International Journal of Computational Linguistics and Applications, vol. 5, no. 2, 2014, pp. 95–113
Received 5/01/2014, accepted 6/02/2014, final 10/12/2014.
ISSN 0976-0962, http://ijcla.bahripublications.com

Unsupervised Active Learning of CRF Model for Cross-Lingual Information Extraction

MOHAMED FAROUK ABDEL HADY, ABUBAKRELSEDIK KARALI, ESLAM KAMAL, AND RANIA IBRAHIM

Microsoft Research, Egypt

ABSTRACT

Manual annotation of the training data of information extraction models is a time consuming and expensive process but necessary for the building of information extraction systems. Active learning has been proven to be effective in reducing manual annotation efforts for supervised learning tasks where a human judge is asked to annotate the most informative examples with respect to a given model. However, in most cases reliable human judges are not available for all languages. In this paper, we propose a cross-lingual unsupervised active learning paradigm (XLADA) that generates high-quality automatically annotated training data from a word-aligned parallel corpus. To evaluate our paradigm, we applied XLADA on English-French and English-Chinese bilingual corpora then we trained French and Chinese information extraction models. The experimental results show that XLADA can produce effective models without manually-annotated training data.

KEYWORDS: *Information extraction, named entity recognition, cross-lingual domain adaptation, unsupervised active learning.*

1 INTRODUCTION

Named Entity Recognition (NER) is an information extraction task that identifies the names of locations, persons, organizations and other named

entities in text, which plays an important role in many Natural Language Processing (NLP) applications such as information retrieval and machine translation. Numerous supervised machine learning algorithms, such as Maximum Entropy, Hidden Markov Model, and Conditional Random Field (CRF) [1], have been adopted for NER and achieved high accuracy. They usually require large amount of manually annotated training examples. However, it is time-consuming and expensive to obtain labeled data to train supervised models. Moreover, in sequence modeling like NER task, it is more difficult to obtain labeled training data since hand-labeling individual words and word boundaries is really complex and need professional annotators. Hence, the shortage of annotated corpora is the obstacle of supervised learning and limits the further development, especially for languages for which such resources are scarce.

Active learning is the method which, instead of relying on random sampling from the large amount of unlabeled data, reduces the cost of labeling by actively guiding the selection of the most informative training examples: an oracle is asked for labeling the selected sample. There are two settings depending on the oracle type: supervised setting [2], which requires human annotators as oracle for manual annotation, and the unsupervised setting, where the oracle is an automation process. Using different settings, active learning may find much smaller and most informative subset of the unlabeled data pool. The difference between unsupervised active learning and semi-supervised learning [3] is that the former depends on an oracle to automatically annotate the most informative examples with respect to the underlying model. The later depends on the underlying model to automatically annotate some unlabeled data, to alleviate mislabeling noise the model selects the most confident examples.

For language-dependent tasks such as information extraction, to avoid the expensive re-labeling process for each individual language, cross-lingual adaptation, is a special case of domain adaptation, refers to the transfer of classification knowledge from one source language to another target language.

In this paper, we present a framework for incorporating unsupervised active learning in the cross-lingual domain adaptation paradigm (*XLADA*) that learns from labeled data in a source language and unlabeled data in the target language. The motivation of *XLADA* is to collect large-scale training data and to train an information extraction model in a target language without manual annotation but with the help of an effective information extraction system in a source language, bilingual corpus and word-level alignment model.

2 Related Work

2.1 *Cross-lingual Domain Adaptation*

Guo and Xiao [4] developed a transductive subspace representation learning method to address domain adaptation for cross-lingual text classifications. The proposed approach is formulated as a non-negative matrix factorization problem and solved using an iterative optimization procedure. They assume there is a shared latent space over the two domains, such that one common prediction function can be learned from the shared representation for both domains.

Prettenhofer and Stein [5] presented a new approach to cross-lingual domain adaptation that builds on structural correspondence learning theory for domain adaptation. The approach uses unlabeled documents, along with a simple word translation oracle, in order to induce task specific, cross-lingual word correspondences. The analysis reveals quantitative insights about the use of unlabeled data and the complexity of inter language correspondence modeling.

Wan et al. [6] present a transfer learning approach to tackle the cross-lingual domain adaptation. They first align the feature spaces in both domains utilizing some online translation service, which makes the two feature spaces under the same coordinates. They propose an iterative feature and instance weighting (Bi-Weighting) method for domain adaptation. The main idea here is to select features which have distinguished utility for classification from source language and make distributions of source and target languages as similar as possible. In this way, the features useful for classifying instances in source could also be functional for classification on target.

2.2 *Automatic Generation and Annotation of Training Data for Information Extraction*

Yarowsky et al. [7] used word alignment on parallel corpora to induce several text analysis tools from English to other languages for which such resources are scarce. An NE tagger was transferred from English to French and achieved good classification accuracy. However, Chinese NER is more difficult than French and word alignment between Chinese and English is also more complex because of the difference between the two languages.

Some approaches have exploited Wikipedia as external resource to generate NE tagged corpus. Richman and Schone [8] and Nothman et

al. [9] used similar methods to create NE training data. They transformed Wikipedia's links into named entity annotations by classifying the target articles into common entity types. But the article classification seeds also had to be hand-labeled in advance. Kim et al. [10] 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. The model outperforms both standard annotation projection methods and methods based solely on Wikipedia metadata. *XLADA* does not leverage Wikipedia because its content is poor in some languages like Chinese.

Fu et al. [11] presents an approach to generate large-scale Chinese NER training data from an English-Chinese discourse level aligned parallel corpus. It first employs a high performance NER system on one side of a bilingual corpus. And then, it projects the NE labels to the other side according to the word level alignment. At last, it selects labeled sentences using different strategies and generate an NER training corpus. This approach can be considered as passive domain adaptation while *XLADA* is active learning framework that filters out the auto-labeled data and selects the most informative training sentences.

2.3 *Active Learning for Information Extraction*

Muslea et al. [12] introduced *Co-Testing*, a multi-view active learning framework, where two models are trained on two independent and sufficient sets of features. The most informative sentences are the points of disagreement between the two models that could improve their performance and a human judge is asked for labeling them. On the other hand, *XLADA* looks for the most informative sentences for the target model and we don't have judges.

Jones et al. [3] adapted semi-supervised learning *Co-EM* to information extraction tasks to learn from both labeled and unlabeled data that makes use of two distinct feature sets (training document's noun phrases and context). It is interleaved in the supervised active learning framework *Co-Testing*. *XLADA* differs in that cross-lingual label propagation on a parallel corpus is interleaved for automatic annotation instead of using *Co-EM* approach and that it adopts an unsupervised active learning strategy.

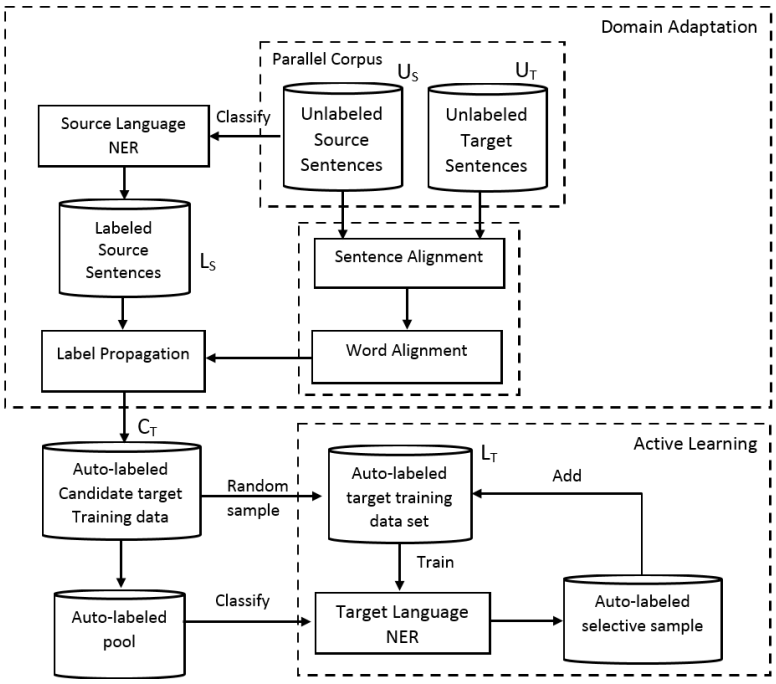

Fig. 1. Architecture of cross-lingual active domain adaptation (*XLADA*)

XLADA is more practical than the framework proposed by Li et al. [13] that depends on cross-lingual features extracted from the word-aligned sentence pair in training the target language CRF model. Hence, it isn't possible to extract named entities from a sentence in the target language unless it is aligned with a sentence in the source language.

3 Algorithmic Overview

The architecture of the proposed combination of cross-lingual domain adaptation and active learning paradigm *XLADA* is shown in Figure 1.

3.1 *Initial Labeling*

Source Language NER An effective source language NER is applied on the source-side of the bilingual corpus U_S to identify named entities

Fig. 2. Projection of named-entity tags from English to Chinese and French sentences

such as person, location, organization names, denote the output L_S. In our experiments, the source language is English and English Stanford NER[1] is used. The system is based on linear chain CRF [1] sequence models that can recognize three types of named entities (Location, Person and Organization).

WORD ALIGNMENT OF PARALLEL CORPUS Sentence alignment and word alignment is performed on the given unlabeled bilingual corpus U_S and U_T. First, sentence level alignment is performed then we applied word dependent transition model based HMM (WDHMM) for word alignment [14].

LABEL PROPAGATION We project the NE labels to the target side of the parallel corpus to automatically annotate target language sentences, according to the result of word alignment, as shown in Figure 2. The output is a set of candidate training sentences. A target sentence is filtered out from the set of candidate training sentences if the number of named entities after label propagation is less than the number of named entities in the source sentence.

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

3.2 *Unsupervised Active Learning*

The amount of auto-labeled sentences in the target language training is too huge to be used for training the information extraction model. Also they are noisy because of the errors in source language NER or word-level alignment. Unsupervised active learning is adopted for selecting high quality training sentences used to train CRF model. The manual annotation of the selected sentences by human judges is replaced with the alignment-based automatic annotation.

We randomly select a set of auto-labeled training sentences L_T. An initial CRF model is trained with L_T. Since a random set of auto-labeled sentences is not sufficient to train a good prediction model in the target language, additional labeled data is required to reach a reasonable prediction model. Afterward, *XLADA* will proceed in an iterative manner.

A pool CP_T of the large amount of auto-labeled sentences is selected. There are two ways to select the sentences in the pool, either a random sample or by assigning a score for each target sentence and finally choose sentences with the highest score (most confident sentences).

The score of each target sentence depends on the score given to its corresponding source sentence in the parallel corpus, as follows:

$$score(S) = \min_{w_i \in S} \max_{c_j \in classes} P(c_j|w_i, \theta_{src})$$

The source NER model θ_{src} assigns probability for each token of how likely it belongs to each entity type: person, location, organization or otherwise. Then, the entity type for each token is the class with maximum probability $P(c_j|w_i, \theta_{src})$. We apply the forward-backward algorithm to compute them.

In each round of active learning, the current target NER model θ_{tgt} tags each target sentence in the auto-labeled pool CP_T. The critical challenge lies in how to select the most informative sentences for labeling. Based on different measurements of target sentence informativeness, we propose the following metric to measure how informative is a given sentence S.

$$inform(S) = \frac{1}{N(S)} \sum_{w_i \in S} \hat{y}(w_i) P(\hat{y}_{tgt}(w_i)|w_i, \theta_{tgt})$$

where $\hat{y}(w_i) = I(\hat{y}_{src}(w_i) \neq \hat{y}_{tgt}(w_i))$ the indicator boolean function between

$$\hat{y}_{src}(w_i) = \arg \max_{c_j \in classes} P(c_j|w_i, \theta_{src})$$

the NE label propagated from the source NER model θ_{src} through alignment information and

$$\hat{y}_{tgt}(w_i) = \arg \max_{c_j \in classes} P(c_j|w_i, \theta_{tgt})$$

the NE tag assigned by the current target NER model θ_{tgt} to the i^{th} word in S.

The most informative sentences are the ones that the target NER model θ_{tgt} didn't learn yet (least confident on its NE prediction) and mismatch with the source NER model θ_{src} where $N(S)$ is the number of tokens in S the two models disagree. Then we select the top N sentences or the ones less than a predefined threshold, add them to L_T with the automatic labels propagated from the source NER model θ_{src} and remove them from the pool CP_T. After new labels being acquired, the target model is retrained on the updated L_T.

3.3 *Conditional Random Field*

Conditional Random Fields (CRFs) [15], similar to the Hidden Markov Models (HMMs) [16] , are a type of statistical modeling method used for labeling or parsing of sequential data, such as natural language text and computer vision. CRF is a discriminative undirected probabilistic graphical model that calculates the conditional probability of output values for a given observation sequence. HMMs made strong independence assumption between observation variables, in order to reduce complexity, which hurts the accuracy of the model while CRF does not make assumptions on the dependencies among observation variables.

Figure 3 shows the graphical representation of liner chain CRFs. Because of its linear structure, linear chain CRF is frequently used in natural language processing to predict sequence of labels Y for a given observation sequence X. The inference of a linear-chain CRF model is that given an observation sequence X, we want to find the most likely sequence of labels Y. The probability of Y given X is calculated as follows:

$$P(Y|X) = \frac{1}{Z(X)} \exp(\sum_{t=1}^{T} \sum_{i=1}^{n} w_i f_i(y_{t-1}, y_t, X, t))$$

where

$$Z(X) = \sum_{Y'} \exp(\sum_{t=1}^{T} \sum_{i=1}^{n} w_i f_i(y'_{t-1}, y'_t, X, t))$$

In the equation, the observation sequence $X = (x_1, \ldots, x_T)$, the label sequence $Y = (y_1, \ldots, y_T)$ where y_t is the label for position t, state feature function is concerned with the entire observation sequence, the transition feature between labels of position $t - 1$ and t on the observation sequence is also considered. Each feature f_i can either be state feature function or transition feature function. The coefficients w_is are the weights of features and can be estimated from training data. $Z(X)$ is a normalization factor.

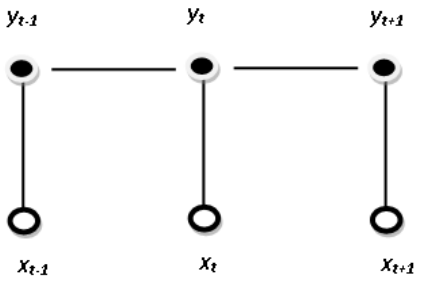

Fig. 3. Graphical representation of linear-chain CRF

4 EXPERIMENTS

4.1 *Datasets*

The performance of *XLADA* is evaluated on the unsupervised learning of Chinese and French NER for named entity recognition of three entity types, person (PER), location (LOC) and organization (ORG). To achieve this goal, unlabeled training data set and labeled test data set is required for each target language. As unlabeled training data, two bilingual parallel corpora is used. The English-Chinese corpus is 20 million parallel sentences and the English-French corpus contains 40 million parallel sentences. The corpora involve a variety of publicly available data sets including United Nations proceedings[2], proceedings of the European

[2] http://catalog.ldc.upenn.edu/LDC94T4A

Parliament[3], Canadian Hansards[4] and web crawled data. Both sides of each corpus were segmented (in Chinese) and tokenized (in English and French).

Table 1. Corpora used for performance evaluation

test set	Chinese	French
#sentences	5,633	9,988
#Person	2,807	3,065
#Location	7,079	3,153
#Organization	3,827	1,935

Table 1 shows a description of the corpora used as labeled test data for *XLADA*. One is the Chinese OntoNotes Release 2.0 corpus [5] and the second is a French corpus manually labeled using crowd sourcing. A group of five human annotators was asked to label each sentence then the majority NE tag is assigned to each token.

4.2 *Setup*

A widely used open-source NER system, Stanford Named Entity Recognizer is employed to detect named entities in the English side of the English-Chinese and English-French parallel corpora.The number of sentences that has at least one named entity detected by the Stanford NER is around 4 million sentences for Chinese and 10 million sentences for French. The features used to train the CRF model are shown in Figure 2. It's worth mentioning that the trainer used here is a local implementation of CRF (not Stanford's implementation) since Stanford's implementation is very slow and memory consuming.

As baselines for comparison, we have studied the following data selection techniques:

- *random sample*: The first NER model was trained on randomly sample of 340,000 and 400,000 sentences from the four million auto-labeled sentences and ten million sentences for Chinese and French language, respectively (upper horizontal dashed lines in Figures 4–6).

[3] http://www.statmt.org/europarl/

[4] http://www.isi.edu/natural-language/download/hansard/

[5] http://catalog.ldc.upenn.edu/LDC2008T04

Table 2. Features used for Named Entity Recognition CRF model

type	extracted features
Shape Features	WordString, WordShape, StartsWithCapital, AllCapital Character N-Grams, Shape Character N-Grams
Brown Clusters [17]	Levels 0, 4, 8, 12
Bags of Words	Date Tokens, Punctuations, Personal Titles, Stop Words
Contextual	Previous 3 words and next 2 word features

– *most confident sample*: the second NER model was trained on the set of the top 340,000 and the top 400,000 most confident sentences (based on the min-max score function defined in Section 3) for Chinese and French, respectively (lower horizontal dashed lines in Figures 4–6).

For active learning, we have randomly chosen 100,000 auto-labeled sentences to train the initial NER for Chinese and French, respectively. And then, we have created a pool (set) of two million sentences where we have two experiments:

– *random pool*: one with a pool of randomly chosen sentences regardless of tagging confidence.
– *most confident pool*: another experiment with a pool of target sentences corresponding to the most confident source sentences selected by min-max score function.

The initial NER is applied on the pool and the informativeness of each sentence is measured using the function defined in section 3.

– *informative addition*: The most informative sentences are the sentences with score less than 0.9. At the end of the first iteration, the labeled training set is augmented with the newly-selected most informative sentences and the target NER is re-trained, this process is repeated for 20 iterations where the final NER for Chinese and French has been trained on 340,000 sentences and 400,000 sentences, respectively.
– *random addition*: another baseline for comparison where in each iteration, a number of auto-labeled sentences in the target language, Chinese or French, is randomly selected, equals to the number of most informative sentences selected in the same iteration at the *informative addition* experiment.

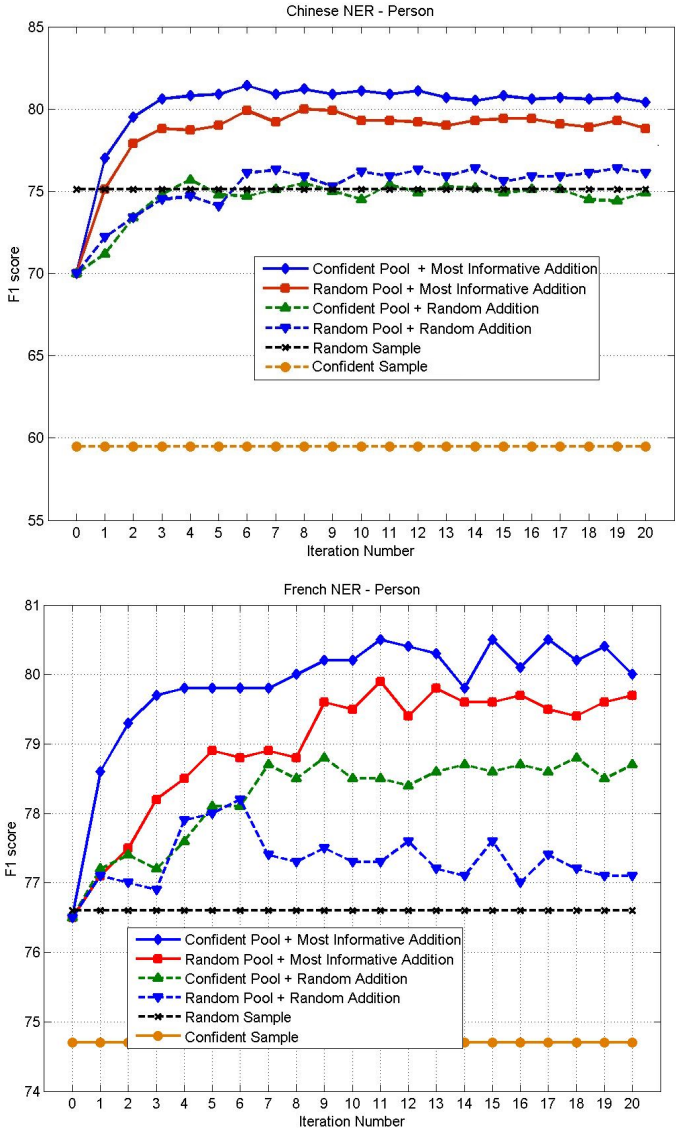

Fig. 4. Performance of unsupervised Chinese and French NER: person

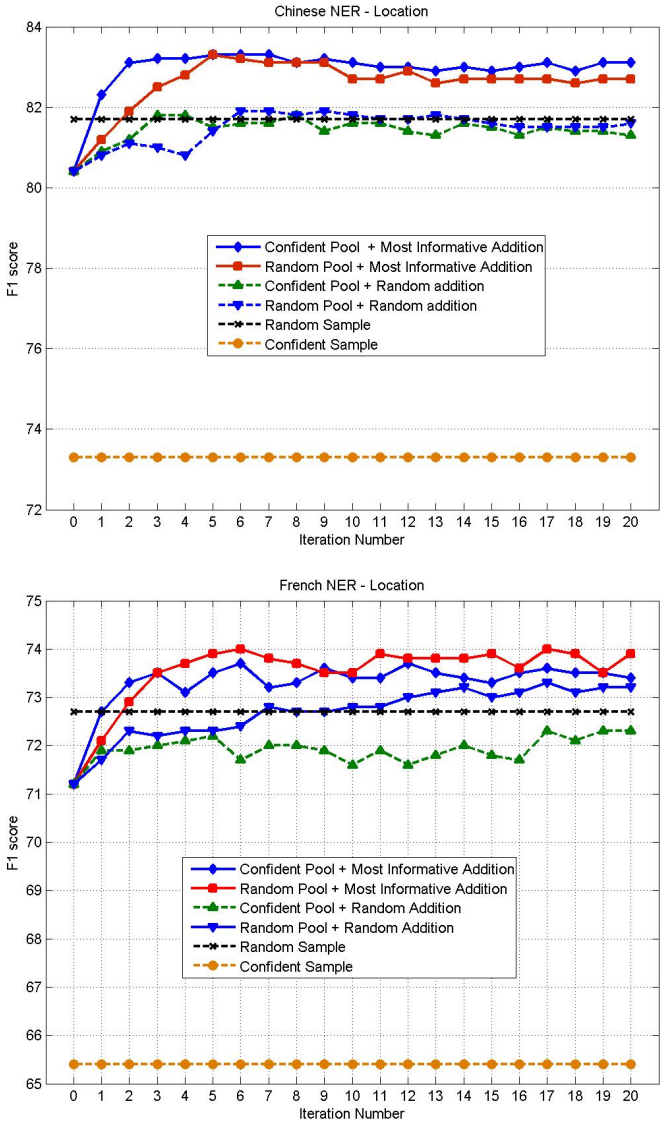

Fig. 5. Performance of unsupervised Chinese and French NER: location

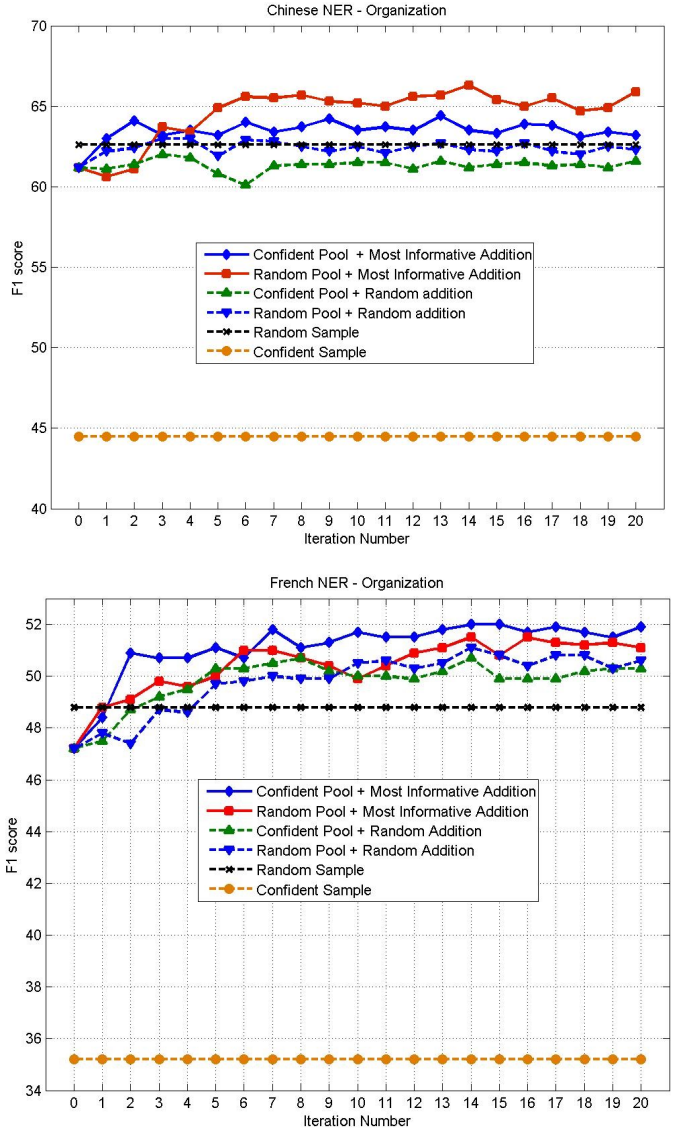

Fig. 6. Performance of unsupervised Chinese and French NER: organization

Table 3. The performance of unsupervised French NER models trained using *XLADA* compared to baselines

	selection method	XLADA		Baseline			
		informative addition		*random addition*		most	random
Entity type	Pool type	*most confident*	*random*	*most confident*	random	confident sample	sample
PER	Precision	78.1	**79.7**	78.6	79.0	77.9	77.0
	Recall	**82.0**	79.6	78.8	75.4	71.9	76.2
	F1	**80.0**	79.7	78.7	77.1	74.7	76.6
LOC	Precision	82.9	82.9	83.2	83.8	73.0	**84.6**
	Recall	65.9	**66.6**	64.0	64.9	59.2	63.8
	F1	73.4	**73.9**	72.3	73.2	65.4	72.7
ORG	Precision	54.1	50.1	52.0	51.7	**59.1**	49.5
	Recall	50.0	**52.2**	48.7	49.6	25.0	48.1
	F1	**51.9**	51.1	50.3	50.6	35.2	48.8

4.3 *Results*

The performance of unsupervised Chinese and French NER systems is reported in Table 4 and Table 3, respectively where the best performing data selection technique is bold faced. Figures 4–6 show the learning curve of target NER models using the different training data selection techniques for Chinese and French, respectively. The F1 measure of both *random sample* NER and *most confident sample* NER is drawn as a horizontal dashed line. The results show that *XLADA* outperforms the random sample baseline.

FOR CHINESE NER For person NE, *XLADA* with *informative addition* using *most confident pool* achieves the highest F1-score 80.4% compared to 59.5% for *most confident sample* and 75.1% for random sample. This is attributed to the increase in person recall from 43.6% and 63.2% to 69.7% and 68.0% respectively. For location NE, *XLADA* with *informative addition* using *most confident pool* achieves the highest F1-score 83.1% compared to 73.3% for *most confident sample* and 81.7% for random sample. This is attributed to the increase in location recall from 64.6% and 74.0% to 76.4% and 75.0% respectively. For organization NE, *XLADA* with *informative addition* using *random pool* achieves the highest F1-score 65.9% compared to 44.5% for *most confident sample* and 62.6% for random sample. This is attributed to the increase in organization recall from 29.4% and 50.3% to 55.2%, respectively.

Table 4. The performance of unsupervised Chinese NER models trained using *XLADA* compared to baselines

	selection method	XLADA		Baseline			
		informative addition		*random addition*		most	random
Entity type	Pool type	*most confident*	*random*	*most confident*	random	confident sample	sample
PER	Precision	**94.9**	93.5	92.9	91.8	93.4	92.4
	Recall	**69.7**	68.0	62.7	65.0	43.6	63.2
	F1	**80.4**	78.8	74.9	76.1	59.5	75.1
LOC	Precision	91.1	**92.2**	90.9	91.1	84.9	91.1
	Recall	**76.4**	75.0	73.6	73.9	64.6	74.0
	F1	**83.1**	82.7	81.3	81.6	73.3	81.7
ORG	Precision	80.8	81.7	87.2	78.6	**91.4**	83.0
	Recall	51.9	**55.2**	47.7	51.6	29.4	50.3
	F1	63.2	**65.9**	61.6	62.3	44.5	62.6

FOR FRENCH NER For person NE, *XLADA* with *informative addition* using *most confident pool* achieves the highest F1-score 80.0% compared to *most confident sample* with 74.7% and *random sample* with 76.6% . This is attributed to the increase in person recall from 71.9% and 76.2% to 82.0%. For location NE, *XLADA* with *informative addition* using *random pool* achieves the highest F1-score 73.9% compared to domain adaptation without active learning: *most confident sample* of 65.4% and *random sample* of 72.7%. This is attributed to the increase in location recall from 59.2% and 63.8% to 66.6%. For organization NE, *XLADA* with *informative addition* using *most confident pool* achieves the highest F1-score 51.9% compared to *most confident sample* of 35.2% and *random sample* of 48.8%. This is attributed to the significant improvement of organization recall from 25.0% and 48.1% to 50.0%.

4.4 *Discussion*

The improvement in recall means increase in the coverage of the trained NER model. This is attributed to the high quality of the training sentences selected by the proposed selective sampling criterion compared to random sampling. In addition, it is better than selecting target sentences where the English NER model is most confident about their corresponding English ones. The reason is that although the English NER model is most confident, this does not alleviate the passive nature of the target NER model as it has no control on the selection of its training data

based on its performance. That is, it implies that the selected sentences do not carry new discriminating information with respect to the target NER model. In all cases, the *random sample* outperforms the *most confident sample*. The reason that selecting only the most confident sentences tends to narrow the coverage of the constructed NER. Figures 4–6 show that *XLADA* achieves the most significant performance improvement in the early iterations, then the learning curve starts to saturate.

In general, the results of *organization* NE type are lower than the results of *Person* and *Location*. The reason is that ORG names are more complex than *Person* and *Location* names. They usually consist of more words, which may result in more word alignment errors and then lead to more training sentences being filtered out. Another reason behind this is that ORG names mostly consist of a combination of common words. Not only for French and Chinese but also English ORG entity recognition is more difficult, which also results in more noise among the ORG training sentences.

5 CONCLUSIONS

The manual annotation of training sentences to build an information extraction system for each language is expensive, error-prone and time consuming. We introduced an unsupervised variant of active learning in the cross-lingual automatic annotation framework that replaces the manual annotation with the alignment-based automatic annotation. It depends on the existence of high quality source language NER model, bilingual parallel corpus and word-level alignment model. A modified score function is proposed as the criterion for selecting the most informative training sentences from the huge amount of automatically annotated sentences. Although the reported results are on the recognition of three entity types, the framework can be generalized to any information extraction task.

REFERENCES

1. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of CoNLL. (2003)
2. Esuli, A., Marcheggiani, D., Sebastiani, F.: Sentence-based active learning strategies for information extraction. In: Proceedings of the 2nd Italian Information Retrieval Workshop (IIR 2010). (2010) 41–45

3. Jones, R., Ghani, R., Mitchell, T., Rilo, E.: Active learning for information extraction with multiple view. In: Proceedings of the European Conference in Machine Learning (ECML 2003). Volume 77. (2003) 257–286
4. Guo, Y., Xiao, M.: Transductive representation learning for cross-lingual text classification. In: Proceedings of IEEE 12th International Conference on Data Mining (ICDM 2012). (2012)
5. Prettenhofer, P., Stein, B.: Cross-lingual adaptation using structural correspondence learning. ACM Transactions on Intelligent Systems and Technology **3(1)** (2012)
6. Wan, C., Pan, R., Li, J.: Bi-weighting domain adaptation for cross-language text classification. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011). (2011)
7. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: In Human Language Technology Conference, pages 109-116. (2001)
8. Richman, A.E., Schone, P.: Mining wiki resources for multilingual named entity recognition. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. (2008) 1–9
9. Nothman, J., Curran, J.R., Murphy, T.: Transforming Wikipedia into named entity training data. In: Proceedings of the Australian Language Technology Workshop. (2008) 124–132
10. Kim, S., Toutanova, K., Yu, H.: Multilingual named entity recognition using parallel data and metadata from Wikipedia. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. (2012)
11. Fu, R., Qin, B., Liu, T.: Generating Chinese named entity data from a parallel corpus. In: Proceedings of the 5th International Joint Conference on Natural Language Processing. (2011) 264–272
12. Muslea, I., Minton, S., Knoblock, C.A.: Active learning with multiple views. Journal of Artificial Intelligence Research **27** (2006) 203–233
13. Li, Q., Li, H., Ji, H.: Joint bilingual name tagging for parallel corpora. In: Proceedings of CIKM 2012. (2012)
14. He, X.: Using word-dependent transition models in HMM based word alignment for statistical machine translation. In: Proceedings of the Second Workshop on SMT (WMT), Association for Computational Linguistics. (2007)
15. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of International Conference on Machine Learning (ICML). (2001) 282–289
16. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. In: Proceedings of the IEEE. Volume 77. (1989) 257–286
17. Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Computational Linguistics **18(4)** (1992)

MOHAMED FAROUK ABDEL HADY
MICROSOFT RESEARCH,
CAIRO, EGYPT
E-MAIL: <MOHABDEL@MICROSOFT.COM>

ABUBAKRELSEDIK KARALI
MICROSOFT RESEARCH,
CAIRO, EGYPT

ESLAM KAMAL
MICROSOFT RESEARCH,
CAIRO, EGYPT

RANIA IBRAHIM
MICROSOFT RESEARCH,
CAIRO, EGYPT