# Aspect Marker Generation in English-to-Chinese Machine Translation

**Yang Ye**
Linguistics Department
University of Michigan
Ann Arbor, MI, USA
yye@umich.edu

**Karl-Michael Schneider**
Textkernel
Nieuwendammerkade 28A17
1022AB Amsterdam, Netherlands
karlmicha@gmail.com

**Steven Abney**
Linguistics Department
University of Michigan
Ann Arbor, MI, USA
abney@umich.edu

## Abstract

This paper reports on a pilot study of aspect marker generation in an English-to-Chinese translation scenario. Our classifier combines a number of linguistic features in a Maximum Entropy framework and achieves an overall accuracy of 78%. We also investigate the impact of different clusters of linguistic features; we find that syntactic features have the highest utility and lexical aspectual properties associated with verbs do not have significant contribution to the prediction of aspect markers. Furthermore, we have demonstrated converging evidence that there is only marginal sequential dependency between the aspect markers of different verbs in the same sentence.

## 1 Introduction

It is well documented in linguistics that natural languages alternate between two foci of temporal reference expressions: The first focus is based on precedence of events, that is, an event is earlier or later than another event. The second focus is based on the relative positioning of an event with respect to the following three time parameters proposed by Reichenbach (1947): speech time (S), event time (E) and reference time (R).

Languages like English use morphologically marked tense to express when an event happens, ignoring its temporal length (the English progressive is aspectually marked, though). In contrast, Chinese possesses a set of active aspect markers to express whether an event has finished or is still ongoing, ignoring when it happens.

Correct translation of tense and aspect is crucial for translation quality, not only because tense and aspect are relevant for all sentences, but also because temporal information is essential for a correct delivery of the meaning of a sentence. When the translation output is pipelined with other functions for higher level tasks, errors in tense and aspect translation might result in problems such as inaccurate returns for queries in Temporal Question Answering or inappropriate merging of distinct events in Multi-Document News Summarization.

The different temporal strategies in English and Chinese pose a challenge for tense and aspect marker generation in automatic translation. Since Chinese aspect markers are represented as separate lexemes rather than morphemes, neither word-based alignment nor phrase aligning algorithms can capture the mapping between the tense markers of English verbs and the aspect markers of the corresponding Chinese verbs.

The following examples are randomly selected sentences that were translated using the Google English-to-Chinese MT system:

I first went to the post office, then I went back home.

| | | | | | | |
|---|---|---|---|---|---|---|
| 先 | 去 | 郵政局, | 那 | 我 | 回 | 家 |
| first | go | post office | then | I | go back | home |

I sent her a letter.

| | | | | | |
|---|---|---|---|---|---|
| 我 | 寄 | 給 | 她 | 的 | 信 |
| I | mail | to | her | DE | letter |

I sent a letter to her.

| | | | | |
|---|---|---|---|---|
| 我 | 寫 | 信 | 給 | 她 |
| I | write | letter | to | her |

I sent a letter to him.

| | | | | |
|---|---|---|---|---|
| 我 | 寫 | 信 | 給 | 他 |
| I | write | letter | to | him |

I went to Beijing last summer.

| | | | | |
|---|---|---|---|---|
| 去年 | 夏天, | 我 | 去 | 北京 |
| last year | summer | I | go | Beijing |

I phoned my mom yesterday evening.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 我 | 昨天 | 晚上 | 打 | 電話 | 給 | 我 | 媽媽 |
| I | yesterday | evening | make | phonecall | to | my | mother |

She refused him.

| | | |
|---|---|---|
| 她 | 拒絶 | 他 |
| she | refuse | him |

I believed him.

| | | |
|---|---|---|
| 我 | 相信 | 他 |
| I | believe | him |

They lost the game.

| | | |
|---|---|---|
| 他們 | 失去 | 游戲 |
| they | lose | game |

He broke his leg last month.

| | | | | |
|---|---|---|---|---|
| 他 | 上個 | 月 | 爆發 | 腿部 |
| he | last | month | break | leg |

In all sentences, the LE aspect marker (marking complete aspect) is missed in the MT output. These sentences are typical sentences in daily use. However, because they are possibly very sparse in the training data of the MT system, the alignment between the English tensed verbs and the n-grams containing the corresponding Chinese verb and the LE aspect marker cannot be captured by the alignment model of the MT system.

The following examples from the Google English-to-Chinese MT system illustrate how the translation of aspect markers in even short and simple sentences depends on the frequency of alignments between the specific tensed English verb and the n-gram in Chinese in the training data:

> I sent her a letter.
> 我寄給她的信。

> I sent him a letter.
> 我寄了一封信。

While the two sentences differ only in the gender of the object, LE is translated correctly in one sentence but not in the other.

Treating aspect marker generation as a separate task, rather than as part of a complete MT system, makes it more feasible to correctly predict Chinese aspect markers with a moderate data set. Furthermore, the lack of precise correspondence between tense, aspect markers and linguistic cues makes a learning-based approach more viable.

Previous work has left the problem of Chinese aspect marker generation largely untouched. This paper investigates aspect marker generation in English-to-Chinese translation. In particular, we address the following task: given an English sentence, predict appropriate insertions of Chinese aspect markers following the verbs in the Chinese translation. Our approach is based on a combination of different types of features extracted from both the English source sentences and the Chinese target sentences. The central issues are: (1) Which features are most useful for aspect marker generation in English-to-Chinese translation? And (2) How strong is the sequential dependency among the aspect markers of verbs in multi-verb sentences?

## 2 Related Work

The extensive body of literature on temporal information processing in the past few decades has focused on issues such as event ordering, time-stamping, and temporal connection words generation. Little work, however, has addressed cross-lingual temporal reference distinction mapping, and the challenge of this mapping for some language pairs has not received much attention. In the rest of the section, we review work that is more closely related to the theme of the current study.

Campbell et al. (2002) proposed a language-neutral framework, the Language Neutral Syntax (LNS), as a syntactic representation for tense. Based on the observation that grammatical tenses in different languages do not necessarily mean the same thing, they interpret semantic tense to be largely a representation of event sequence. The tense node in the LNS tree contains either a global tense feature or an anchorable tense feature; thus compound tenses are represented by primary and secondary tense features. The tense in an embedded clause is anchored to the tense in the matrix clause.

Pustejovsky et al. (2004) reported a temporal annotation scheme, the TimeML metadata, for the markup of events and their anchoring in documents. The challenge of human labeling of links among eventualities was also discussed extensively. Automatic “time-stamping” was attempted on a small sample of text in an earlier work of Mani et al. (2005). The result was not particularly promising, showing the need for a larger amount of training data as well as more predictive features, particularly at the discourse level.

Li et al. (2004) reported a computational model based on machine learning and heterogeneous collaborative bootstrapping, which classifies temporal relations in Chinese multiple-clause sentences. The core model is a set of rules that map the combined effects of a set of linguistic features onto one class of temporal relations for an event pair. This work showed promising results for combining machine learning algorithms and linguistic features for the purpose of temporal relation resolution.

Olsen et al. (2001) addressed tense reconstruction on a binary taxonomy (present and past) for Chinese text in Chinese-to-English MT. Besides the more overt features, their work made use of the telicity information encoded in the lexicon through the use of Lexical Conceptual Structures (LCS). Based on the dichotomy of grammatical aspect and lexical aspect, they proposed that past tense corresponds to the telic LCS, which is either inherently telic or derived telic. While grammatical aspect markings supersede the LCS, in the absence of grammatical aspect marking, verbs that have telic LCS are translated into past tense, whereas verbs without telic LCS are translated into present tense. Their work, while pushing tense reconstruction towards the semantic level, oversimplifies the temporal reference situation in Chinese by adopting a one-to-one mapping between grammatical aspect marking and tense.

Ye et al. (2005; 2006) showed that for the task of Chinese-to-English translation, a tense classifier can be trained on a moderate size of data with promising overall accuracy by combining a number of linguistic features with a standard classification algorithm. They also reported on the high utility of lexical aspectual features in the classification task. This confirms Olsen et al.’s (2001) report on the significance of the verb telicity feature in tense reconstruction in Chinese-to-English translation as well as Ye et al.’s (2006) analysis that lexical aspectual features significantly affect the inter-annotator agreement rate for tense annotation.

## 3 Problem Formulation and Feature Sets

We formulate aspect marker generation as a classification problem, in which we learn the mapping from a set of features onto different aspect marker classes from the training data set. The taxonomy of Chinese aspect markers includes the following four aspect marker tags: LE, ZHE and GUO, which encode complete, progressive, and experiential aspect, respectively, and in the absence of any of the above, a NULL marker.

All of LE, ZHE and GUO are obligatory in some contexts and optional in others, and human annotators do not always agree on the distinction between them. Obligatory aspect markers do not carry extra semantic information; they can be predicted by the features in the context, given that the features can be reliably extracted, which is not always the case. Optional aspect markers, on the other hand, can change the meaning of a sentence slightly when added, but the difference is so subtle that even native speakers do not always agree on the exact meaning. Therefore we think it does not make sense to have a machine make the distinction. The grey area between optional and obligatory aspect markers motivated us to combine them in our study. For the reasons above, aspect marker generation discussed in this paper is mostly about the well-formedness of MT output.

The aspect marker of a Chinese verb is jointly decided by various properties associated with the verb. The presence of certain aspect markers following a verb is a function of a range of features that human annotators depend on to judge the goodness of adding a certain aspect marker. Below we discuss the features explored in our experiments of aspect marker generation in several groups.

Because the current study attempts to improve upon the aspect marker resolution of an MT system output, we are primarily interested in features from the target language, which are more important in predicting aspect markers than features from the source language. In a non-translation scenario, for example, in assisting foreign language learners to decide upon aspect markers when producing Chinese sentences, features will be exclusively from the Chinese text.

### 3.1 Syntactic Features

Syntactic features are either features of the phrase structure the verb is embedded in or features of the neighboring phrase structures. Syntactic features can influence the verb's tendency to take an aspect marker. The syntactic features employed in our experiments include:

- what type of object the verb takes;
- in what type of embedding structure the current verb occurs;
- whether the verb is in a sequential verbal construction;
- whether the verb is embedded in a "shi…de" structure, which is the structure for emphasizing a fact;
- what part of speech the previous word is;
- what part of speech the next word is.

### 3.2 Positional Features

Conceptually, verbs in the vicinity of each other will interact with one another and possibly change each other's inclination to take an aspect marker. We experiment with a group of features that are related to a verb's position in the sentence as well as its position with regard to other verbs. The features we explore include:

- whether clauses occur before the current verb;
- whether the verb is followed immediately by a comma;
- whether the verb is at the end of the sentence;
- the distance between the current verb and the previous verb.

### 3.3 Signal Lexeme Features

LE, ZHE and GUO are each compatible with certain signal auxiliary words, demonstrating certain lexical co-occurrence patterns. The presence of these words before the verb increases the probability of the verb taking certain aspect markers. These features include:

- whether "yi3(jing1)" (already) exists before the verb;
- whether "ceng2(jing1)" (once) occurs before the verb;
- whether "gang1(gang1)" (just now) occurs before the verb;
- whether negation occurs before the verb;
- whether "jiang1(yao4)" (be going to) occurs before the verb.

### 3.4 Lexical Aspectual Features

Tense and aspect are intrinsically associated with lexical aspectual features specifying the verb's proneness to being bound temporarily by a limited time span or a time point (Vendler, 1967). Previous research has reported the importance of lexical aspectual properties for tense generation in Chinese-to-English translation (Olsen et al., 2001; Ye et al., 2006). One would assume that lexical aspectual features would have a significant impact on aspect marker generation in English-to-Chinese translation as well, and we investigate the following two features in particular:

- the punctuality feature of the verb;
- the telicity feature of the verb.

The major challenge of assigning lexical aspectual features to verbs is that a verb can denote situations of multiple aspectual types in different contexts. Although a verb is typically associated with certain inherent aspectual features in isolation, these features are volatile once the verb enters a sentence. Therefore, we annotated these two features based on whole sentences rather than verbs in isolation.[1]

### 3.5 Phonological Feature

Aspect markers in Chinese are incompatible with idioms that typically have four Chinese characters. Therefore, we experiment with the feature of whether the verb is a four-character verb.

### 3.6 English Tense Feature

All of the above features are from the Chinese verbs under investigation. In addition to those target language features, we also include the tense of the corresponding English verb in our feature set. Presumably, the English morphological tense is a very informative feature in our task because it has a strong relation to Chinese aspect (both express temporal reference).

## 4 Empirical Experiments and Results

### 4.1 Data

We used 250 parallel news articles of English and Chinese from the LDC Multiple Translation Corpora (LDC2002T01, LDC2003T17 and LDC2004T07). The source articles are in Chinese and have been translated into English by several translation teams. We chose the best

[1] The aspect markers were removed from the texts when the annotation was performed in order to avoid the bias the annotators might have while annotating the features when they see certain aspect markers.

human English translation as selected by the LDC. This way we tried to approximate a corpus with English source sentences and Chinese translations (such a corpus was not available directly).[2]

There was a total number of 2723 verbs in the whole data set. We manually annotated the aspect marker tags for the verbs in the Chinese texts. For the annotation, the annotators were instructed to tag both optional and obligatory aspect markers, but for the experiments, the distinction between obligatory and optional was removed.

All features are automatically extracted except the English tense feature and the lexical aspectual features. Although English tense can be identified with high accuracy using simple rules, aligning the English tense with the Chinese verbs using automatic tools such as GIZA++ would introduce additional noise. Therefore we obtained the English tense feature manually. Chinese verbs that were nominalized in the English translation were assigned a "null" tag to the English tense feature.

To validate the human annotator's reliability of assigning aspect markers to Chinese verbs, we carried out a pilot annotation experiment in which three Chinese native speakers were recruited to annotate the Chinese aspect markers for a small-scale data set consisting of 12 news articles and 250 verbs. The Kappa score, which is the de facto measurement of inter-annotator agreement rate, was used to calibrate the reliability of the annotation experiment (Cohen, 1960). It is defined by the following formula, where $P(A)$ is the observed agreement among the annotators and $P(E)$ is the expected agreement:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

The expected agreement is computed assuming that each annotator has an individual distribution of labels (Cohen, 1960).[3] A Kappa score of 0 indicates agreement merely by pure chance, whereas 1 indicates perfect agreement.

When the annotators performed the annotation under the provided guidelines, a Kappa score of 0.929 was achieved on the reduced taxonomy that combines obligatory and optional aspect markers, indicating a very good agreement rate for Chinese aspect marker annotation. Given this high inter-annotator agreement rate, one of the three annotators was asked to annotate the aspect markers for the whole data set as an oracle.

In order to make maximum usage of the available data, all experiments are performed using five-fold cross-validation. We suspect that verbs with a shorter distance to the previous verb (equal or less than 10 words) in the same sentence might be more difficult to tag compared to verbs with a longer distance to the previous verb (more than 10 words) and verbs in single-verb sentences. Thus, the data was split into five parts so that there was an approximately equal distribution of single-verb sentences, multiple-verb sentences with short distance between two verbs and multiple-verb sentences with longer distance between two verbs.

For our experiments, verbs in news headlines were removed from the data set based on the observation that verbs in news headlines behave differently from other verbs. The distribution of the four classes of aspect markers is shown in Table 1.

| Aspect Marker | LE | ZHE | GUO | NULL |
|---|---|---|---|---|
| Frequency | 1216 | 186 | 33 | 1174 |

Table 1: Aspect marker distribution in our data.

### 4.2 Classifier Learning and Evaluation

We use Conditional Random Fields (CRFs) (Lafferty et al., 2001) to learn the mapping from sets of features to aspect markers. CRFs are undirected graphical models in which a set of output variables is conditioned on a set of input variables. In our case, the output variables are the aspect markers attached to the verbs in a sentence while the input variables are the features derived from the sentence. Using a CRF we can compute the conditional probability of a given set of aspect markers for one sentence given its features. CRFs employ a globally normalized log-linear model and can thus combine features without relying on independence assumptions (Lafferty et al., 2001).

We choose sentences, not verbs, as the unit in our experiments because while we assume there is no interaction between the aspect markers of verbs across sentence boundaries, we are interested in possible within-sentence dependencies.[4]

In our experiments we use two different types of CRF structures. In a linear-chain CRF, the output variables are connected by edges in a linear chain. These models make a first-order Markov assumption; they can be roughly understood to correspond to conditionally trained hidden Markov models. Linear-chain CRFs are a globally normalized extension to Maximum Entropy Markov Models (McCallum et al., 2000). In our scenario this means that the aspect marker for a verb depends solely on the features of that verb and on the aspect marker of the previous verb.

In addition to the linear-chain structure, we also use a simpler structure in which output nodes are not connected at all, roughly corresponding to zero-order maximum entropy models. These models assume no direct dependence between the aspect markers of different verbs in a sentence with multiple verbs. However, due to the globally normalized nature of CRFs, co-occurrence dependencies between aspect markers for verbs that appear in the same sentence might still be captured. Both first-order and zero-order linear-chain CRFs predict the whole set of aspect markers for one sentence based on its features.

[2] Alternatively, we could have chosen to use automatic translations of English source sentences, but those translations would have been naturally imperfect, and we think that feature extraction from such imperfect translations would be very unreliable and therefore may not be useful, especially when the purpose of this work is not only to improve aspect marker translation quality, but also to investigate the utility of different features.

[3] An alternative definition of the Kappa score assumes that there is a single hypothetical distribution of labels for all annotators (Eugenio, 2004).

[4] Previous research (Ye et al., 2005) reported equal accuracy of sentence sequence and paragraph sequence for tense classification in the opposite (Chinese-to-English) scenario.

|  | LE | ZHE | GUO | NULL |
|---|---|---|---|---|
| P | 0.7872 | 0.5 | 0.65 | 0.8038 |
| R | 0.8244 | 0.5018 | 0.5 | 0.787 |
| F | 0.8051 | 0.4986 | 0.5343 | 0.7862 |

Table 2: Results using all features.

All experiments are performed using MALLET (McCallum, 2002), a Java implementation of CRFs. We use the SimpleTagger class from MALLET which implements linear-chain CRFs and has a command line option to set the Markov order.

For the evaluation we use standard classification accuracy by comparing the predicted aspect markers to the annotated (gold-standard) aspect markers. To measure the classifier performance on each individual aspect marker, we also report precision, recall, and F score, for each aspect marker. Precision for some aspect marker is the number of times that this marker is correctly predicted, divided by the total number of times that this marker is predicted. Recall is the number of times that a marker is correctly predicted divided by the number of occurrences of the marker in the gold standard. F score is the harmonic mean of precision and recall. All results are averaged over the trials in the five-fold cross-validation, and we report the p-values from a paired t-test along with evaluation measures.

### 4.3 Aspect Marker Prediction Using All Features

We trained a first-order CRF model using all the features discussed in Section 3. This model achieved an overall accuracy of 77.25%. Table 2 shows precision, recall and F score for each aspect marker. A simple baseline that always assigns the most frequent aspect marker (LE) achieves an overall accuracy of only 46.6%.

### 4.4 Feature Utility Ranking

In order to evaluate the usefulness of different types of features, we ran additional experiments that used subsets of the features, leaving out one group of features each time. We experimented with five groups of features of concern: syntactic features, positional features, signal lexeme features, lexical aspectual features, and English tense feature. Table 3 summarizes the evaluation results of classifiers trained on different feature subsets.

It can be observed that the least accurately tagged aspect markers are ZHE and GUO, which occur sparsely in most data domains. We believe that these two aspect markers are not intrinsically more challenging to predict, and that using larger data sets with more occurrences of these markers will yield more accurate classifiers.

Nevertheless, the evaluation results for LE and NULL, as well as the overall accuracy, reveal the impact of different feature groups. Removing syntactic features results in the largest drop in classification accuracy (F score for LE by 3.4 percentage points with a p-value of 0.004, F score for NULL by 8.3 percentage points with a p-value of 0.001, overall accuracy of the classifier by 5 percentage points with a p-value of 0.002). The utility of the English tense feature is evident too: removing it causes a significant drop in the classifier's performance (2 percentage points for the F score of LE with a p-value of 0.013, 1 percentage point for the F score of NULL with a p-value of 0.036, 3 percentage points for the overall accuracy with a p-value of 0.01).

|  |  | Short | Long | Single |
|---|---|---|---|---|
| LE_F | (All Features) | 0.7593 | 0.8046 | 0.8051 |
| LE_F | (No Positional) | 0.7473 | 0.7927 | 0.7966 |
| NULL_F | (All Features) | 0.7268 | 0.7597 | 0.7907 |
| NULL_F | (No Positional) | 0.7008 | 0.7294 | 0.7766 |

Table 4: Impact of positional features on aspect marker prediction depending on inter-verb distance.

The features of signal words are significant for NULL (a drop in F score of 1.6 percentage points with a p-value of 0.036) and for overall accuracy (a drop of 1.5 percentage points with a p-value of 0.031) but not for LE (a drop of less than 1 percentage point with a p-value of 0.25).

The positional features are significant too, removing them resulted in a drop of 1 percentage point for overall accuracy (p = 0.014). We suspected that positional features would have a larger impact on predicting a NULL aspect marker insertion than predicting on the LE tag. The experimental results confirmed this by demonstrating a bigger drop in the classification performance for NULL (1.7 percentage points, p = 0.008) than for LE (1 percentage point, p = 0.05) when positional features were removed from the feature set.

Surprisingly, lexical aspectual features (i.e., telicity and punctuality features) do not have a significant impact on the classification performance (except for ZHE). There is only a very slight (less than 1 percentage point) drop for the F score of LE (p = 0.293) and for overall accuracy (p = 0.351), while NULL and GUO remain almost unaffected. However, lexical aspectual features seem to be important for ZHE, since its F score drops considerably (from 0.4986 to 0.4115, though with a p-value of 0.895). This result is in contrast to the high utility of lexical aspectual features for tense classification in Chinese-to-English translation (Ye et al., 2006).

To determine whether there is interaction between the positional features and different inter-verb distances, we also examined the effect of removing positional features on verbs with different inter-verb distances, as shown in Table 4. Statistical significance testing showed that removing positional features did not result in significant performance change between verbs with different inter-verb distances, except that the change is marginally significant between verbs with long distance and those with short distance for the LE tag (p = 0.07). This indicates that compared to verbs that are closer to the previous verb, it is slightly more difficult to predict an insertion of LE compared to verbs further away from the previous verb.

## 5 Sequential Dependencies of Aspect Markers

One could expect a certain degree of sequential dependency for aspect marker prediction among neighboring verbs. Ye

| | all features | no syntactic | no positional | no signal lexeme | no aspectual | no tense |
|---|---|---|---|---|---|---|
| LE_P | 0.7872 | 0.7296 | 0.7726 | 0.7727 | 0.7642 | 0.7651 |
| LE_R | 0.8244 | 0.8178 | 0.8207 | 0.8249 | 0.8338 | 0.8098 |
| LE_F | 0.8051 | 0.7710 | 0.7955 | 0.7976 | 0.7973 | 0.7864 |
| ZHE_P | 0.5 | 0.5623 | 0.5159 | 0.4899 | 0.4903 | 0.3705 |
| ZHE_R | 0.5018 | 0.4382 | 0.5106 | 0.4789 | 0.3646 | 0.3031 |
| ZHE_F | 0.4986 | 0.4899 | 0.5092 | 0.481 | 0.4115 | 0.3324 |
| GUO_P | 0.65 | 0.7 | 0.6633 | 0.4 | 0.6343 | 0.6917 |
| GUO_R | 0.5 | 0.6676 | 0.4812 | 0.4 | 0.4994 | 0.6062 |
| GUO_F | 0.5343 | 0.6593 | 0.5205 | 0.4 | 0.53 | 0.6038 |
| NULL_P | 0.8038 | 0.738 | 0.7919 | 0.7905 | 0.8052 | 0.7826 |
| NULL_R | 0.787 | 0.6728 | 0.7486 | 0.7511 | 0.7663 | 0.7626 |
| NULL_F | 0.7862 | 0.7034 | 0.7693 | 0.7699 | 0.7851 | 0.7719 |
| Accuracy | 0.7725 | 0.7235 | 0.7613 | 0.7571 | 0.7653 | 0.749 |

Table 3: Results for different feature sets.

| | LE | ZHE | GUO | NULL |
|---|---|---|---|---|
| P_order 1 | 0.7872 | 0.5000 | 0.65 | 0.8038 |
| R_order 1 | 0.8244 | 0.5018 | 0.4776 | 0.7699 |
| F_order 1 | 0.8051 | 0.4986 | 0.5343 | 0.7862 |
| accuracy | 0.7725 | | | |
| P_order 0 | 0.8004 | 0.6058 | 0.6857 | 0.8091 |
| R_order 0 | 0.8465 | 0.5251 | 0.6274 | 0.7794 |
| F_order 0 | 0.8226 | 0.5546 | 0.6429 | 0.7939 |
| accuracy | 0.7899 | | | |

Table 5: Comparison of first and zero order CRFs.

| | LE | ZHE | GUO | NULL |
|---|---|---|---|---|
| P | 0.8024 | 0.4638 | 0.7143 | 0.8058 |
| R | 0.8432 | 0.5056 | 0.5844 | 0.7862 |
| F | 0.8219 | 0.5382 | 0.615 | 0.7957 |

Table 6: Results for individual verb classification.

et al. (2006) reported that the sequential dependency between the tenses of neighboring verbs is not significant based on the observation that a CRF tense classifier does not outperform a non-sequential classifier. In order to test the sequential dependency hypothesis for aspect marker prediction, we carry out experiments with three different model structures.

### 5.1 Sequential Dependency

In the first set of experiments, we compare first-order and zero-order linear-chain CRFs (see Section 4.2). In the presence of sequential dependency, we would expect a first-order model to perform better than a zero-order model, where the probability of the current verb taking a certain aspect marker does not depend on the marker of the previous verb.

Our experiments showed that a classifier based on a first-order model does not outperform the zero-order model classifier. We found that a zero-order model classifier achieved slightly better performance than the first-order model classifier, as shown in Table 5. The differences are, however, not statistically significant (the p-value for the F score differences are 0.074 for LE and 0.38 for NULL). We attribute this slight increase in accuracy to specific properties of our data (e.g., overfitting).

### 5.2 Global Optimization

CRFs are trained to model the joint probability of all output variables (i.e., aspect markers) given the values of all input variables (i.e., sentence features). A classifier for aspect marker prediction based on a first-order or zero-order linear-chain CRF thus predicts the sequence of aspect markers (first-order) or the set of aspect markers (zero-order) for all verbs in a sentence. In order to determine whether the global normalization in a CRF benefits aspect marker prediction, we trained a maximum-entropy classifier with all verbs in our data as individual instances.[5]

Our results showed that not only did the performance of the classifier not drop, but it improved slightly (Table 6). The overall accuracy is 78.91%, an increase of 1.7% over the first-order linear-chain model. The performance difference for LE is marginally significant (the p-value for the F score is 0.097) and not significant for NULL (the p-value for the F score 0.143).

### 5.3 Contextual Features

In our default set-up, each aspect marker is conditioned solely on features derived from the same verb, not on features derived from other verbs. We expanded the feature vector of each verb by including the features from the previous and the following verb; the results showed no improvement by adding these contextual features.

From the above three experiments, together with the findings from the previous section that positional features are significant in improving the overall accuracy, and that the effect of removing positional features is marginally significant between verbs with long distance and those with short distance for the LE tag, we conclude that the sequential dependency between the aspect markers of verbs in the same sentence is marginally significant.

[5] More precisely, we trained a linear-chain CRF using sequences of length one.

|   | LE | ZHE | GUO | NULL |
|---|---|---|---|---|
| P | 0.7524 | 0.5361 | 0.62 | 0.7656 |
| R | 0.8126 | 0.4451 | 0.4361 | 0.7283 |
| F | 0.7813 | 0.483 | 0.4856 | 0.7464 |

Table 7: Results with syntactic features extracted from the output of a Chinese parser.

## 6 Results with Automatically Extracted Syntactic Features

State-of-the-art parsers for Chinese are far less accurate than parsers for English. Since syntactic features are high-utility features in aspect marker prediction, from an application-oriented point of view, we are interested in seeing how well an aspect marker classifier can perform when it uses syntactic features that are extracted fully automatically by a Chinese parser. We trained Daniel Bikel's statistical parser (Bikel, 2004) for Chinese on the LDC Chinese treebank data. We used the trained Chinese parser to parse our data set and extracted the syntactic features described in Section 3 from the output of the parser automatically.

Table 7 shows the performance of the classifier with automatically extracted syntactic features using the Chinese parser along with other features. The overall accuracy is 74.33%, a drop of 3 percentage points compared to the classifier using manually annotated syntactic features. But the automatically extracted syntactic features still gain 2 percentage points in general accuracy over a classifier without syntactic features.

## 7 Conclusions and Future Work

As the very first study on Chinese aspect marker generation in English-to-Chinese translation, this paper reports on an aspect marker classifier based on a maximum entropy model with promising classification accuracy. The current study also investigates the utility of different clusters of features. The results show that syntactic features associated with the verb have the highest utility in the classification task; in contrast, lexical aspectual features have the highest utility in tense classification in the opposite scenario. Additionally, empirical evidence suggests a marginally significant dependency between the aspect markers of different verbs within a sentence.

In future work we want to extend the feature space and explore additional features that might help shorten the gap between the aspect marker classifier discussed in the current paper and human performance, such as classes of temporal expressions associated with the event described by the verb. We are also interested in investigating the efficiency of the current approach in temporal reference tagging tasks for other languages pairs.

A more interesting topic for future work would be to predict whether an aspect marker actually does appear in the text, rather than could appear, thus making hand annotation of aspect markers unnecessary. This way, the size of the training data can be much larger.

## Acknowledgements

The authors are grateful to the anonymous reviewers for many useful comments and suggestions.

## References

D. Bikel, 2004. Multilingual Statistical Parsing Engine, University of Pennsylvania, URL: http://www.cis.upenn.edu/∼dbikel/software.html

R. Campbell, T. Aikawa, Z. Jiang, C. Lozano, M. Melero, A. Wu, 2002. A Language-Neutral Representation of Temporal Information. Workshop on Annotation Standards for Temporal Information in Natural Language, LREC 2002, 13–21.

J. Cohen, 1960. A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, 20, 37–46.

B. D. Eugenio, M. Glass, 2004. The Kappa statistic: A second look. Computational Linguistics, 30(1), 95–101.

J. Lafferty, A. McCallum, F. Pereira, 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 18th International Conference on Machine Learning, 282–289.

W. Li, K. Wong, C. Hong, C. Yuan, 2004. Applying Machine Learning to Chinese Temporal Relation Resolution. 42nd Annual Meeting of the Association for Computational Linguistics, 582–588.

A. McCallum, 2002. MALLET: A Machine Learning for Language Toolkit, URL: http://mallet.cs.umass.edu/

A. McCallum, D. Freitag, F. Pereira, 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. 17th International Conference on Machine Learning, 591–598.

I. Mani, J. Pustejovsky, R. Gaizauskas. 2005. The Language of Time, Oxford University Press.

M. Olsen, D. Traum, C. Van Ess-Dykema, A. Weinberg, 2001. Implicit Cues for Explicit Generation: Using Telicity as a Cue for Tense Structure in a Chinese to English MT System. Machine Translation Summit VIII.

J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani, 2004. The Specification Language TimeML. The Language of Time: A Reader. Oxford University Press, 185–96.

H. Reichenbach, 1947. Elements of Symbolic Logic. Macmillan, New York, N.Y.

Z. Vendler, 1967. Verbs and Times. Linguistics in Philosophy, 97–121.

Y. Ye, Z. Zhang, 2005. Tense Tagging for Verbs in Cross-Lingual Context: a Case Study. Proceedings of IJCNLP 2005, 885–895.

Y. Ye, V. Li Fossum, S. Abney, 2006. Latent Features in Temporal Reference Translation. Fifth SIGHAN Workshop on Chinese Language Processing, 48–55.