Bilingual Active Learning for Relation Classification via Pseudo Parallel Corpora

Longhua Qian   Haotian Hui   Ya'nan Hu   Guodong Zhou*   Qiaoming Zhu

Natural Language Processing Lab

School of Computer Science and Technology, Soochow University

1 Shizi Street, Suzhou, China 215006

`{qianlonghua,20134227019,20114227025,gdzhou,qmzhu}@suda.edu.cn`

Abstract

Active learning (AL) has been proven effective to reduce human annotation efforts in NLP. However, previous studies on AL are limited to applications in a single language. This paper proposes a bilingual active learning paradigm for relation classification, where the unlabeled instances are first jointly chosen in terms of their prediction uncertainty scores in two languages and then manually labeled by an oracle. Instead of using a parallel corpus, labeled and unlabeled instances in one language are translated into ones in the other language and all instances in both languages are then fed into a bilingual active learning engine as pseudo parallel corpora. Experimental results on the ACE RDC 2005 Chinese and English corpora show that bilingual active learning for relation classification significantly outperforms monolingual active learning.

1   Introduction

Semantic relation extraction between named entities (*aka*. entity relation extraction or more concisely relation extraction) is an important subtask of Information Extraction (IE) as well as Natural Language Processing (NLP). With its aim to identify and classify the semantic relationship between two entities (ACE 2002-2007), relation extraction is of great significance to many NLP applications, such as question answering, information fusion, social network construction, and knowledge mining and population etc.

In the literature, the mainstream research on relation extraction adopts statistical machine learning methods, which can be grouped into supervised learning (Zelenko et al., 2003; Culotta and Soresen, 2004; Zhou et al., 2005; Zhang et al., 2006; Qian et al., 2008; Chan and Roth, 2011), semi-supervised learning (Zhang et al., 2004; Chen et al., 2006; Zhou et al., 2008; Qian et al., 2010) and unsupervised learning (Hasegawa et al., 2004; Zhang et al., 2005) in terms of the amount of labeled training data they need. Usually the extraction performance depends heavily on the quality and quantity of the labeled data, however, the manual annotation of a large-scale corpus is labor-intensive and time-consuming. In the last decade researchers have turned to another effective learning paradigm-- active learning (AL), which, given a small number of labeled instances and a large number of unlabeled instances, selects the most informative unlabeled instances to be manually annotated and add them into the training data in an iterative fashion. Essentially active learning attempts to decrease the quantity of labeled instances by enhancing their quality, gauged by their informativeness to the learner. Since its emergence, active learning has been successfully applied to many tasks in NLP (Engelson and Dagan, 1996; Hwa, 2004; Tomanek et al., 2007; Settles and Craven, 2008).

It is trivial to validate, as we will do later in this paper, that active learning can also alleviate the annotation burden for relation extraction in one language while retaining the extraction performance. However, there are cases when we may exploit relation extraction in multiple languages and there are corpora with relation instances annotated for more than one language, such as the ACE RDC 2005 English and Chinese corpora. Hu et al. (2013) shows that supervised relation extraction in one language (e.g. Chinese)

---

can be enhanced by relation instances translated from another language (e.g. English). This demonstrates that there is some complementariness between relation instances in two languages, particularly when the training data is scarce. One natural question is: Can this characteristic be made full use of so that active learning can maximally benefit relation extraction in two languages? To the best of our knowledge, so far the issue of joint active learning in two languages has yet been addressed. Moreover, the success of joint bilingual learning may lend itself to many inherent multilingual NLP tasks such as POS tagging (Yarowsky and Ngai, 2001), name entity recognition (Yarowsky et al., 2001), sentiment analysis (Wan, 2009), and semantic role labeling (Sebastian and Lapata, 2009) etc.

This paper proposes a bilingual active learning (BAL) paradigm to relation classification with a small number of labeled relation instances and a large number of unlabeled instances in two languages (non-parallel). Instead of using a parallel corpus which should have entity/relation alignment information and is thus difficult to obtain, this paper employs an off-the-shelf machine translator to translate both labeled and unlabeled instances from one language into the other language, forming pseudo parallel corpora. These translated instances along with the original instances are then fed into a bilingual active learning engine. Findings obtained from experiments with relation classification on the ACE 2005 corpora show that this kind of pseudo-parallel corpora can significantly improve the classification performance for both languages in a BAL framework.

The rest of the paper is organized as follows. Section 2 reviews the previous work on relation extraction while Section 3 describes our baseline systems. Section 4 elaborates on the bilingual active learning paradigm and Section 5 discusses the experimental results. Finally conclusions and directions for future work are presented in Section 6.

2 Related Work

While there are many studies in monolingual relation extraction, there are only a few on multilingual relation extraction in the literature.

Monolingual relation extraction: A wide range of studies on relation extraction focus on monolingual resources. As far as representation of relation instances is concerned, there are feature-based methods (Zhao et al., 2004; Zhou et al., 2005; Chan and Roth, 2011) and kernel-based methods (Zelenko et al., 2003; Zhang et al., 2006; Qian et al., 2008), mainly for the English language. Both methods are also widely used in relation extraction in other languages, such as those in Chinese relation extraction (Che et al., 2005; Li et al., 2008; Yu et al., 2010).

Multilingual relation extraction: There are only two studies related to multilingual relation extraction. Kim et al. (2010) propose a cross-lingual annotation projection approach which uses parallel corpora to acquire a relation detector on the target language. However, the mapping of two entities involved in a relation instance may leads to errors. Therefore, Kim and Lee (2012) further employ a graph-based semi-supervised learning method, namely Label Propagation (LP), to indirectly propagate labels from the source language to the target language in an iterative fashion. Both studies transfer relation annotations via parallel corpora from the resource-rich language (English) to the resource-poor language (Korean), but not vice versa. Based on a small number of labeled instances and a large number of unlabeled instances in both languages, our method differs from theirs in that we adopt a bilingual active learning paradigm via machine translation and improve the performance for both languages simultaneously.

Active Learning in NLP: Active learning has become an active research topic due to its potential to significantly reduce the amount of labeled training data while achieving comparable performance with supervised learning. It has been successfully applied to many NLP applications, such as POS tagging (Engelson and Dagan, 1996; Ringger et al., 2007), word sense disambiguation (Chan and Ng, 2007; Zhu and Hovy, 2007), sentiment detection (Brew et al., 2010; Li et al., 2012), syntactical parsing (Hwa, 2004; Osborne and Baldridge, 2004), and named entity recognition (Shen et al., 2004; Tomanek et al., 2007; Tomanek and Hahn, 2009) etc.

Different from these AL studies on a single task, Reichart et al. (2008) introduce a multi-task active learning (MTAL) paradigm, where unlabeled instances are selected for two annotation tasks (i.e. named entity and syntactic parse tree). They demonstrate that MTAL in the same language outperforms one-sided and random selection AL. From a different perspective, we propose an active learning framework for the same task, but across two different languages.

Another related study (Haffari and Sarkar, 2009) deals with active learning for multilingual

machine translation, which make use of multilingual corpora to decrease human annotation efforts by selecting highly informative sentences for a newly added language in multilingual parallel corpora. While machine translation inherently deals with multilingual parallel corpora, our task focuses on relation extraction by pseudo parallel corpora in two languages.

3 Baseline Systems

This section first introduces the fundamental supervised learning method, and then describes a baseline active learning algorithm.

3.1 Supervised Learning

We adopt the feature-based method for fundamental supervised relation classification, rather than the tree kernel-based method, since active learning needs a large number of iterations and the kernel-based method usually performs much slower than the feature-based one. Following is a list of our used features, much similar to Zhou et al. (2005):

a) Lexical features of entities and their contexts
   WM1: bag-of-words in the 1st entity mention
   HM1: headword of M1
   WM2: bag-of-words in the 2nd entity mention
   HM2: headword of M2
   HM12: combination of HM1 and HM2
   WBNULL: when no word in between
   WBFL: the only one word in between
   WBF: the first word in between when at least two words in between
   WBL: the last word in between when at least two words in between
   WBO: other words in between except the first and last words when at least three words in between
b) Entity type
   ET12: combination of entity types
   EST12: combination of entity subtypes
   EC12: combination of entity classes
c) Mention level
   ML12: combination of entity mention levels
   MT12: combination of LDC mention types
d) Overlap
   #WB: number of other mentions in between
   #MB: number of words in between
  M1>M2 or M1<M2: flag indicating whether M2/M1 is included in M1/M2.

3.2 Active Learning Algorithm

We use a pool-based active learning procedure with uncertainty sampling (Scheffer et al., 2001;

Culotta and McCallum, 2005; Kim et al., 2006) for both Chinese and English relation classification as illustrated in Fig. 1. During iterations a batch of unlabeled instances are chosen in terms of their informativeness to the current classifier, labeled by an oracle and in turn added into the labeled data to retrain the classifier. Due to our focus on the effectiveness of bilingual active learning on relation classification, we only use uncertainty sampling without incorporating more complex measures, such as diversity and representativeness (Settles and Craven, 2008), and leave them for future work.

Algorithm uncertainty-based active learning

Input:
  - *L*, labeled data set
  - *U,* unlabeled data set
  - *n*, batch size
Output:
  - *SVM*, classifier
Repeat:
  1. Train a single classifier *SVM* on *L*
  2. Run the classifier on *U*
  3. Find at most *n* instances in *U* that the classifier has the highest prediction uncertainty
  4. Have these instances labeled by an oracle
  5. Add them into *L*
Until: certain number of instances are labeled or certain performance is reached

---

Figure 1. Pool-based active learning with uncertainty sampling

Since the SVMLIB package used in this paper can output probabilities assigned to the class labels on an instance, we have three uncertainty metrics readily available, i.e., least confidence (**LC**), margin (**M**) and entropy (**E**). The NER experimental results on multiple corpora (Settles and Craven, 2008) show that there is no single clear winner among these three metrics. This conclusion is also validated by our preliminary experiments on the task of active learning relation extraction, thus we adopt the **LC** metric for simplicity. Specifically, with a sequence of K probabilities for a relation instance at some iteration, denoted as $\{p_1, p_2, \ldots p_K\}$ in the descending order, the **LC** metric of the relation instance can be simply picked as the first one, i.e.

$$H^{LC} = p_1 \qquad (1)$$

Where K denotes the total number of relation classes. Note that this metric actually reflects prediction reliability (i.e. reverse uncertainty) rather than uncertainty in order to facilitate joint

confidence calculation for two languages (cf. §4.4). Intuitively, the smaller the H^{LC} is, the less confident the prediction is.

4 Bilingual Active Learning for Relation Classification

In this section, we elaborate on the bilingual active learning for relation extraction.

4.1 Problem Definition

With Chinese and English (designated as c and e) as two languages used in our study, this paper intends to address the task of bilingual relation classification, i.e., assigning relation labels to candidate instances that have semantic relationships. Suppose we have a small number of labeled instances in both languages, denoted as L_c and L_e (non-parallel) respectively, and a large number of unlabeled instances in both languages, denoted as U_c and U_e (non-parallel). The test instances in both languages are represented as T_c and T_e. In order to take full advantage of bilingual resources, we translate both labeled and unlabeled instances in one language to ones in the other language as follows:

$$L_c \rightarrow L_{et}$$
$$U_c \rightarrow U_{et}$$
$$L_e \rightarrow L_{ct}$$
$$U_e \rightarrow L_{ct}$$

The objective is to learn SVM classifiers in both languages, denoted as SVM_c and SVM_e respectively, in a BAL fashion to improve their classification performance.

4.2 Bilingual Active Learning Framework

Currently, AL is widely used in NLP tasks in a single language, i.e., during iterations unlabeled instances least confident only in one language are picked and manually labeled to augment the training data. The only exception is AL for machine translation (Haffari et al., 2009; Haffari and Sarkar, 2009), whose purpose is to select the most informative sentences in the source language to be manually translated into the target language. Previous studies (Reichart et al., 2008; Haffari and Sarkar, 2009) show that multi-task active learning (MTAL) can yield promising overall results, no matter whether they are two different tasks or the task of machine translation on multiple language pairs. If a specific NLP task on two languages, such as relation classification, can be regarded as two tasks, it is reasonable to argue that these two tasks can benefit

each other when jointly performed in the BAL framework. Yet, to our knowledge, this issue remains unexplored.

An important issue for bilingual learning is how to obtain two language views for relation instances from multilingual resources. There are three solutions to this problem, i.e. parallel corpora (Lu et al., 2011), translated corpora (*aka.* pseudo parallel corpora) (Wan 2009), and bilingual lexicons (Oh et al., 2009). We adopt the one with pseudo parallel corpora, using the machine translation method to generate instances from one language to the other in the BAL paradigm, as depicted in Fig. 2.
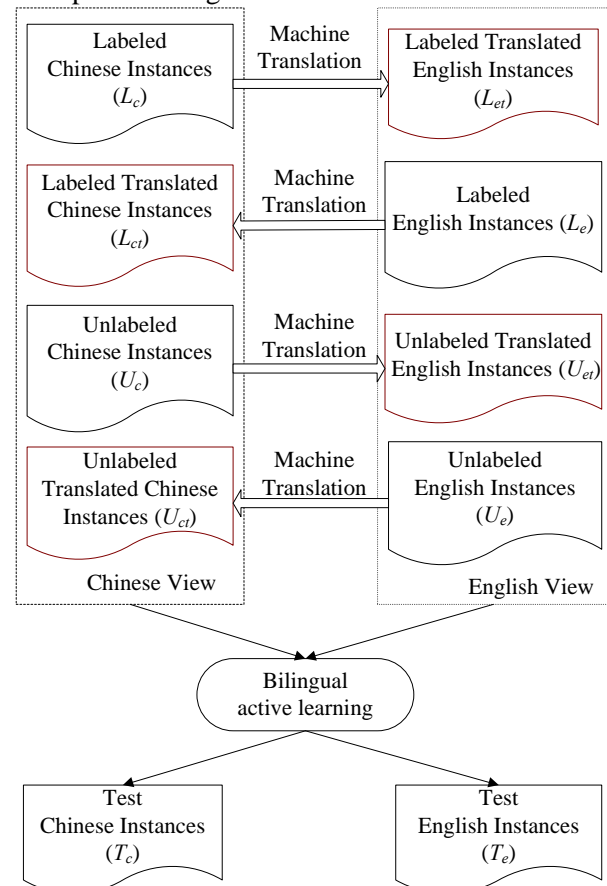

Figure 2. Framework of bilingual active learning

In order to make full use of pseudo parallel corpora, translated labeled and unlabeled instances are augmented in the following two ways:

- For labeled Chinese instances (L_c) and English instances (L_e), their translated counterparts (L_{et} and L_{ct}), along with their labels, are directly added into the labeled instances in the other language;
- For unlabeled Chinese instances (U_c) and English instances (U_e), during an active learning iteration the top n unlabeled instances in U_c and U_{et} which are least confidently jointly

predicted by SVM_c and SVM_e are labeled by an oracle and added to L_c and L_e respectively. (cf. §4.4)

4.3 Instance Projection via MT

Among the several off-the-shelf machine translation services, we select the Google Translator[1] because of its high quality and easy accessibility. Both the mentions of relation instances and the mentions of two involved entities are first translated into the other language via machine translation. Then, two entities in the original instance are aligned with their counterparts in the translated instance in order to form an aligned bilingual relation instance pair.

Instance translation

All the positive instances in the ACE 2005 Chinese and English corpora are translated to another language respectively, i.e. Chinese to English and vice versa. The relation instance is represented as the word sequence between two entities. This word sequence, rather than the whole sentence, is then translated to another language by the Google Translator. The reason is that, although this sequence loses partial contextual information of the relation instance, its translation quality is supposed to be better. Our preliminary experiments indicate that the addition of contextual information fail to benefit the task. After translation, word segmentation is performed on Chinese instances translated from English while tokenization is needed for translated English instances.

Entity alignment

The objective of entity alignment is to build a mapping from the entities in the original instances to the entities in the translated instances. Put in another way, entity alignment automatically marks the entity mentions in the translated instance, thereby the feature vector corresponding to the translated instance can be constructed. Entity alignment is vital in cross-language relation extraction whose difficulty lies in the fact that the same entity mention as an isolated phrase and as an integral phrase in the relation instance can be translated to different phrases. For example, the Chinese entity mention "官员" (*officer*) is translated to "officer" in isolation, it is, however, translated to "officials" when in the relation instance "叙利亚 官员" (*Syrian officials*).

Algorithm entity alignment

Input:
- M_e, entity mention in English
- R_e, relation instance in English
- M_{ct}, translation of M_e in Chinese
- R_c, translation of R_e in Chinese
- L, a lexicon consisting of entries like (e_i, c_i, p_i), where p_i is the translation probability from e_i to c_i
- α, probability threshold

Output:
- M_c, the counterpart of M_e in R_c

Steps:
1. If M_{ct} can be exactly found in R_c, then return M_{ct}
2. If the rightmost part of M_{ct} can be found in R_c, then this part can be returned
3. For very word w_e in M_e,
   a) If there exists a word w_c in R_c and (w_e, w_c, p) in L and $p>\alpha$, then (w_e, w_c) is a match of two words
   b) Return a successive sequence of matching words w_c
4. Return null

Figure 3. Entity alignment algorithm

Therefore, we devise some heuristics to align entity mentions between Chinese and English. The basic idea is that the word sequence in one mention successively matches the word sequence in the other mention. Take entity alignment from English to Chinese as an example, given entity mention M_e in relation instance R_e in English and their respective translations M_{ct} and R_c in Chinese, the objective of entity alignment is to find M_c, the counterpart of M_e in R_c. The procedure of entity alignment algorithm can be described in Fig. 3.

In the algorithm, the probability threshold α is empirically set to 0.002 where the precision and recall of entity alignment are balanced. Our lexicon is derived from the FBIS parallel corpus (#LDC2003E14), which is widely used in machine translation between English and Chinese. It should be noted that the process of relation translation and entity alignment are far from perfection, leading to reduction in the number of instances being mapping to the other language, i.e.

$$|L_c| > |L_{et}|$$
$$|U_c| > |U_{et}|$$
$$|L_e| > |L_{ct}|$$
$$|U_e| > |L_{ct}|$$

4.4 Bilingual Active Learning Algorithm

The basic idea of our BAL paradigm is that, while unlabeled instances uncertain in one lan-

guage are informative to the learner in that language, unlabeled instances jointly uncertain in both languages are informative to the learners in both languages, thus potentially improving classification performance for both languages more than their individual active learners do. This idea is embodied in the BAL algorithm in Fig. 4, where n is the batch size, i.e., the number of instances selected, labeled and augmented at each iteration.

Algorithm bilingual active learning

Input:
   - L_c and U_c, labeled and unlabeled instances in Chinese, and L_{et} and U_{et}, their respective translation counterparts in English
   - L_e and U_e, labeled and unlabeled instances in English, and L_{ct} and U_{ct}, their translation counterparts in Chinese
   - n, batch size

Output:
   - SVM_c and SVM_e, two classifiers for Chinese and English respectively

Initialize:
   1. Add instances L_{ct} to L_c
   2. Add instances L_{et} to L_e

Repeat:
   1. Learn the Chinese classifier SVM_c from L_c
   2. Use SVM_c to classify instances in U_c and U_{ct}
   3. Learn the English classifier SVM_e from L_e
   4. Use SVM_e to classify instances in U_e and U_{et}
   5. Choose the n least confidently jointly predicted instance pairs $\{E_c|E_{et}\}$ from $\{U_c|U_{et}\}$, and have them labeled by an oracle
   6. Choose the n least confidently jointly predicted instance pairs $\{E_e|E_{et}\}$ from $\{U_e|U_{ct}\}$, and have them labeled by an oracle
   7. Remove E_c from U_c and E_e from U_e
   8. Add instances $E_c \cup E_{ct}$ to L_c with their manual labels
   9. Add instances $E_e \cup E_{et}$ to L_e with their manual labels

Until certain number of instances are labeled or certain performance is reached

---

Figure 4. Bilingual active learning algorithm

The key point of this algorithm lies in Step 5 and Step 6, where unlabeled instances from U_c and U_e are selected and labeled respectively. Take Chinese for an example, when gauging the prediction uncertainty for an unlabeled instance in U_c, not only its own uncertainty measure H_c predicted by SVM_c is considered, but also the uncertainty measure H_{et} for its translation counterpart in U_{et}, which is predicted by SVM_e, is considered. Generally, in order to jointly consider these two measures, there are three methods to compute their means, namely, arithmetic mean, geometric mean and harmonic mean. Preliminary experiments show that among these three means, there is no single winner, so we simply take the geometric mean defined as follows:

$$H_g = \sqrt{H_c * H_{et}} \qquad (2)$$

Considering that we adopt the **LC** measure as the uncertainty score, when an instance in U_c can't find its translation counterpart in U_{et} due to translation error or entity alignment failure, H_{et} is set to 1, i.e. the maximum. Since the bigger H is, the more confident the prediction is, the less likely the instance will be chosen, in this way we discourage the unlabeled instances without translation counterparts.

5 Experimentation

We have systematically evaluated our BAL paradigm on the relation classification task using ACE RDC 2005 RDC Chinese and English corpora.

5.1 Experimental Settings

Corpora and Preprocessing

We use the ACE 2005 RDC Chinese and English corpora as the benchmark data (hereafter we refer to them as the Chinese corpus (ACE2005c) and the English corpus (ACE2005e) respectively). Both corpora have the same entity/relation hierarchies, which define 7 entity types, 6 major relation types. However, the Chinese corpus contains 633 documents and 9,147 positive relation instances while the English corpus only contains 498 files and 6,253 positive instances. Therefore, in order to balance the corpus scale to fairly evaluate bilingual active learning impact on relation classification, we randomly select 458 Chinese files and thus get 6,268 positive instances, comparable to the English corpus.

Preprocessing steps for both corpora include sentence splitting and tokenization (word segmentation for Chinese using ICTCLAS[2]). Then, positive relation mentions with word sequences between two entities and their feature vectors are extracted from sentences while negative relation mentions are simply discarded because we focus on the task of relation classification. After entity and relation mentions in one language are trans-

---

[2] http://ictclas.org/

lated into the other language using the Google translator, entity alignment is performed between relation mentions and their translations. Finally 4,747 Chinese relation mentions are successfully translated and aligned from English and vice versa, 4,936 English relation mentions are translated and aligned from Chinese.

SVMLIB (Chang and Lin, 2011) is selected as our classifier since it supports multi-class classification. The training parameters C (SVM) is set to 2.4 according to our previous work on relation extraction (Qian et al., 2010). Relation classification performance is evaluated using the standard Precision (P), Recall (R) and their harmonic average (F1) as well as *deficiency* measure (cf. latter in this section.). Overall performance scores are averaged over 10 runs. For each run, 1/40 and 1/5 randomly selected instances are used as the training and test set respectively while the remaining instances are used as the unlabeled set for further labeling during active learning iterations.

Methods for Comparison

For fair comparison, two baseline methods of supervised learning are included to augment their training sets with labeled instances during iterations. However, these labeled instances are chosen randomly from the corpus.

SL-MO (Supervised Learning with monolingual labeled instances): only the monolingual labeled instances are fed to the SVM classifiers for both Chinese and English relation classification respectively. The initial training data only contain L_c and L_e for Chinese and English respectively.

SL-CR (Supervised Learning with cross-lingual labeled instances): in addition to monolingual labeled instances (**SL-MO**), the training data for supervised learning contain labeled instances translated from the other language. That is, the initial training data contain L_c and L_{ct} for Chinese, or L_e and L_{et} for English. More important, at each iteration not only the labeled instances are added to the training data of its own language, but their translated instances are also added to the training data of the other language.

AL-MO (Active Learning with monolingual instances): labeled and unlabeled data for active learning only contain monolingual instances. No translated instances are involved. That is, the data contain L_c and U_c for Chinese, or L_e and U_e for English respectively. This is the normal active learning method applied to a single language.

AL-CR (Active Learning with cross-lingual instances): both the manually labeled instances and their translated ones are added to the respective training data. The initial training data contain L_c and L_{ct} for Chinese, or L_e and L_{et} for English. At each iteration, the n least confidently classified instances in U_c and U_e are labeled and added to the Chinese/English training data respectively. Their translated instances in U_{et} and U_{ct} are also added to the English/Chinese training data respectively.

AL-BI (Active Learning with bilingual labeled and unlabeled instances): similar to **AL-CR** with the exception that the unlabeled instances are chosen not by uncertainty scores in one language, but by the joint uncertainty scores in two languages. (cf. §4.4)

Evaluation Metric

Although learning curves are often used to evaluate the performance for active learning, it is preferable to quantitatively compare various active learning methods using a statistical metric *deficiency* (Schein and Ungar, 2007) defined as:

$$def_n(AL, REF) = \frac{\sum_{i=1}^{n}(F_n(REF) - F_i(AL))}{\sum_{i=1}^{n}(F_n(REF) - F_i(REF))} \quad (3)$$

Where n is the number of iterations involved in active learning and F_i is the F1-score of relation classification at the ith iteration. *REF* is the baseline active learning method and *AL* is an improved variant of *REF*, such as **AL-CR** or **AL-BI**. Essentially this *deficiency* metric measures the degree to which *REF* outperforms *AL*. Thus, smaller deficiency value (i.e. <1.0) indicates AL outperforms *REF* while a larger value (i.e. >1.0) indicates *AL* underperforms *REF*.

5.2 Experimental Results and Analysis

Comparison of overall deficiency

Table 1 compares the deficiency scores of relation classification on the Chinese (ACE2005c) and English corpora (ACE2005e) for various learning methods, i.e., **SL-CR**, **AL-MO**, **AL-CR** and **AL-BI**. Particularly, **SL-MO** is used as the baseline system against which deficiency scores for other methods are computed. The batch size n is set to 100 and iterations stop after all the unlabeled instances have run out of. Deficiency scores are averaged over 10 runs and the best ones are highlighted in bold font. Each run has a different test set and a different seed set.
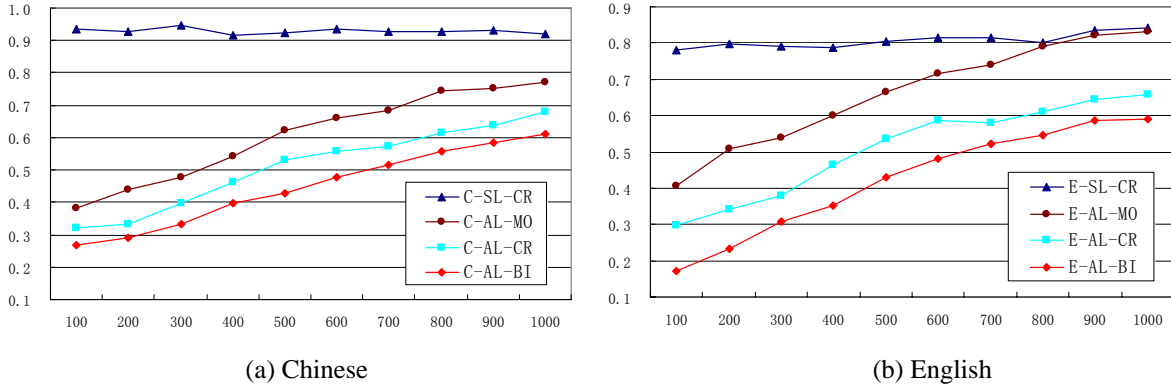
(a) Chinese

(b) English

Figure 5. Deficiency comparison for different batch sizes
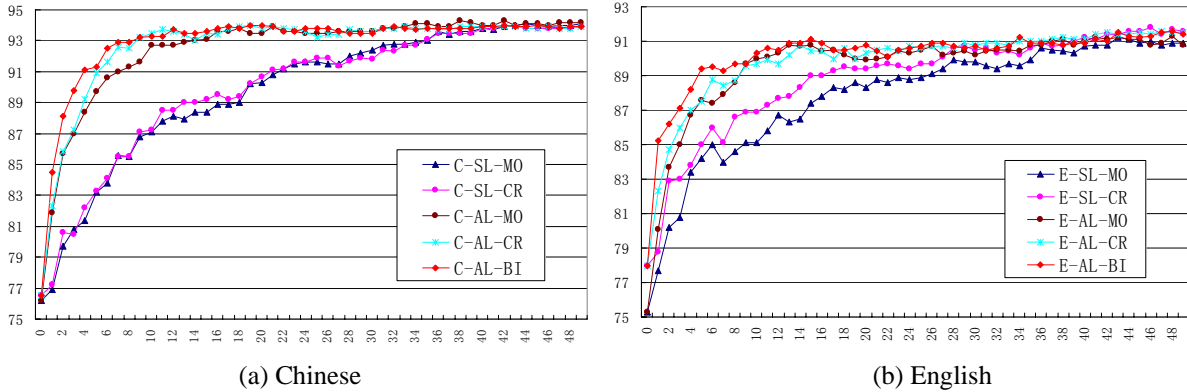

(a) Chinese

(b) English

Figure 6. Learning curves for different methods

The table shows that among the three active learning methods, bilingual active learning (**AL-BI**) achieves the best performance for both Chinese and English relation classification. This demonstrates that, bilingual active learning with jointly selecting the unlabeled instances can not only enhance relation classification for its own language, but also help relation classification for the other language due to the complementary nature of relation instances between Chinese and English.

Corpora	SL-CR	AL-MO	AL-CR	AL-BI
ACE2005c	0.934	0.383	0.323	**0.254**
ACE2005e	0.779	0.405	0.298	**0.160**

Table 1. Deficiency comparison of different methods

The table also shows the consistent utility of cross-lingual information for relation classification for both languages. When cross-lingual information is augmented, **SL-CR** outperforms **SL-MO** and **AL-CR** outperforms **AL-MO**.

Comparison of different batch sizes

Figure 5(a) and 5(b) illustrate the deficiency scores for four learning methods (**SL-CR, AL-**MO, **AL-CR** and **AL-BI**) against the **SL-MO** method with different batch sizes (n), where prefixes "C" and "E" denote Chinese and English respectively. The horizontal axes denote the range of n ($<=1000$) while the vertical ones denote the deficiency scores.

The figures show that the deficiency scores for three active learning methods run virtually parallel with each other while they increase monotonously with the batch size n. This suggests that for both Chinese and English **AL-BI** consistently performs best against other methods across a wide range of batch sizes, though the overall advantage of three active learning methods generally diminish.

Comparison of learning curves

In order to gain an intuition into how the performance evolves when the labeled instances are added into the training data during iterations, we depict the learning curves for various learning methods on the Chinese and English corpora in Fig. 6(a) and 6(b) respectively. The horizontal axes denote learning iterations while the vertical ones denote F1-scores. For simplicity of illustration the F1-scores are collected from one of the 10 runs.

The figures clearly demonstrate the performance difference for both languages among five methods at the beginning of iterations while F1-scores converge at the end of iterations. Particularly at the very outset, **AL-BI** outperforms other methods, quickly jumps to a very high point comparable to its best performance. However, after the 10th iteration the performance scores for the three AL variants tend to show trivial difference probably because most highly informative instances have already been added to the training data.

Comparison of annotation scale

In order to better compare BAL with other AL methods Figure 7 zooms out partial data on three AL methods in Fig. 6 and rescale the data for **AL-MO**, where "C" and "E" denote Chinese and English respectively. Likewise, the vertical axis denotes F1-scores while the horizontal axis denotes the number of instances labeled for **AL-CR** and **AL-BI**. However, for **AL-MO** that number is doubled. This figure tries to answer the question: to label n respective instances in both languages for BAL or to labeled $2n$ instances in just one language for monolingual AL, can the former rival the latter?
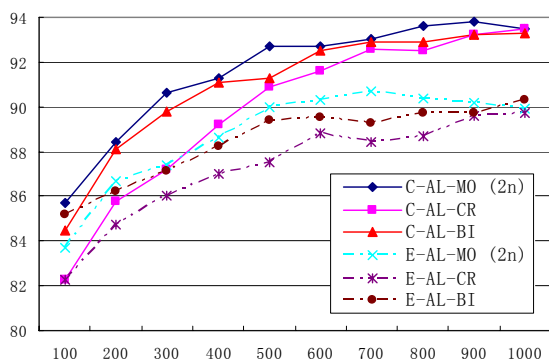

Figure 7. Comparison of annotation scale among three AL methods

The figure shows that for both Chinese and English, when the number of instances (n) to be labeled is no greater than 400, **AL-BI** with n instances can achieve comparable performance with **AL-MO** with $2n$ instances. It implies that when the labeled instances are limited, labeling instances, half in one language and half in the other for BAL, is competitive against labeling the same total number of instances in just one language for monolingual AL, not to mention that the former can generate two relation extractors on two languages.

6   Conclusion

This paper proposes a bilingual active learning paradigm for Chinese and English relation classification. Given a small number of relation instances and a large number of unlabeled relation instances in both languages, we translate both the labeled and unlabeled instances in one language to the other as pseudo parallel corpora. After entity alignment, these labeled and unlabeled instances in both languages are fed into a bilingual active learning engine. Experiments with the task of relation classification on the ACE RDC 2005 Chinese and English corpora show that bilingual active learning can significantly outperforms monolingual active learning for both Chinese and English simultaneously. Moreover, we demonstrate that BAL across two languages can compete against monolingual AL when the annotation scale is limited, though the overall number of labeled instances remains the same.

For future work, on one hand, we plan to combine uncertainty sampling with diversity and informativeness measures; on the other hand, we intend to combine BAL with semi-supervised learning to further reduce human annotation efforts.

Reference

ACE. 2002-2007. Automatic Content Extraction. http://www.ldc.upenn.edu/Projects/ACE/

A. Brew, D. Greene, and P. Cunningham. 2010. Using crowdsourcing and active learning to track sentiment in online media. *ECAI'2010*: 145–150.

Y.S. Chan and D. Roth. 2011. Exploiting Syntactico-Semantic Structures for Relation Extraction. *ACL'2011*: 551-560

Y.S. Chan and H.T. Ng. 2007. Domain adaptation with active learning for word sense disambiguation. *ACL'2007*.

C.C. Chang and C.J. Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27):1-27.

W.X. Che, T. Liu, and S. Li. 2005. Automatic Extraction of Entity Relation (in Chinese). *Journal of Chinese Information Processing*, 19(2): 1-6.

J.X. Chen, D.H. Ji, and C. L. Tan. 2006. Relation Extraction using Label Propagation-based Semi-supervised Learning. *ACL/COLING'2006*: 129-136.

A. Culotta and J. Sorensen. 2004. Dependency tree kernels for relation extraction. *ACL'2004*: 423-439.

A. Culotta and A. McCallum. 2005. Reducing labeling effort for stuctured prediction tasks. *AAAI'2005*: 746–751.

S. P. Engelson and I. Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. ACL'1996: 319–326.

G. Haffari, M. Roy, and A. Sarkar. 2009. Active learning for statistical phrase-based machine translation. *NAACL'2009*: 415–423.

G. Haffari and A. Sarkar. 2009. Active learning for multilingual statistical machine translation. *ACL/IJCNLP'2009*: 181–189.

T. Hasegawa, S. Sekine, and R. Grishman. 2004. Discovering Relations among Named Entities from Large Corpora. *ACL'2004*.

Y.N. Hu, J.G. Shu, L.H. Qian, and Q.M. Qiao. 2013. Cross-lingual Relation Extraction based on Machine Translation (in Chinese). *Journal of Chinese Information Processing*, 27(5): 191-197.

R. Hwa. 2004. Sample selection for statistical parsing. *Computational Linguistics*, 30(3): 253–276.

S. Kim, M. Jeong, J. Lee, and G.G. Lee. 2010. A Cross-lingual Annotation Projection Approach for Relation Detection. *COLING'2010*: 564-571.

S. Kim and G.G. Lee. 2012. A Graph-based Cross-lingual Projection Approach for Weakly Supervised Relation Extraction. *ACL'2012*: 48-53.

S. Kim, Y. Song, K. Kim, J.W. Cha, and G.G. Lee. 2006. MMR-based active machine learning for bio named entity recognition. *HLT-NAACL'2006*: 69–72.

W.J. Li, P. Zhang, F.R. Wei, Y.X. Hou, and Q. Lu. 2008. A Novel Feature-based Approach to Chinese Entity Relation Extraction. *ACL'2008*: 89-92.

S.S. Li, S.F. Ju, G.D. Zhou, and X.J. Li. 2012. Active learning for imbalanced sentiment classification. *EMNLP-CoNLL'2012*: 139-148.

B. Lu, C.H. Tan, C. Cardie, and B.K. Tsou. 2011. Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. *ACL'2011*: 320-330.

J. Oh, K. Uchimoto, and K. Torisawa. 2009. Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition. *ACL'2009*: 432-440.

M. Osborne and J. Baldridge. 2004. Ensemble based active learning for parse selection. *HLT-NAACL'2004*: 89–96.

L.H. Qian, G.D. Zhou, F. Kong, and Q.M. Zhu. 2010. Clustering-based Stratified Seed Sampling for Semi-Supervised Relation Classification. *EMNLP2010*: 346-355.

L.H. Qian, G.D. Zhou, Q.M. Zhu, and P.D. Qian. 2008. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. *COLING'2008*: 697-704.

R. Reichart, K. Tomanek, U. Hahn, and A. Rappoport. 2008. Multi-task active learning for linguistic annotations. *ACL'2008*: 861-869.

E. Ringger, P. McClanahan, R. Haertel, G. Busby, M. Carmen, J. Carroll, K. Seppi, and D. Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop at ACL'2007*: 101–108.

T. Scheffer, C. Decomain, and S. Wrobel. 2001. Active hidden Markov models for information extraction. In Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA), pages 309–318.

A. I. Schein and L. H. Ungar. 2007. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3): 235-265.

P. Sebastian and M. Lapata. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*, 36(1): 307-340.

B. Settles and M. Craven. 2008. An Analysis of Active Learning Strategies for Sequence Labeling Tasks. *EMNLP'2008*: 1070–1079.

D. Shen, J. Zhang, J. Su, G.D. Zhou and C.-L. Tan. 2004. Multi-criteria-based active learning for named entity recognition. *ACL'2004*.

K. Tomanek and U. Hahn. 2009. Semi-Supervised Active Learning for Sequence Labeling. *ACL-IJCNLP'2009*: 1039-1047.

K. Tomanek, J. Wermter, and U. Hahn. 2007. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. *EMNLP-CoNLL'2007*: 486–495.

X.J. Wan. 2009. Co-Training for Cross-Lingual Sentiment Classification. *ACL-AFNLP'2009*: 235-243.

D. Yarowsky and G. Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. *NAACL'2001*: 1-8.

D. Yarowsky, G. Ngai, and R. Wicentorski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. *HLT'2001*:1-8.

H.H. Yu, L.H. Qian, G.D. Zhou, and Q.M. Zhu. 2010. Chinese Semantic Relation Extraction based on Unified Syntactic and Entity Semantic Tree (in Chinese). *Journal of Chinese Information Processing*, 24(5): 17-23.

D. Zelenko, C. Aone, and A. Richardella. 2003. Kernel Methods for Relation Extraction. *Journal of Machine Learning Research*, 3: 1083-1106.

Z. Zhang. 2004. Weakly-supervised relation classification for Information Extraction. *CIKM'2004*.

M. Zhang, J. Su, D. M. Wang, G. D. Zhou, and C. L. Tan. 2005. Discovering Relations between Named Entities from a Large Raw Corpus Using Tree Similarity-Based Clustering. *IJCNLP'2005*: 378-389.

M. Zhang, J. Zhang, J. Su, and G.D. Zhou. 2006. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features. *ACL/COLING'2006*: 825-832.

S.B. Zhao and R. Grishman. 2005. Extracting relations with integrated information using kernel methods. *ACL'2005*: 419-426.

G.D. Zhou, J.H. Li, L.H. Qian, and Q.M. Zhu. 2008. Semi-Supervised Learning for Relation Extraction. *IJCNLP'2008*: 32-38.

G.D. Zhou, J. Su, J. Zhang, and M. Zhang. 2005. Exploring various knowledge in relation extraction. *ACL'2005*: 427-434.

J.B. Zhu and E. Hovy. 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. *EMNLP-CoNLL'2007*: 783-790.