Learning from Chinese-English Parallel Data for Chinese Tense Prediction

Feifan Liu
The University of Wisconsin-Milwaukee
liuf@uwm.edu

Fei Liu and Yang Liu
The University of Texas at Dallas
feiliu,yangl @hlt.utdallas.edu

Abstract

Tense prediction can be useful for many language processing tasks, such as temporal inference and machine translation. In this paper, we investigate using diverse contextual features for Chinese tense prediction under a statistical learning framework. Because of lack of annotated training data, we propose to leverage Chinese-English parallel corpora to automatically generate reference tense for model training. We also propose to use an iterative learning framework to deal with the noisy reference data to improve learning. Evaluation is performed using both automatically generated reference data and a manually annotated set with verb tense. Our results demonstrate the effectiveness of our proposed learning framework that maps annotation from one language to another using parallel data. Furthermore, we show better performance using our proposed iterative bootstrapping learning method compared to using the original automatically created training data.

1 Introduction

Tense is used in languages to indicate the time at which an action or event described by the sentence takes place. Lacking correct tense information can cause confusion and misunderstanding in communication. Predicting tense information is very useful for different natural language processing tasks: both monolingual applications such as speech recognition (Karlgren, 1996) and multilingual applications such as machine translation (MT) (Buschbeck et al., 1991).

In inflectional languages like English, tense is often expressed by verb inflections, which can be easily recognized. For example, "*I worked till 5pm yesterday*" uses the past-tense verb "*worked*" to describe an event in the past. However, in languages such as Chinese, no verb inflections exist to indicate any tense information (Xiao and McEnery, 2002). The morphology of a verb itself never changes when used to express different tenses. The following examples (1a) and (1b) indicate a past-tense sentence and a future-tense sentence respectively; however, the form of the Chinese verb "到(arrive)" is the same in the two cases.

Example 1:
1a. 几天前(several days ago)我(I)到(arrive)
了(n/a)上海(Shanghai).
I arrived at Shanghai several days ago. (past)
1b. 今天晚上(tonight)7点(7 pm)我(I)将(will)
到(arrive)上海(Shanghai).
I will arrive at Shanghai at 7 pm tonight. (future)

This lack of inflections in Chinese verbs imposes some challenges in many applications. For instance, in a Chinese-English machine translation (MT) system, it is almost impossible to determine the tense for the corresponding English verb if solely based on the Chinese verb morphology. However, tense information does exist in Chinese language, and it is expressed lexically instead of morphologically. There are useful contextual cues that can help determine the tense for the whole Chinese sentence or individual verbs, such as temporal adverbs or phrases, aspect auxiliary words and prepositions. For instance, in example (1a), the aspect particle "了(a particle word in Chinese, there is no literal translation to English)" and temporal phrase "几天前(several days ago)" together indicate the past tense of the sentence, and thus the correct translation of the sentence is "*I arrived at Shanghai several days ago*".

Most of the previous work on tense prediction (Li et al., 2004; Cao et al., 2004; Ye and Zhang, 2005; Lin, 2006) has been conducted using relatively small data sets (e.g., hundreds of

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1116–1124,
Chiang Mai, Thailand, November 8 – 13, 2011.

Chinese sentences) and typically news article domain. They often require hand crafted rules in the systems. In this paper, we adopt a statistical classification framework for Chinese tense prediction. This study is different from previous work in that (a) we propose to utilize the parallel Chinese-English corpora to automatically generate reference tense information, which overcomes the limitation of insufficient training data as in previous work; (b) we evaluate a variety of linguistic contextual features for this task; and (c) we propose to use a modified bootstrapping iterative learning method in order to select more reliable and informative instances, which addresses the problem that the automatically derived training data is noisy. We evaluate our system by comparing with the automatically derived references as well as human annotations. Our experimental results have shown that our tense prediction system performs reasonably well, suggesting we can leverage the parallel data to learn information for one language from the other language, and that the iterative learning approach we proposed is effective.

2 Related Work

Time Expression Recognition and Normalization (TERN) was a task evaluated in the Automatic Content Extraction (ACE) program. It requires that certain temporal expressions mentioned in the documents be detected and recognized. Such temporal expressions include both absolute and relative expressions, durations, event-anchored expressions, and sets of times. The tense prediction task we investigate in this paper is related to the TERN task in that identifying temporal expressions may be helpful to determine the verb tense. Lin (Lin, 2003; Lin, 2006) outlined a framework for temporal interpretation in Chinese from a theoretical perspective. Hundreds of rules and situations were induced to determine the appropriate tense for Chinese sentences, based on information such as viewpoint aspect, verbal semantics, temporal adverbials, and complement and relative clauses. (Li and Wong, 2002; Li et al., 2001) used a rule-based approach to combine different types of temporal indicators for temporal relation classification. These indicators include time word, time position word, temporal adverb, auxiliary word, preposition word, auxiliary verb, trend verb, and some special verbs. (He et al., 2008) explored an error-driven strategy to derive heuristic rules to recognize time expression. (Cheng, 2008) investigated using dependency structure analysis for temporal relation identification.

Machine learning methods have been adopted for the tense prediction task in recent years. (Li et al., 2004) and (Cao et al., 2004) investigated linguistic features including eleven temporal indicators and one event class, and compared several classifiers (e.g., decision trees, naive Bayes classifier). Their results showed that adopting collaborative bootstrapping approach was able to reduce the human efforts required for the task, though it also degraded the classification accuracy. (Ye and Zhang, 2005) applied conditional random fields for tense classification on Chinese news documents, and reported an overall sentence and paragraph level accuracy of around 58%.

In this paper, we attempt to automatically predict tense information using a classification framework with diverse contextual information, especially around the verbs in the sentence. To address the problem of lacking annotated data, we develop effective methods that leverage the aligned English sentences to obtain reference tense for Chinese. Similar mapping methods have been investigated recently for some natural language processing applications (Bentivogli et al., 2004; Ye and Zhang, 2005; Feldman et al., 2006; Pado and Lapata, 2009; Chen and Ji, 2009; Schwarck et al., 2010), where there are parallel corpora and annotation is only available for one language, thus allowing us to derive information for the other language based on the alignment. Since such automatically extracted labels are noisy, we further adopted a modified bootstrapping method for more effective learning. Bootstrapping or self-training has shown promising results in many different tasks by utilizing labeled and unlabeled data, such as named entity classification (Collins and Singer, 1999), parsing (McClosky et al., 2006), web page classification (Blum and Mitchell, 1998), relation and pattern extraction (Ravichandran and Hovy, 2002; Pasca et al., 2006), machine translation (Ueffing, 2006) and ontology population (Carlson et al., 2009).

3 Approaches for Chinese Tense Prediction

In this study, we focus on the prediction of absolute tense information, which indicates the rela-

tionship between the speech time and the event time under the Reichenbachian theory (Thompson, 2005). We consider four basic tenses: present, past, future, and infinitive. Other detailed tense and aspect information such as progressive and perfect is not used in this paper.

3.1 Problem Definition

The tense prediction task can be formulated as a classification task. We use a verb-based tense prediction setup, where our goal is to develop a classification system to assign each verb in the sentence a tense tag (from 4 basic tenses mentioned above). As illustrated in Example 2, tags of "1,2,3,4" denote infinitive, past, present, future tense respectively. Note that in the example the aspect particle "了(a particle word in Chinese, there is no literal translation to English)" not only is the indicator of past tense (as in Example 1), but also is used in the context of future tense. This shows the ambiguity and the challenges of the tense prediction problem.

Example 2:
上星期(last week) 没有(not) 完成(finish)/**2** 任务(the task)，我们(we) 计划(plan)/**3** 招募(recruit)/**1** 更多(more) 人(people)，因为(because) 下周五(next Friday) 我们(we) 就要(will) 汇报(submit)/**4** 结果(the results) 了。

Last week we did not finish the task, we plan to recruit more people because we will submit the results next Friday.

Compared to sentence-based framework, this verb-based setting is more flexible and has advantages especially for complex sentences. In addition, we expect that the verb-based setup is more beneficial for other applications, such as providing richer annotation for machine translation.

3.2 Leveraging Contextual Features for Tense Prediction

We use a supervised learning framework for tense classification of every verb. In this initial study, we investigate using some basic lexical features (i.e., words) and simple syntactic features (POS tags). The following lists all the features we used.

Bag of words (BOW) features: For each verb, we consider words before and after that verb within a predefined window length (5 in our experiments).

Words and POS patterns (WP): These include combinations of the word/POS tag of the current verb and those from either the previous or the following adjacent word. These features are expected to capture some expression patterns unique to some tense class. For example, the pattern "verb+了(a particle word with no literal translation to English)" means a verb followed by a word "了", which is often a good indicator of past tense.

Local bigram features: We hypothesize that in addition to information from the immediate previous and the following words of the verb, other adjacent words and patterns around the verb may also be good indicators for tense. They may form a temporal expression or part of it, such as "几天前(several days ago)". We thus extract the word bigrams before and after the verb within a 3-word window.

Global bigram features: We observe that useful cues to predict a verb's tense can appear anywhere in a sentence, therefore, we include all the bigrams in the sentence as the global feature.

Dependency features (DEP): We automatically derived dependency features for each verb as features, which are expected to capture long-distance temporal evidence through dependency relation.

Note that we limited the feature scope to within one sentence since we noticed that information from other sentences is often noisy and not helpful to determine the tense for verbs in the current sentence.

3.3 Automatic Extraction of Tense Reference based on Parallel Data

An important part of supervised learning is the collection of labeled data. For the tense prediction task, we need reference tense information for each Chinese verb. Currently there is no labeled data publicly available for this task. To avoid the time-consuming manual labeling efforts, we propose to leverage the parallel Chinese-English corpus to automatically obtain the tense annotation for Chinese verbs using the corresponding English data. The following describes our procedure.

POS tag and parse the English sentences, and generate tense information for each verb.

POS tag the Chinese sentences.

Align the English and Chinese sentences at the word level.

For each identified verb in the Chinese sentences, if it is aligned to one English verb,

or if it is aligned to multiple English verbs but with no conflicting tense types, then we use the corresponding English verb tense as the reference tag for that Chinese verb; otherwise, we do not include that Chinese verb in the training set.

This process of generating reference tense for the Chinese verbs is not perfect due to errors in English parsing and tense assignment, word alignment between Chinese and English, and Chinese POS tagging. In addition, tenses are not always well-alignable between two different languages. However, it can effectively make use of the large amount of existing parallel corpora and provides an efficient way to create a large set of labeled data automatically. In addition, most learning frameworks have the potential power to deal with noisy data and thus we expect this data set may still allow us to build an effective model for tense prediction. Furthermore, since one of our future plans is to use tense information in Chinese to help Chinese to English machine translation, we believe that using information derived from English suits this goal. (Ye and Zhang, 2005) also used a parallel Chinese-English corpus to obtain tense information, however, it was done manually. In contrast, our method is automated, which allows us to utilize a large amount of existing parallel data.

3.4 Iterative Learning Using Noisy Training Data

Since the reference data automatically derived from parallel corpora is noisy, we propose to use an iterative method to address this problem for more effective learning. Our method is similar to bootstrapping (or self-training), but different in that: (i) we do not have a gold-standard seed set to start with; (ii) our data is not fully unlabeled, instead it has the labels obtained from the mapping process using parallel data; and (iii) we use different data labeling and selection methods in the iterative learning process. Our algorithm is shown in Algorithm 1, with more detailed description below.

Algorithm 1 Iterative learning algorithm from noisy data

Let be the automatically labeled training data
Train classifier on
Select initial training set based on self-testing using
for to **do**
  Train classifier on
  Classify using
  Assign a label to each instance in
  Rank instances in each class
  Select top data samples in each class,
  Update training data,
end for

Create an initial set.

Unlike self-training, we do not have an initial human annotated seed set. Therefore we will select a reliable small set from our noisy data for iterative learning. Our expectation is that when testing the classifier on the data set used for training, the noisy data points (i.e., instances with incorrect initial labels) are likely to have wrong predictions or low confidence (note that there are other impacting factors such as the features and the classifier used). We define the confidence score for the classifier's output as the ratio between the probability of the predicted label and the probability of the second most probable label. Using this measure, we created an initial set containing 10% of the entire data that have the highest confidence scores. We also compared this confidence measure with using the standard posterior probabilities from the classifier, and found this performed better.

In each iteration, assign labels to instances in the set of unselected samples.

In each iteration, we first apply the currently trained classifier to label each instance that is not yet selected by the iterative process, and then assign a final label to an instance using information based on the current iteration classifier ( ) and the initially self-trained classifier (, which is trained using the entire data set). This is different from traditional bootstrapping where there is only one classifier during the iterations. When the two classifiers agree on the most likely tag for an instance, it is straightforward to assign this tag. When there is a disagreement, we need to resolve the conflict. For this, we use confidence score from each classifier (the same confidence measure as used above), and use the label from the classifier with a higher confidence score. To take into account of the difference of the score range, we normalize the

confidence score: _ ————.
Rank instances.

This is needed for better selection of reliable and informative instances for model training. We rank instances in each tense category (their labels are determined from the previous step) based on the confidence score from the current classifier . We found this performed better than other metrics, such as Kullback-Leibler distance and Gap ratio.
Select top-ranked instances to add to the training set.

Based on the above ranking, we selected the top (p is empirically set to 0.5 in this work) of instances in each class and added them to the training set for next iteration training. Some previous studies (e.g., (Carlson et al., 2009)) showed that constraints are useful for self-learning or semi-supervised learning in terms of quality control. Therefore, in addition to the ranking scores above, during the selection process we consider more constraints. Specifically, for each instance we compare the labels generated based on three sources: original automatically derived label, prediction from the current classifier, and prediction from the initial self-trained classifier. We filter out cases using two different constraints: (a) Constraint I: neither the prediction from the current classifier nor the self-trained one is the same as the automatically derived label. (b) Constraint II: none of those three labels agree with others, that is, they are all different.

4 Experiments and Results

4.1 Data and Experimental Setup

The training data we used is from the Chinese-English parallel MT data collection provided by LDC for the DARPA GALE program. It comes from various domains: broadcast news, broadcast conversation, newswire, weblog, and newsgroup. We split the data into sentences when there is a period, question mark, or exclamation mark. All of the English sentences were parsed using the Charniak parser (Charniak and Johnson, 2005). The English verbs were then labeled with tense tags based on the parses. We used the Chinese POS tagger from (Huang et al., 2007) and Chinese dependency parser from (Chang et al., 2009). After preprocessing the source and target language data (mainly word segmentation for Chinese and tokenization for English), we used GIZA++ (Och and Ney, 2003) to obtain a word-level alignment. After applying heuristic rules to eliminate verbs that have no aligned English verbs or are aligned to multiple verbs with conflicting tenses (see Section 3.3 for reference tense creation), we finally created a training set consisting of 279,379 verb instances and 38,087 sentences.

We chose the maximum entropy (ME) model as the classifier for tense prediction, because ME can effectively utilize many features and performs competitively with other approaches in many classification tasks. We used the ME implementation from (Zhang, 2006) with a Gaussian prior of 0.1 and 100 iterations in the model training.

4.2 Cross Validation Evaluation Using Automatically Generated Reference

First we directly use the automatically created training set and measure the cross validation performance, mainly to evaluate the modeling approach and features for tense prediction. We use classification accuracy as the performance measurement in this experiment. Table 1 shows the 5-fold cross validation results using different feature sets described in Section 3.2. The baseline is calculated when assigning the majority tag in the training set to all the test instances.

Features	Accuracy (%)
Baseline	34.70
BOW	56.89
+ WP	63.13
+ Local-Bigram	63.39
+ Global-Bigram	64.90
+ Dependency	65.02

Table 1: Classification accuracy using different feature sets for verb-based tense prediction. Results are based on 5-fold cross-validation using automatically generated tense labels.

We found that adding contextual feature sets improves performance incrementally. When only bag of words (BOW) features are used, the accuracy is 56.89%, substantially better than the baseline of 34.70%. Adding "word and POS pattern" (WP) around verbs yields significant improvement, resulting in the accuracy of 63.13%, 7.8% gain compared to using BOW features only. This suggests that such combination of word and POS may represent some syntactic characteristics

and is more indicative of tense information. As expected, adding local bigram features slightly improves the performance and global bigram features can further boost the tense classification performance, yielding an accuracy of 64.90%. This shows that for verb tense prediction, global bigrams can provide some helpful information complementary to local features in recognizing tense information. Interestingly, adding dependency features only slightly improved the performance, which could be due to the low quality of dependency extraction on speech data. On the other hand, incrementally adding features might not clearly indicate the contribution of each feature type due to overlap problem. We further conducted experiments to see the effects of excluding each feature type at one time. The results show similar patterns to the above, showing the "BOW" and "Global-Bigram" features are most important features, and "Local-Bigram" and "Dependency" features are least important features.

Note that the training data is noisy due to the automatic process. To have a better understanding of the data quality, we decided to sample a small subset of the data and create human annotation. We randomly selected 2 files in the training data and asked a native Chinese speaker to manually label the first 150 utterances of each file. The annotator was asked to decide which verbs to label and assign the appropriate tense labels. No other explicit instructions were provided. During annotation, the human annotator found that many cases are ambiguous and felt tense labeling is quite subjective, suggesting creating human annotation for tense is very challenging as investigated in (Xue et al., 2008). Using this small data set, we found that the percentage of correct tense labels by the automatic alignment process was 78.52% (329 out of 419 instances), which seems quite satisfying, although further inspection is still needed.

4.3 Evaluation on Human Annotated Data

Another setting we use to evaluate the automatic tense prediction performance is by comparing to human annotation. We performed a pilot annotation for verb tense. The same native Chinese speaker as that who created the small sample training set manually annotated the verbs with tense types using 900 utterances from 6 randomly selected files in the GALE MT 2007 development set. In total, 2484 verbs were labeled by the human annotator, and these are used as the references to measure the performance of our automatic tense prediction system.

Because the POS tagger may miss some verbs and thus there will be no system hypothesis for those verbs, two different metrics are used in this evaluation. One is the labeled recall rate, defined as the percentage of the correctly labeled verbs out of the human labeled verbs; the other metric is the tense classification accuracy measured using only those verbs that are identified by both the POS tagger and the human annotator. As dependency features do not seem to be reliable enough, we trained two models from the automatically generated data set to validate its effectiveness on test data: one is using all the features (ALL) and the other one excludes dependency features (W/O Dependency). The results are shown in Table 2. To make a more competitive baseline system, instead of using the majority category of all the verbs on the training set, we used the majority tag for each individual verb.

System	Recall(%)	Accuracy(%)
Baseline	52.79	56.52
ALL	71.50	74.80
W/O Dependency	73.09	76.49

Table 2: Verb tense prediction performance on the human annotated test data.

We can see that the POS tagging errors (i.e., missing verbs) have a noticeable impact on system performance. Note that there is a difference between the baseline performance using the human annotation vs. the automatically created data set. There are two reasons for this. First is the different class distributions in the two data sets, for example, about 28%, 34%, 32%, and 16% for infinitive, past, present, and future tense respectively in the automatically labeled set, and 15%, 35%, 42%, and 8% in human annotation. Second, the baseline results in Table 2 are word-based, rather than the majority class for the entire set.

Our results show that using parallel data to derive reference annotations is a promising approach for a statistical learning framework, achieving the best performance of 76.49% compared with the baseline of 56.52%. Although adding dependency features obtained small improvement for the cross validation result, it degraded the performance on test set (see last row in Table 2). Therefore we exclude that feature for subsequent experiments. We

also observe that the performance is much better compared to Table 1, even though the class distributions are different in training and testing (which may affect statistical learning). One explanation might be that noisy labels from automatically generated training data may harm the cross validation performance. In addition, statistical learning can deal with the noisy data for training a robust model on unseen test data. We will show how the effects of noisy labels can be further reduced by iterative learning in next section.

4.4 Iterative Learning Results

Table 3 shows the best results for three iterative learning settings evaluated using the human annotated data set. These are from different numbers of iterations with different combinations of label assignment and instance ranking methods. We can see that using iterative learning generally improved tense classification in different settings – achieving the best accuracy of 77.49% compared to the original 76.49%, and the best recall rate of 74.02% compared to the original 73.09%. It suggests that our proposed learning method can effectively overcome some negative impact of the noisy training data, by iteratively selecting more likely trustworthy instances for model training. The best performance was obtained by applying constraint II (74.02%/77.49%), while adding constraint I degraded the performance a lot, which could be because using constrain I mislead the iterative process as more rules would miss out some useful training instances in the early stage. It suggests that adding appropriate constrains in the iterative bootstrapping process can provide better quality control for improved performance.

Selection Constraint	Recall (%)	Accuracy (%)
W/O Constraints	73.85	77.31
Constraint I	72.31	75.72
Constraint II	74.02	77.49

Table 3: Verb tense prediction performance on the human annotated test data using iterative learning. The results using the original training data are 73.09 and 76.49 for recall and accuracy.

For the iterative learning algorithm, we also examined its learning curve. Figure 1 shows the curves for the above best system using Constraint II. The horizontal dotted line is the baseline result when we just use the original data without iterative learning. We notice that performance in the early iterations generally increases (with some fluctuations) as more data samples are added for model training. After certain number of iterations, the performance starts dropping, since some added instances are noisy and do not help learning. Based on the trend shown in the curve, we expect that when the classifier itself improves, such as when incorporating more discriminative features, the learning curve could potentially increase further. The curves for other settings show similar patterns, with the best results achieved at different number of iterations.

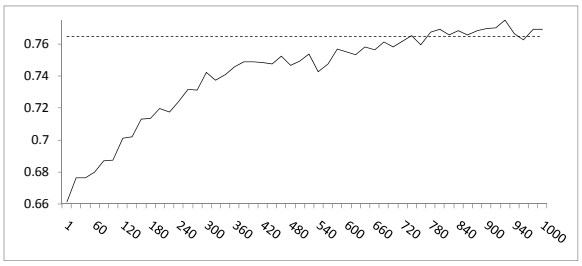

Figure 1: Learning curve of the iterative learning approach.

Another question we try to answer is how much we can attribute the performance improvement to iterative learning, or whether simply resampling can achieve similar performance. We examined re-sampling methods based on confidence measure (as described in iterative learning) to select similar number of samples as used in the iterative learning system above (using its optimal number of iterations), and then evaluated the classifier trained using this subset of the data. The results shows that simple re-sampling only obtained the accuracy of 74.58%, which is even worse than using the entire data set. Therefore we can conclude that the improvement we observe is mainly from the iterative learning method.

4.5 Error Analysis

Example 3 illustrates some examples of errors from our analysis. We identified several reasons causing the errors of our tense prediction system:

> Imperfect Chinese word segmentation and POS tagging results can directly affect the correct identification of Chinese verbs. In example (3a), the named entity "戴尔(Dell)" was incorrectly segmented as two Chinese words, and "戴" was further tagged as a verb due to its verb meaning "wear". In

Example 3:
3a) 戴/3 尔(Dell) 公司(Inc.) 新闻(media) 发言人 (spokesman) JessBlackburn 表示(indicate)/2 ， 该(the) 公司 (company) 正在(is) 评估(evaluating)/3 Google 的 软件 (software) 。

The spokesman of Dell Inc. indicated that, the company is evaluating the software from Google.

3b) 小时(hour) 工部(department) 分(divide)/3 职位(position) 技能(skill) 要求(require)/3 不(not) 高(high)/1， 如(such as) 食品(food) 促销 员(salesman)/3 、 理 货 员(shelf stockers) 等(etc) ， 失业(unemployment) 人员(people) 经过(by) 简单 (simple)/3 的 培训(training) 即可(can) 上(take)/3 岗(job) 。

Some positions for hourly workers require less job skills, such as the food salesman, shelf stockers, etc., the unemployed can take the job after simple training.

3c) 负责人(manager) 称(claim)/2，他们(they) 会(will) 在 (by) 2005 年(year) 底(end) 完成(finish)/4 这项(the) 工程 (project) 。

The manager claimed that they would finish the project by the end of 2005.

example (3b), the phrase "小 时 工(hourly worker)部 分(some)职 位(position)" was incorrectly segmented and the resulting word "分(divide)" was incorrectly tagged as a verb. In addition, some adjectives were also wrongly recognized as verbs, such as "高(high)", "简单(simple)".

There is limitation using our current feature set and it needs deep understanding to derive the correct tense. In example (3c), the system assigns future tense to the Chinese verb "完成(finish)" mainly because of the evidence "会(will)". To correct this, the system needs to determine that "the end of 2005" already passed based on knowledge or other long distance context (e.g., "But they didn't." in the following sentence).

The Chinese verb may not always be translated into English verbs, and vice versa. Training noises plus the imperfect alignment tools and other propagated errors from other preprocessing modules contributed another portion of errors.

Human annotation is not perfect. In some cases, both the system's output and human annotation are acceptable but they are not consistent. On the one hand, it shows the challenges of this task; on the other hand it suggests that investigation of inter-agreement among human subjects is needed in the future.

5 Conclusion and Future Work

In this paper, we have developed a classification approach to predict tense for Chinese verbs. We proposed an automatic mechanism to generate reference tense for the Chinese verbs that utilizes the Chinese-English parallel corpora, thus can efficiently create a large training set without relying on the time-consuming human annotation efforts. Our experimental results have shown that various contextual features around verbs can be effectively used to determine tense information, and this method of leveraging parallel corpora is feasible for Chinese tense prediction. In addition, the bootstrapping approach we explored in this paper further improves performance, proving to be an effective way to select training samples iteratively from noisy data.

In our future work, we will refine the process to automatically create the training set to better deal with problems from POS tagging, parsing, and alignment. A better data set will likely improve the tense prediction model. In addition, inter-agreement of human annotation on tense information is worth studying, especially for conversational style data. Finally, we will investigate using richer syntactic information and other semi-supervised methods (e.g. co-training in (Bergsma et al., 2011)) for tense prediction, and more importantly, develop methods using tense information derived in our system for machine translation and other applications.

Acknowledgments

We thank Eugene Charniak for providing his tools to obtain the tense information for the English verbs, Mary Harper for providing the parses for English and POS tagger for Mandarin, and Mari Ostendorf for useful discussions. This work was supported by DARPA under Contract No. HR0011-06-C-0023. Distribution is unlimited.

References

Luisa Bentivogli, Pamela Forner, and Emanuele Pianta. 2004. Evaluating cross-language annotation transfer in the multisemcor corpus. In *Proceedings of COLING*.

S. Bergsma, D. Yarowsky, and K. Church. 2011. Using large monolingual and bilingual corpora to improve coordination disambiguation. In *Proceedings of ACL/HLT*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*, pages 92–100.

Bianka Buschbeck, Renate Henschel, Iris Höser, Gerda Klimonow, Andreas Küstner, and Ingrid Starke. 1991. Limits of a sentence based procedural approach for aspect choice in German-Russian MT. In *Proceedings of EACL*, pages 269–274.

Guihong Cao, Wenjie Li, Kam-Fai Wong, and Chunfa Yuan. 2004. Combining linguistic features with weighted bayesian classifier for temporal reference processing. In *Proceedings of COLING*, pages 702–708.

Andrew Carlson, Justin Betteridge, Estevam R. Hruschka, Jr., and Tom M. Mitchell. 2009. Coupling semi-supervised learning of categories and relations. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 1–9.

P. Chang, H. Tseng, D. Jurafsky, and C.D. Manning. 2009. Discriminative reordering with chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL*, pages 173–180.

Zheng Chen and Heng Ji. 2009. Can one language bootstrap the other: A case study on event extraction. In *Proceedings of HLT-NAACL 2009 Workshop on Semi-supervised Learning for Natural Language Processing*.

Y. Cheng. 2008. Constructing a temporal relation identification system of Chinese based on dependency structure analysis. Thesis in Nara Institute of Science and Technology.

M. Collins and Y. Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of EMNLP*.

Anna Feldman, Jirka Hana, and Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of LREC*.

R. He, B. Qin, T. Liu, Y. Pan, and S. Li. 2008. A novel heuristic Error-Driven learning for recognizing Chinese time expression. *Journal of Chinese Language and Computing*, 18(4):139 - 159.

Zhongqiang Huang, Mary Harper, and Wen Wang. 2007. Mandarin part-of-speech tagging and discriminative reranking. In *Proceedings of EMNLP*.

Jussi Karlgren. 1996. Tense prediction for speech recognition purposes. Technical Report in New York University.

Wenjie Li and Kam-Fai Wong. 2002. A word-based approach for modeling and discovering temporal relations embedded in Chinese sentences. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(3):173–206.

Wenjie Li, Kam-Fai Wong, and Chunfa Yuan. 2001. A model for processing temporal references in Chinese. In *Proceedings of the workshop on Temporal and spatial information processing*, pages 1–8.

Wenjie Li, Kam-Fai Wong, Guihong Cao, and Chunfa Yuan. 2004. Applying machine learning to Chinese temporal relation resolution. In *Proceedings of ACL*, pages 582–588.

JoWang Lin. 2003. Temporal reference in mandarin Chinese. *Journal of East Asian Linguistics*, 12(3):259–311.

JoWang Lin. 2006. Time in a language without tense: The case of Chinese. *Journal of Semantics*, 23(1):1–53.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of HLT-NAACL*, pages 152–159.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sebastian Pado and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*, 26:307–340.

Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. 2006. Names and similarities on the web: fact extraction in the fast lane. In *Proceedings of ACL*, pages 809–816.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of ACL*, pages 41–47.

Florian Schwarck, Alexander Fraser, and Hinrich Schuetze. 2010. Bitext-based resolution of german subject-object ambiguities. In *Proceedings of HLT/NAACL*.

Ellen Thompson. 2005. *Time in Natural Language: Syntactic Interfaces with Semantics and Discourse*. Walter de Gruyter.

Nicola Ueffing. 2006. Self-training for machine translation. In *Proceedings of NIPS workshop on Machine Learning for Multilingual Information Access*.

R. Z. Xiao and A. M. McEnery. 2002. A corpus-based approach to tense and aspect in English-Chinese translation. In *The 1st International Symposium on Contrastive and Translation Studies between Chinese and English*.

N. Xue, H. Zhong, K. Y Chen, and M. Marrakech. 2008. Annotating "tense" in a tense-less language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Yang Ye and Zhu Zhang. 2005. Tense tagging for verbs in cross-lingual context: A case study. In *Proceedings of IJCNLP*, pages 885–895.

Le Zhang. 2006. Maximum entropy modeling toolkit for Python and C++. http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.