Automatic Inference of the Tense of Chinese Events Using Implicit Linguistic Information

Yuchen Zhang
Brandeis University
415 South Street
Waltham, MA
yuchenz@brandeis.edu

Nianwen Xue
Brandeis University
415 South Street
Waltham, MA
xuen@brandeis.edu

Abstract

We address the problem of automatically inferring the tense of events in Chinese text. We use a new corpus annotated with Chinese semantic tense information and other implicit Chinese linguistic information using a "distant annotation" method. We propose three improvements over a relatively strong baseline method – a statistical learning method with extensive feature engineering. First, we add two sources of implicit linguistic information as features – eventuality type and modality of an event, which are also inferred automatically. Second, we perform joint learning on semantic tense, eventuality type, and modality of an event. Third, we train artificial neural network models for this problem and compare its performance with feature-based approaches. Experimental results show considerable improvements on Chinese tense inference. Our best performance reaches 68.6% in accuracy, outperforming a strong baseline method.

1 Introduction

As a language with no grammatical tense, Chinese does not encode the temporal location of an event directly in a verb, while in English, the grammatical tense of a verb is a strong indicator of the temporal location of an event. In this paper we address the problem of inferring the semantic tense, or the temporal location of an event (e.g., present, past, future) in Chinese text. The semantic tense is defined relative to the utterance time or document creation time, and it does not always agree with the grammatical tense in languages like English where there is grammatical tense. Inferring semantic tense potentially benefits natural language processing tasks such as Machine Translation and Information Extraction (Xue, 2008; Reichart and Rappoport, 2010; Ye et al., 2006; Ye, 2007; Liu et al., 2011), but previous work has shown that automatic inference of the semantic tense of events in Chinese is a very challenging task (Xue, 2008; Ye et al., 2006; Liu et al., 2011).

There are at least two reasons why this is a difficult problem. First, since Chinese does not have grammatical tense which could serve as an important clue when annotating the semantic tense of an event, generating consistent annotation for Chinese semantic tense has proved to be a challenge. Xue and Zhang (2014) use a "distant annotation" method to address this problem. They take advantage of an English-Chinese parallel corpus with manual word alignments (Li et al., 2012) , and perform annotation on the English side, which provides more explicit information such as grammatical tense that helps annotators decide the appropriate semantic tense. The annotations are then projected to the Chinese side via the word alignments. They show consistent annotation agreements on semantic tense. Second, the lack of grammatical tense also makes automatic inference of Chinese semantic tense challenging since the grammatical tense would be an important source of information for predicting the semantic tense. Previous work has shown that it is very difficult to achieve high accuracy using standard machine learning techniques such as Maximum Entropy and Conditional Random Field classifiers combined with extensive feature engineering.

We address these challenges in two ways. First of all, we take advantage of the newly annotated corpus described in (Xue and Zhang, 2014) in which semantic tense is annotated together with eventuality type and modality using the distant annotation method. This makes it possible to use these two additional sources of information to help predict tense. Eventuality type and modality are intricately tied to tense. For example, Smith and

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1902–1911,
October 25-29, 2014, Doha, Qatar.

Erbaugh (2005) show that states by default hold in the present but (episodic) events occur by default in the past. This means knowing the eventuality type of an event would help determine the tense. Eventuality type and modality are also annotated on the English side and then projected onto the Chinese side via manual word alignments, taking advantage of the rich morphosyntactic clues in English. High inter-annotator agreement scores are also reported on eventuality type and modality.

We experimented with two ways of using eventuality type and modality information. In the first approach, we first train statistical machine learning models to predict eventuality type and modality and then use these two sources of information as features to predict semantic tense. In the second approach we trained joint learning models between semantic tense and eventuality type, and between semantic tense and modality. We show both approaches improve the tense inference accuracy over a baseline where these two sources of information are not used. Second, in our statistical machine learning experiments on tense inference using feature engineering, we find that the design of feature templates has great influence on the results. So in order to explore more possible feature combinations and mitigate the feature engineering work, we apply artificial neural network models to this problem. This shows improvements on tense inference accuracy as well in some of the experiment settings.

The rest of the paper is organized as follows. Section 2 discusses related work in automatic tense inference. Section 3 briefly introduces the distant annotation method. In section 4, we describe our experiments and analyze the experimental results. We conclude this paper in section 7.

2 Related Work

Inferring the semantic tense of events in Chinese text is not a new topic. There have been several attempts at it, yet high accuracy in this task has proved to be elusive. Using a corpus with tense annotated directly in Chinese text, Xue (2008) performed extensive feature engineering in a machine learning framework to address this problem. They used both local lexical features and structured features extracted from manually annotated syntactic parsing trees. In our baseline method, we adopt most of their features as the baseline, only on a new corpus in which semantic tenses are not annotated directly on Chinese events but projected from annotations from the English side of a parallel Chinese-English corpus. In our experiments, we also use structural features extracted from automatic parse trees, so our experimental settings are more realistic.

Ye et al. (2006) took a similar approach in which they predict tense with feature engineering in a statistical learning framework. They also used a Chinese-English parallel corpus and projected tense for English events onto Chinese events via human alignments. The main difference between their data and ours is that they used the grammatical tense of the English events, while we use human-annotated semantic tense which we believe are more "transferrable" across languages as it is free of the language-specific idiosyncrasies of grammatical tense. In addition, they also used human annotated linguistic information as "latent" features in their work, which are similar to our implicit linguistic features. However, the "latent" features that they used in their system are human-annotated, while the eventuality type and modality features in our system are predicted automatically. Another difference is that they ignored events that are not verbs. For example, they excluded verbal expressions in Chinese that are translated into nominal phrases in English. In contrast, we kept all events in our data, and they can be realized as verbs, nouns, as well as words in other parts of speech. We performed separate experiments on events realized as verbal expressions and events not in verbal expressions to investigate their impact on semantic tense inference.

Liu et al. (2011) introduced more global features in a machine learning framework, and on top of that proposed an iterative learning algorithm which better handles noisy data, but they also ignored events that are not realized as verbal expressions, or events that are verbal expressions but have more than one verb in them. They mainly focused on events that are one-verb expressions.

In a similar work on inferring tense in English text, Reichart and Rappoport (2010) aimed at inferring fine-grained semantic tenses for events in English. They introduced a fine-grained sense taxonomy for tense in a more general Tense Sense Disambiguation (TSD) task to annotate and disambiguate semantic tenses. The underlying senses include "things that are always true", "general and repeated actions and habits", "plans, expectations

and hopes", etc., which encode a combination of tense, eventuality type and modality. In the corpus that we use, the same information is organized in a more structured manner along three dimensions – semantic tense, eventuality type, and modality.

3 Distant Annotation

Figure 1 shows the distant annotation procedure from (Xue and Zhang, 2014). Starting with a word-aligned parallel English-Chinese corpus, all sentences are part-of-speech (POS) tagged first and then all verb instances in the English text as well as expressions aligned with verb instances on the Chinese side are targeted for annotation. As we will show in Section 4, these expressions include verbs as well as nouns, prepositions and word sequences "headed" by a verb. We consider those expressions as events. Annotators work only on the English side and tag every event with a pre-defined semantic tense label. These labels are then projected from the English side to the Chinese side via word alignments. The resulting corpus contains events annotated with semantic tense labels in both languages. Categories for semantic tense are "Past", "Present", "Future", "Relative Past", "Relative Present", "Relative Future", and "None".

Events annotated with relative tenses are also linked to another event that serves as the temporal reference for the event in question. In some cases the relative tense can be resolved to an absolute tense. For example, if an event is annotated with a "relative past" tense to a reference event that is annotated with a present tense, then the semantic tense of that event can resolve to an absolute "past" tense. In other cases, they can not be resolved. For example, if an event is labeled with a "relative future" tense and the reference event has a past tense, then its tense cannot be resolved to an absolute tense, which is defined with regard to the utterance time or document creation time. In our work, where possible, we resolve these links and keep only absolute tense labels. For events with relative tenses that can not be resolved (i.e. events which are "Relative Future" to "Past" events, or events which are "Relative Past" to "Future" events), we use "None" as the default label.

Eventuality type and modality are labeled in the same way as auxiliary annotation that can help with the inference of tense. Labels for eventuality type include "Episodic", "Habitual", "State", "Progressive", "Completed", and "None". Labels for modality are "Actual", "Intended", "Hypothetical", "Modalized", and "None". Readers are refered to (Xue and Zhang, 2014) for detailed explanations of each label.

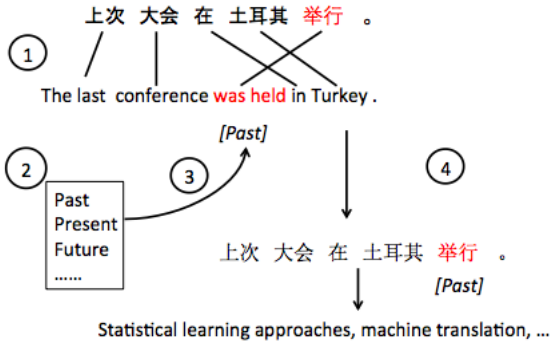

Figure 1: Distant annotation procedure.

As we mentioned in Section 2, in this corpus not only verbs but also their counterparts on the opposite language are considered as events, yielding events that may not be verbs. For example, in the following sentence pair (1), the Chinese verb (VV) "利用" is aligned with an English noun (NN) "use". In the sentence pair (2), the English verb (VBG) "opening" is aligned with an Chinese noun (NN) "开放".

(1) Statistics show that , in the past five years , Guangxi's foreign trade and its **use** of foreign investments has expanded rapidly.

统计资料显示，过去五年广西对外贸易和**利用(li4yong4)**外资规模迅速扩大。

(2) Beihai has already become a bright star arising from China's policy of **opening** up to the outside world.

北海已成为中国对外**开放(kai1fang4)**中升起的一颗明星。

In this corpus, events could be either one verb, or a verb compound, or a verb sequence "headed" by a verb, or even nouns and words of other parts of speech.

4 Experiments

4.1 Experimental Setting

Xue and Zhang (2014) annotated semantic tense, eventuality type and modality on top of the Parallel Aligned Treebank (Li et al., 2012), a corpus

of word-aligned Chinese-English sentences treebanked based on the Penn TreeBank (Marcus et al., 1993) and the Chinese TreeBank (Xue et al., 2005) standards. Human annotation of tense is performed on the newswire and webblog sections of this corpus. They report that the average pairwise agreement among three annotators consistently stays above 80% and the average Kappa score consistently exceeds 70%, indicating reliable annotation.

Apart from using the entire corpus, we also conducted experiments on three different subsets of the corpus. An examination of the data indicates that newswire data is grammatically more formal and complete than webblog data, so we also conducted separate experiments on newswire data only. Considering that the diversity of the parts of speech of the events may affect the inference accuracy and that most of our features extracted from the parse trees assume that our events being verbs, we also conducted experiments exclusively on "v_events". "v_events" consist of two parts. One part is events that are realized as a single word and the word is a verb; the other one is events which have multiple words but there is only one verb among them. In the latter case, we stripped off words tagged with other parts of speech and only keep the verbs as events. This makes it more effective to use features from previous work that are designed for single verbs. One such feature is the aspect marker. Distinctions between newswire and webblog data and between v_events and other events are further explored in Section 5.1 and Section 5.2. Table 1 presents the statistics for each subset of the experimental data.

dataset	# of v_events	# of all events
nw	6,686	8,268
all	17,153	20,885

Table 1: Statistics of four subsets of the annotated corpus (Chinese side). "nw" denotes the newswire data. "v_events" denotes events that consist of or can be reduced to only a single verb.

For each subset, randomly selected 80% were used as the training set, while 10% were used as the development set and 10% were used as the test set.

4.2 Baseline

Based on previous approaches on Chinese tense inference, we used a Maximum Entropy model with extensive feature engineering as our baseline method. We use the implementation of the Maximum Entropy algorithm in Mallet [1] for our experiments. The corpus is parsed using the Berkeley Parser for the purpose of extracting structure features. Since the Parallel Alignment TreeBank is a subset of the Chinese TreeBank (CTB) 8.0, we automatically parsed the CTB 8.0 by doing a 10-fold cross validation. The bracketing F-score is 80.5%. Feature extractions are performed on the automatic parse trees. Adopted features include previous word and its POS tag, next word and its POS tag, aspect marker following the event, 得following the event, the governing verb of the event, the character string of the main verb in the previous clause that is coordinated with the clause the event is in, whether the event is in quote, and left modifiers of the event including head of adverbial phrases, temporal noun phrases, prepositional phrases, localizer phrases, as well as subordinating clauses. Readers are referred to (Xue, 2008) for details of these features. Since in this corpus an event can span over more than one verb, we also use the character string and the POS string of the entire event instead of one word and one POS tag as features.

- The character string of an event – it could be one or more words. In our corpus, only 69.7% events consist of single word (e.g. "居住", "live"), the other 30.3% of the events are expressed with two or more words (e.g. "引发+了", "have caused").
- The POS string of an event – it could be verbs, nouns, or POS sequences of other word sequences. Table 2 shows the top ten POS tag or POS tag sequences with example word or word sequences.

Other features that we used in the baseline system are as follows.

- DEC – if the word immediately following an event has the POS tag "DEC", use its character string as a feature. In most cases, "DEC" is the POS tag for "的" when it used as a complementizer marking the boundary in a relative clause. This feature implies that an event

[1] http://mallet.cs.umass.edu/

POS	freq	examples
VV	48.2%	居住(live)
NN	5.8%	开放(opening)
VC	5.2%	是(is)
VV+AS	5.2%	引发+了(have caused)
VV+DEC	3.2%	隔绝+的(isolated)
AD+VV	3.0%	正在+建议(is suggesting)
VA	3.0%	大(is big)
AD	2.0%	似乎(seemed)
VE	1.9%	有(there is)
P	1.8%	根据(according to)

Table 2: Frequencies and examples of the ten most frequent POS tag or POS tag sequences for events in our corpus.

is inside a relative clause modifying a noun phrase and it is more often stative than eventive.

- Determiners – we find the subject of an event from its parse tree and extract the determiner of the subject, if there is one, as a feature. This feature indicates different types of agents, and different types of agents often signal different types of events. For example, individual agents tend to perform one-time episodic actions which are by default located in the past or described by a state in the present, while multiple agents tend to involved in habitual actions that spans over a long period of time.

Baseline results are reported in Table 5 and Table 6, in MaxEnt_b rows.

4.3 Eventuality Type and Modality as Features

Xue and Zhang (2014) reports that gold eventuality type and modality labels significantly help the inference of tense in Chinese, improving the accuracy by more than 20%. However, it is unrealistic to expect to have human annotated eventuality type and modality labels in a random new data set if we want to use these two sources of implicit linguistic information in any Chinese text. So we trained statistical learning models to automatically extract these two labels. We trained Maximum Entropy models and ran a 10-fold cross validation on the entire corpus in order to get automatic labels for every event. Feature used for labeling modality are as follows. Table 3 shows the average accuracies for automatic modality labeling.

- The character string of an event.
- The POS string of an event.
- The character string of an event's governing verb and its POS tag.
- Whether the event is in a conditional clause. If an event is in a subtree with the functional tag "CND", return "True"; otherwise, return "False". This feature indicates that the event's modality label is "Hypothetical".
- Whether the event is in a purpose or reason clause. If an event is in a subtree with the functional tag "PRP", return "True"; otherwise, return "False" as a feature. This feature indicates the event's modality label is "Intended".
- Whether the event string is the start of a sentence. If an event is the start of a sentence, return "True"; otherwise, return "False". Sentences that start with an event is often imperative, and the event generally has "modalized" modality label.

dataset	v_events	all events
nw	81.1%	81.2%
all	75.4%	76.4%

Table 3: Average modality labeling accuracy, using a 10-fold cross validation.

Statistics show that the five labels for modality have a skewed distribution in this corpus. Among all events, 67.3% of them fall in the "Actual" category, while the events of all the other categories are around or less than 10%. Similar distributions are found in all four subsets of the data. Still, compared with always choosing the most frequent label (around 67% accuracy), we still get a big improvement from our statistical model, even though only a very simple set of features are used.

Features used for labeling eventuality type are as follows. Table 4 shows the average accuracies for automatic eventuality type labeling.

- The character string of an event.
- The POS string of an event.

- Adverbs on the left that modifies the event.
- Aspect marker following the event
- Whether the event is Inside a relative clause. If an event is in a CP subtree with the word "的" and POS tag "DEC" as its last node, return "True"; otherwise, return "False". Events in relative clauses modifying a noun phrase and tend to be more often stative than eventive.

dataset	v_events	all events
nw	68.7%	67.7%
all	65.3%	65.1%

Table 4: Average automatic eventuality type labeling accuracy using a 10-fold cross validation.

The six labels of eventuality type are also distributed unevenly. The first group of columns in Figure 3 shows the distribution of all events. Over 65% of events are either "Episodic" or "State", while the other types of events are less than 15%. There are two categories that are even less than 5%. However, even though we only use some simple features, our model still beats the most frequent label baseline (around 35% accuracy) by a big margin, as shown in Table 4.

Tense inference accuracies using automatic eventuality type and/or modality features are reported in Table 5 and Table 6, in MaxEnt_e, MaxEnt_m, and MaxEnt_em rows.

4.4 Joint Learning

Apart from using eventuality type and modality labels as features, we also conducted joint learning experiments on them. Joint learning are applied on 1) tense and eventuality type, and 2) tense and modality. Features used are the union of the two sets of features in inferring each single label. MaxEnt_jle and MaxEnt_jlm rows in Table 5 and Table 6 present the experimental results on joint learning.

4.5 Artificial Neural Network

For each of the experiments using the maximum entropy algorithm, we conducted a neural network experiment using the same setting in order to explore more possible feature combinations and mitigate the feature engineering work. We convert the features in each of our tense inference methods into feature vectors. If a feature is not a word, we use a one-hot representation for that feature (a vector with all 0s except for a 1 at the place of the feature's index in our feature lexicon). If a feature is a word, we convert it into a word embedding. To get a dictionary of word embeddings, we use the word2vec tool [2] (Mikolov et al., 2013) and train it on the Chinese Gigaword corpus (LDC2003T09). For each word embedding, a 300-dimensional vector is used. Artificial neural networks are built using the theano package [3] (Bergstra et al., 2010). We use 5000 hidden units for all networks and set the learning rate α = 0.01. Experimental results are presented in the ANN rows of Tables 5 and 6.

5 Results Analysis

A comparison of the baseline accuracy for the four different subsets of the data shows that (1) tense inference is slightly better on v_events than on all events, but the difference is not substantial; and (2) tense inference on newswire data performs better than on all data by around 8% on v_events and around 5% on all events, verifying our assumption that automatic tense inference is easier on newswire data than webblog data. Although our experiments are performed on different data sets from that of previous work, our baseline method still shows strong results compared with previous work (Xue, 2008; Ye et al., 2006; Liu et al., 2011).

Adding automatic eventuality type and modality labels as features for semantic tense inference leads to improvements over the baseline on all four data subsets. In fact they provide considerable improvements (around 2% increase) on newswire v_events dataset. MaxEnt_e rows report results when only automatic eventuality type is added as a feature, and MaxEnt_m rows report results when only automatic modality is added as a feature. They both outperform (or, in several datasets, match) the baseline results on all datasets. MaxEnt_em rows report results when both automatic linguistic labels are added as features, and they show further improvements over when only one source of information is used. Analysis of the results shows again that tense inference accuracy is higher than webblog data under this experiment condition. The results also show that after adding eventuality type and modality as features,

[2] http://code.google.com/p/word2vec/
[3] http://deeplearning.net/software/theano/

method	all data	nw data
MaxEnt_b	58.9%	66.8%
MaxEnt_e	59.5%	67.9%
MaxEnt_m	59.5%	67.1%
MaxEnt_em	59.6%	**68.6%**
MaxEnt_jle	59.6%	63.5%
MaxEnt_jlm	**60.5%**	66.9%
MaxEnt_ge	74.6%	**77.4%**
MaxEnt_gm	66.6%	70.0%
MaxEnt_gem	**76.2%**	76.9%
ANN_b	63.4%	67.2%
ANN_e	62.6%	66.1%
ANN_m	**63.4%**	59.8%
ANN_em	59.7%	**68.3%**
ANN_jle	62.7%	64.5%
ANN_jlm	62.0%	65.6%

Table 5: Accuracy of tense inference on v_events. Best performances for each group of methods are in bold.

method	all data	nw data
MaxEnt_b	59.7%	65.1%
MaxEnt_e	59.9%	65.1%
MaxEnt_m	59.9%	65.4%
MaxEnt_em	59.9%	65.5%
MaxEnt_jle	59.7%	62.7%
MaxEnt_jlm	**60.4%**	**65.6%**
MaxEnt_ge	75.3%	**76.1%**
MaxEnt_gm	67.1%	69.0%
MaxEnt_gem	**76.2%**	75.9%
ANN_b	63.0%	64.0%
ANN_e	**63.2%**	**66.9%**
ANN_m	60.1%	64.7%
ANN_em	57.8%	66.1%
ANN_jle	61.4%	63.0%
ANN_jlm	62.9%	63.5%

Table 6: Accuracy of tense inference on all events. Best performances for each group of methods are in bold.

the improvements on v_events (0.7% and 1.8%) are much bigger than that on all events (0.2% and 0.4%), regardless of the data genre (newswire or weblog).

In order to test the potential for these two new features, we also conducted experiments using gold eventuality type and/or modality labels as features for the Maximum Entropy models (Table 5 and Table 6, MaxEnt_ge, MaxEnt_gm, and MaxEnt_gem rows.). They outperform our best MaxEnt results by around 10% on newswire data and around 15% on all data, indicating strong potentials for more accurately classified automatic eventuality type and modality labels.

Results also show that joint learning with modality proves to be working better than the baseline (Table 5 and Table 6, MaxEnt_jle, MaxEnt_jlm). In fact, on the datasets with all events, joint learning with modality produces the highest accuracy among all approaches. However, joint learning with eventuality is even worse than the baseline. One possible explanation is that the lower eventuality type classification accuracy affects the tense inference accuracy. We also believe there is still room for improvement with features tuned for the joint learning model. Simply adding the features may not be the best strategy.

On the entire dataset, regardless of v_events or other events, results of the neural network models show improvements over the maximum entropy models under most experimental conditions. A clear trend is that artificial neural networks help more on all data than on newswire data only, indicating greater potentials of the neural network models to select and combine features with carefully trained parameters, given noisier but larger training sets.

Experimental results also show significant differences in accuracy between newswire data and webblog data, and smaller but still recognizable difference between v_events and all events. Therefore, we specifically look into distinctions between these data sets.

5.1 Newswire Data vs. Webblog Data

Considering the big gap in accuracy between newswire and webblog data in our baseline results, we delve deeper into the data and found several major distinctions between these two domains that might have contributed to the rather significant difference in performance on tense inference. First, we look into the word frequency distribution of the two datasets. Here by "word" we mean the character string of an event. We find that both datasets have a small portion of words with high frequencies, but the webblog dataset contains much more low-frequent words than the newswire dataset. In Figure 2, the x-axis shows possible frequencies of words and the y-axis shows the number of words at a particular frequency. It can be seen that the num-

ber of words that appear only once in the webblog dataset is about three times as large as that in the newswire dataset. The entire newswire dataset has a vocabulary of only 2671 entries, while the webblog dataset has a vocabulary size of 6117. This greatly reduces the coverage of features extracted from the training dataset on the events in the test dataset.

Second, webblog data contains more events that are "inherently" ambiguous on temporal location. Among four possible labels for tense in this corpus, "None" is for events whose temporal locations are not clear even to human annotators. Statistics show that in webblog data about 13.4% of the events are tagged as "None", while in newswire data only around 6.7% are "None". Another piece of evidence showing webblog data is harder to process is the different inter-annotator agreement scores for tense annotation on newswire and webblog data reported by (Xue and Zhang, 2014). Newswire data has a 89.0% agreement score and a 84.9% kappa score, while webblog data only has a 81.0% agreement score and a 72.7% kappa. Third, automatic parse trees for newswire data is also more accurate than that for webblog data. The bracketing F-score of automatically parsed newswire data is 83.0% while it is only 80.4% for weblog data. Moreover, sentences in newswire data are more grammatically complete. Analysis shows that webblog data has more dropped constituents in sentences. There are around 40.5% sentences in newswire data that have nominal empty categories, while in webblog data the number is 48.1%. Dropped constituents affect the structures of parse trees and some of the features, which can affect tense inference accuracy.

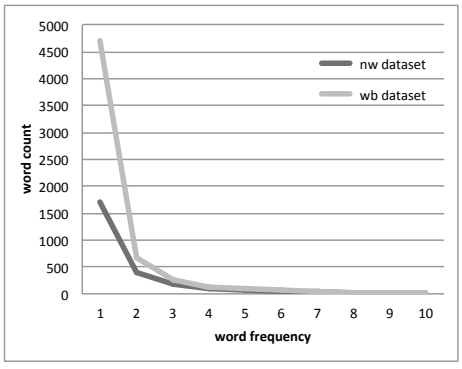

Figure 2: Word frequency distribution in newswire and webblog datasets.

5.2 V_events vs. All Events

In our definition, v_events are (1) events that are single verbs (example 1, 3, 7, 9 in Table 2), and (2) events that are multi-word sequences but only one word among them is a verb and any non-verb words are stripped off (verbs in example 4, 5, 6 in Table 2). Conversely, events that do not fall into this definition include (1) events that have no verbs in their surface form (example 2, 8, 10 in Table 2), and (2) events that have more than one verb in their surface form (e.g. "使+成为(shi3+cheng2wei2)", VV+VV, "make it become"). So from the point of view of a statistical learning algorithm, every v_event has one and only one verb. This makes sure that all features that we used are applicable to v_events. For other events, however, some features may be not applicable. For example, for an event which has a nominal expression, aspect marker, DER, and DEC features are all "None" because these features are only applicable to verbs. Another major distinction between v_events and "other events" is that the distributions of eventuality type labels on them are very different, presented in the second and third groups of columns in Figure 3. There is a rather high percentage of "State" among "other events" and very low percentage of "Completed" and "None". The highly uneven distribution of eventuality type labels make it less effective as a feature for tense inference.

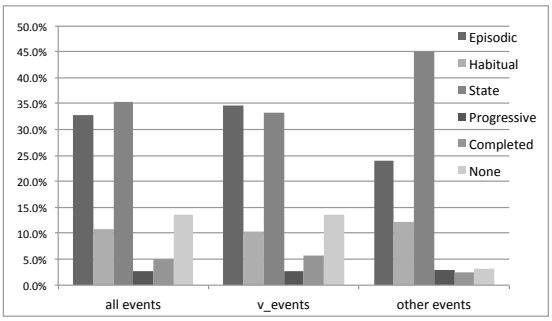

Figure 3: Statistics of eventuality types on different events.

We also find that, on newswire datasets, maximum entropy models and neural network models do not show much difference in performance. To understand this result better, we plot learning curves of the artificial neural network model, trained and tested on newswire v_events dataset. In Figure 4, the black line represents the error rate on training set, and the grey line represents the er-

ror rate on test set. As the size of training data grows, the error rate on the training set gets larger because with more training examples the training set becomes noisier and it gets harder to model all samples with the same number of features; and the error rate on the test set gets smaller because a bigger training set reduces the data sparsity and trains the parameters better. Both lines end at a rather high error rate (around 30%, i.e. only around 70% in accuracy) which means the current network is general enough to cover most cases in the test set, but it is under-fitting the training data. The current model is not specific enough to better capture the fine distinctions between the tense categories. The black line being not very smooth is also understandable, given that there are only around 6000 training examples in the newswire v_events dataset.

Figure 4: Learning curves of the artificial neural network model, trained and tested on newswire v_events dataset.

6 Error Analysis

In order to get a better understanding of the use of eventuality type and modality, we look into the error rates for each error type in greater detail. In Table 7, "Pa" stands for "Past", "Pre" is short for "Present", "Fu" is for "Future", and "No" is "None". For each error type, the left-hand side is the gold-standard tense, and the right-hand side is the wrongly assigned label. Statistics are collected on the newswire v_events data test set. Table 7 compares the different error types between the baseline method and the MaxEnt_em method, the best approach for this dataset. We can see that (1) "Present" and "Past" is the most frequently confused tense pair, and (2) eventuality type and modality information help disambiguate "Present" and "Past" events greatly, and reduce the errors due to mis-classifying "Past" as "Future", or "Future" as "Present", or "None" as "Present".

error type	MaxEnt_b	MaxEnt_em
Pre → Pa	11.7%	11.2%
Pa → Pre	9.8%	9.2%
Pa → Fu	2.5%	2.3%
No → Pa	2.0%	2.0%
Fu → Pre	1.9%	1.6%
Pre → Fu	1.4%	1.4%
Fu → Pa	1.4%	1.4%
No → Pre	1.6%	1.2%
No → Fu	0.5%	0.5%
Pa → No	0.3%	0.3%
Pre → No	0.2%	0.2%
Fu → No	0.0%	0.0%

Table 7: Tense inference error rates for different error types on newswire v_events test set.

A closer examination of the sentences in which events are assigned the wrong tense reveals that "Pre → Pa" error is prone to occur on events in relative clauses. The Chinese verb implies a past episodic event, while the event is actually a present state or habitual event. As a good example, the "生产(sheng1chan3)" event in Sentence (3) is wrongly labeled as "Past" by MaxEnt_b but correctly classified as "Present" by MaxEnt_em with eventuality type "Habitual" and modality tag "Actual" (the underlined part in the Chinese sentence is the relative clause). It is also found that most "Pa → Pre" errors occur on events that are more stative. It is reasonable since classifiers tend to assign "Present" to states and "Past" to episodic events. MaxEnt_em managed to correct some with "episodic" as their correct eventuality type.

(3) 目前该区**生产(sheng1chan3)**此疫苗的 普康公司已形成年产五百万人份的生产规模，这对有效地控制甲肝流行具有重大意义。

At present , the Pu Kang Company , which **produces** the vaccine in this zone , has already formed a production scale of 5 million doses per year , which has great significance in effectively controlling the hepatitis A epidemic .

We are also surprised to see that over 2% "Past" events are classified as "Future" events, ranking

third among all error types. This mistake seems very unlikely, but it is still possible when performing tense inference on a language with no grammatical tense at all. Take the following sentence pair (4) as an example. In the Chinese sentence, MaxEnt_b classifies "讨论(tao3lun4)" as "Future" because there is no grammatical indicator in the Chinese sentence implying that the "discussion" has already happened and it is reasonable to assume the "discussion" is in the near future. However, with eventuality type "Episodic" and modality label "Actual", MaxEnt_em classifies it as "Past" correctly, because episodic events tend to occur in the past and future events tend to get "Intended" or "Hypothetical" modality labels.

(4) 他还说，法国政府"甚至指示它的代表，在联合国安理会**讨论(tao3lun4)** 制裁古巴的议案时不要投赞成票"。

He also said, the French government "even directed its representative not to vote Yes when the Security Council **discussed** the resolution on sanctions on Cuba".

7 Conclusion and Future Work

In this paper, we address the problem of automatic inference of Chinese semantic tense. We took advantage of a new corpus annotated with rich linguistic information, and experimented with three approaches. In the first approach, we use two sources of implicit linguistic information, eventuality type and modality, automatically derived, as features in tense inference. We then conducted joint learning on tense and each of these two information types. Finally, we experimented with using artificial neural networks to train models for tense prediction. All three approaches outperformed a strong baseline, a maximum entropy model with extensive engineering. Our future work will include exploring ways to improve automatic eventuality type and modality labeling accuracy to further improve tense inference accuracy.

Acknowledgments

We would like to thank the three anonymous reviewers for their suggestions and comments. This work is supported by the National Science Foundation via Grant No. 0910532 entitled "Richer Representations for Machine Translation". All views expressed in this paper are those of the authors and do not necessarily represent the view of the National Science Foundation.

References

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June. Oral Presentation.

Xuansong Li, Stephanie Strassel, Stephen Grimes, Safa Ismael, Mohamed Maamouri, Ann Bies, and Nianwen Xue. 2012. Parallel Aligned Treebanks at LDC: New Challenges Interfacing Existing Infrastructures. In *Proceedings of LREC-2012*, Istanbul, Turkey.

Feifan Liu, Fei Liu, and Yang Liu. 2011. Learning from chinese-english parallel data for chinese tense prediction. In *Proceedings of the 5th International Conference on Natural Language Processing*, pages 1116–1124, November.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Roi Reichart and Ari Rappoport. 2010. Tense sense disambiguation: A new syntactic polysemy task. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 325–334, Cambridge, MA, October. Association for Computational Linguistics.

Carlota S. Smith and Mary Erbaugh. 2005. Temporal interpretation in Mandarin Chinese. *Linguistics*, 43(4):713–756.

Nianwen Xue and Yuchen Zhang. 2014. Buy one get one free: Distant annotation of chinese tense, event type, and modality. In *Proceedings of LREC-2014*, Reykjavik, Iceland.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207–238.

Nianwen Xue. 2008. Automatic Inference of the Temporal Location of Situations in Chinese Text. In *EMNLP-2008*, Honolulu, Hawaii.

Yang Ye, Victoria Li Fossum, and Steven Abney. 2006. Latent features in automatic tense translation between Chinese and English. In *The Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing*, Sydney, Australia.

Yang Ye. 2007. *Automatic Tense and Aspect Translation between Chinese and English*. Ph.D. thesis, University of Michigan.