman. Collaborative entity extraction and translation. In *Proc. RANLP*, 2007.

[12] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[13] R. C. Moore. Learning translations of named-entity phrases from parallel corpora. In *Proc. EACL*, 2003.

[14] F. J. Och and H. Ney. Improved statistical alignment models. In *ACL*, 2000.

[15] K. Parton and K. McKeown. Mt error detection for cross-lingual question answering. *Proc. COLING2010*, 2010.

[16] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *CoNLL2009)*, 2009.

[17] M. Snover, X. Li, W.-P. Lin, Z. Chen, S. Tamang, M. Ge, A. Lee, Q. Li, H. Li, S. Anzaroot, and H. Ji. Cross-lingual slot filling from comparable corpora. In *Proc. ACL2011 Worshop on Building and Using Comparable Corpora*, 2011.

[18] C. A. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8:693–723, 2007.

[19] K. Tsuji. Automatic extraction of translational japanese-katakana and english word pairs from bilingual corpora. 15(3), 2002.

[20] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. Tree-based reparameterization for approximate inference on loopy graphs. In *NIPS*, pages 1001–1008, 2001.