Buy one get one free:
Distant annotation of Chinese tense, event type, and modality

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

Abstract

We describe a “distant annotation” method where we mark up the semantic tense, event type, and modality of Chinese events via a word-aligned parallel corpus. We first map Chinese verbs to their English counterparts via word alignment, and then annotate the resulting English text spans with coarse-grained categories for semantic tense, event type, and modality that we believe apply to both English and Chinese. Because English has richer morpho-syntactic indicators for semantic tense, event type and modality than Chinese, our intuition is that this distant annotation approach will yield more consistent annotation than if we annotate the Chinese side directly. We report experimental results that show stable annotation agreement statistics and that event type and modality have significant influence on tense prediction. We also report the size of the annotated corpus that we have obtained, and how different domains impact annotation consistency.

Keywords: Chinese, Tense, Annotation

1. Introduction

Chinese does not have grammatical tense and there have been several attempts at inferring a notional or semantic tense that could potentially benefit natural language processing tasks such as Machine Translation and Information Extraction (Xue et al., 2008; Reichart and Rappoport, 2010; Ye et al., 2006; Ye, 2007; Xue, 2008; Liu et al., 2011). Such a task has been proven to be very challenging. The first challenge is to generate consistently annotated data. Due to the lack of grammatical tense, which could be an important clue for the semantic tense, annotators often have a hard time determining what the appropriate tense should be in a given context. The second challenge is to automatically infer tense with high accuracy, in the absence of grammatical tense markers. We describe an annotation framework aimed at addressing both challenges. To address the first challenge, we adopt a “distant annotation” method where we perform the annotation on the side of a parallel corpus that has richer observable information and project the annotation to the other side that has less such information. In this case, we perform tense annotation on the English side of a parallel Chinese-English corpus and transfer the annotation to the Chinese side[1]. To address the second challenge, we set up an annotation framework where we also annotate event type and modality in the same text to support the inference of tense, based on the observation that event type and modality are crucial in inferring tense in the absence of morphological markers tense. Smith and Erbaugh (2005), for example, shows that states by default hold in the present but (episodic) events occur by default in the past. Events and states, however, are “hidden” information that is not directly observable, and would have to be annotated in order for them to be used in the task of tense inference. The annotation of event type and modality is also performed on English side of the parallel data to take advantage of the more explicit morpho-syntactic clues in English. We show that with our approach we are able to annotate tense with much better consistency, and event type and modality are important sources of information for the task of inferring semantic tense.

The rest of the paper is organized as follows. In Section 2., we describe our annotation procedure in greater detail. In Section 3., we describe our specifications for annotating semantic tense, eventuality type and modality. We present some experimental results in Section 4. that show that we are able to perform this type of annotation with reasonable consistency and that the manually annotated event type and modality are informative indicators for the purpose of predicting tense. We discuss related work in Section 5., and conclude in Section 6..

2. Annotation Procedure

We start with an example that illustrates the gap in the availability of morpho-syntactic information between Chinese and English. In (1), the Chinese verb “举行(ju3xing2)” has no morphological inflections of tense. In contrast, in the English translation of the sentence, tense was grammaticalized in the form of a morpho-syntactic marker on “was held”:

(1)	上次	大会	在土耳其	举行(ju3xing2)	。
	last time	conference	in Turkey	hold	.

“The last conference was held in Turkey.”

Annotating tense consistently in Chinese has been proved to be a challenge (Xue et al., 2008) and we hypothesize that we are more likely to obtain consistent annotation by annotating the English translation rather than the Chinese source directly because the morpho-syntactic clues in English are good indicators of tense and they constrain the

[1] One reviewer pointed out that the quality of the translation may impact the quality of the projected annotation, and we agree tha this is a legitimate concern. This means one needs to make sure that the translation is of high quality before one can attempt such an approach

choices that an annotator has to make during the annotation process. Figure 1 shows our annotation procedure. With a word-aligned Chinese-English parallel corpus, we will be able to annotate tense on the English side and project our annotation back to Chinese. In doing this, we end up with tense annotation on both sides. To do this, we first identify all English text spans that are aligned to a Chinese verb in a word-aligned parallel Chinese-English corpus. Then all the English text spans will be annotated with tense, event type and modality. Note that the resulting English text spans after such mapping may not necessarily be English verbs because a Chinese verb may be translated into an English noun, or words of other parts of speech. Nevertheless, such English text spans can still be treated as "anchors" of tense, event type and modality and be annotated.

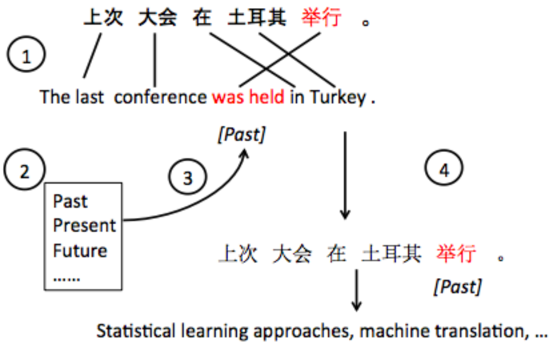
上次 大会 在 土耳其 举行 。
1
The last conference was held in Turkey .
[Past]
2
3
4
Past
Present
Future
......
上次 大会 在 土耳其 举行 。
[Past]
Statistical learning approaches, machine translation, ...

Figure 1: Annotation procedure.

3. Specifications

As described in Section 2., each Chinese verb instance is mapped to a text span in English and then annotation is performed on English text by labeling these text spans with tense and modality categories. Each text span is annotated along three dimensions to support the planned automatic inference of tense and modality on the Chinese side. The first dimension is the *semantic tense*, and the annotator must indicate whether the text span describes a past, present, or future event or state. The second dimension is *event type* that indicates whether the text span represents a habitual, on-going, or episodic event, or a state. The third dimension is *modality*. The modality dimension is broadly construed and it classifies events or states as actual, intended (which encompasses expected, planned events), hypothetical (as in conditional clauses) or modalized. An event or state is modalized if it occurs with a modal verb that indicates possibility, necessity, or ability. These categories are very coarse-grained and we did not get into the finer distinctions of different types of modality. Each of these categories are described in greater detail below.

3.1. Tense

We regard "semantic tense" as the actual occurrence time of an event with respect to the document creation time, and it may or may not be the same as the grammatical tense. In English, it is possible for a grammatical present tense to indicate a future event, in which case the semantic tense is future. This is illustrated in (2), where although "starts" has an inflectional suffix that indicates present grammatical tense, the meeting takes place in the future. The reason for annotating semantic tense rather than grammatical tense is that we expect the semantic tense may transfer across languages but the grammatical tense may not.

(2) The meeting starts at 4pm this afternoon.

We set up six categories for semantic tense, and these are *past*, *present*, *future*, *relative past*, *relative present* and *relative future*. And we claim that these categories are extensive and cover all possible semantic tenses.

Past The text span describes an event or state that happened in the past.

(3) He started an engineering firm and worked with contractors such as ABB and Kellogg , Brown and Root.

Present The text span describes a present event or state. This includes a present state, an event that happens repeatedly in the present, or an on-going event.

(4) It is centered on the Hongshui River hydroelectric plant .

Future The text span describes an event that will happen in the future, or a future state.

(5) Some people will prefer that option because it 's more convenient .

When annotating the semantic tense, some events or states cannot be interpreted in relation to the document creation time and we have to annotate its relative tense. In such cases, we also link this text span to another that it depends on for its temporal interpretation. These links are all in the direction from the *dependent* text span to its *head* span. Such dependent text spans can be tagged as *Relative Past, Relative Present, or Relative Future* when annotating *tense*, and they are typically tagged as *Intended* or *Modalized* when annotating modality.

Relative Future The text span describes an event that happens in the future relative to the event it depends on. In (6), "to strengthen" depends on "has invested" for its temporal interpretation, and "to create" depends on "to strengthen".

(6) It *has invested* more than 130 billion yuan to strengthen the construction of infrastructures so as to create a sound environment .

Relative Present The text span describes an event that happens in the present, or a present state relative to the event it depends on. In (7),"taking up" happens at the same time with "'ve got", i.e. "taking up" is relatively present to "'ve got".

(7) I *'ve got* two dead monitors taking up space in my office .

Relative Past The text span describes an event that happened in the past, or a past state relative to the event it depends on. In (8),"crossing" happened before the time "repeated" happened, i.e. "crossing" is relatively past to "repeated".

(8) After crossing a 30 - foot no man's land we *repeated* the process at the second wall .

Even in English, annotating tense can be challenging in at least three scenarios, and the first one being when there is a mismatch between the grammatical tense and the semantic tense, as illustrated in (2). In this case, the grammatical tense can be deceiving and can be an impediment that prevents the annotator from making the correct decision. The other scenario is when the text span is a verb that takes on a non-finite form. When this happens, tense is underspecified, just like in Chinese. In this case, we find paraphrasing to be a useful tool that helps the annotator make a determination. Where possible, we paraphrase the non-finite form as a finite form, and use that to help make the right decision. For example, in (9), "arising" can be paraphrased as "that is arising", and therefore it should receive the present tense. The third scenario is when a Chinese verb is translated into an (eventive) noun, and obviously English nouns do not have a tense inflection, and context is needed to determine the correct semantic tense.

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

3.2. Event type

We define four event types, and these are *habitual event*, *state*, *on-going event*, and *episodic event*. The event type is set up as a way to help infer tense. Habitual events, on-going events, and states, for example, tend to occur in the present by default, while episodic events tend to occur in the past by default (Smith and Erbaugh, 2005). Given that there is no grammatical tense in Chinese, such a classification may prove to be an important source of information that helps predict tense. Each of the four types is described and illustrated below, and the relevant text spans are underlined:

Habitual Event The text span describes an event that happens repeatedly on a regular basis.

(10) a. I used to drive to work but now I take the bus.

State The text span describes an unchanging situation that will continue unless something happens to change it.

(11) Each enterprise entering this zone has one or more new, high-level technology projects or products.

On-going Event The text span describes an event that is on-going. The progressive aspect marker is generally a good indicator of this type of event.

(12) At the school, where Bush was reading a story to a group of second-graders, the news came on TV that a second jet had hit the World Trade Center.

Episodic Event The text span describes a situation that involves some sort of change or occurrence in a relatively short period of time.

(13) The actual use of foreign investments added up to 3.324 billion US dollars.

3.3. Modality

We define four modality categories and these are *actual event*, *intended event*, *hypothetical event*, and *modalized event*.

Actual Event The text span describes an event or state in the real world that actually happened, happens, is happening or will happen. This includes habitual events that happen repeatedly.

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

Intended Event The text span describes an event or state that does not necessarily happen or hold in the real world. This also covers events that are intended, expected, planned, etc.

(15) It *has also drafted* three documents for attracting foreign capital, strengthening horizontal economic integration and allowing more authority for foreign operations .

Hypothetical Event The text span describes an event or state that is in a conditional (e.g., if, when) clause or takes place conditional on something else, and does not necessarily happen in reality.

(16) Would the experiment have been as successful *if* they had not spent the money ?

Modalized Event The text span follows a modal verb, and describes a possible or necessary event or state, or an ability.

(17) The recent confrontation *could* ignite regional convulsions as Turkey is sucked into Syria, leading to belated actions from the international community.

4. Experiments

We did a series of annotation experiments by using these guidelines to annotate data from 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. We had three annotators (all English native speakers. One speaks some Chinese and the other two do not speak Chinese et al) and after three rounds of training, the average pair-wise agreement consistently stays above 80% and the average *Kappa* score consistently exceeds 70% (Table 1), indicating reliable annotation. Even though it is hard to meaningfully compare agreement statistics, it is still worth noting these numbers are consistently

Round	Instances	Tense	Event	Modality	Average Agreement	Average Kappa
1	167	78.6	76.4	81.4	77.8	73.0
2	102	70.3	74.5	79.4	74.7	72.1
3	92	70.6	68.1	77.5	72.1	60.2
4	200	87.2	79.8	93.3	86.8	80.2
5	154	85.3	82.5	92.6	86.8	81.9
6	209	82.8	79.6	88.7	83.7	75.7
7	186	79.9	79.2	86.9	82.0	70.5

Table 1: Annotation agreement statistics during training sessions (%)

higher than what was reported in (Xue et al., 2008), where they annotate semantic tense directly on Chinese text.

Digging a bit deeper into these statistics, we found that agreement varies on different data domains. Results in Table 2 show that newswire data received the highest average pair-wise agreement scores while the weblog data was the hardest to annotate. These agreements are the average over three rounds of annotation by four annotators who have participated in this project so far.

We have completed the annotation of all weblog and newswire documents in the parallel corpus and we are currently annotating the broadcast data. The amount of data we have annotated so far are listed in Table 3.

Dataset	Sents	Anotated text spans	Words
Weblog	3,699	14,444	86,847
Newswire	2,079	8,560	75,755
Broadcast	511	1,523	10,112
Total	6,289	24,527	172,714

Table 3: Annotated data size, using word and sentence count on the English side.

To test the effect of event type and modality on inferring tense, we trained a CRF model [2] using gold standard event type and modality as features, in addition to time expressions, verb classes and tense information for time expressions from the PKU dictionary [3], as well as features used in (Xue, 2008). Other than event type and modality, all other features are automatically extracted. The results are presented in Table 4 and they show that taking out event type and modality features results in a substantial drop in accuracy. Leaving out eventuality type results in a 12.5% loss while taking out modality results in a 6.2% loss. Taking out both leads to a drop of 22.7%. This means that event type and modality are critically important to tense prediction. In a realistic scenario, of course, event type and modality have to be inferred themselves. This will be our future work.

Features	Accuracy
All	0.796
- gold eventualtiy type	0.671
- gold modality	0.734
- gold eventualtiy and gold modality	0.569

Table 4: Tense prediction experiment results on Chinese.

5. Related Work

The TimeBank (Pustejovsky et al., 2005) also annotates tense, aspect, and modality, as attributes of events. But it is focused on grammatical tense of English verbs instead of semantic tense of both English and Chinese events. However, our approach regards semantic tense as an underlying truth across languages and aims at the semantic tense of events on Chinese. The TempEval evaluations (Verhagen et al., 2007; Verhagen et al., 2010) are aimed at detecting time expressions, events, and the relations among them. They target only on the main event of a sentence and use temporal relations such as “before”, “after”, or “overlap” to represent “abstract tense” between two events or between one event and the document creation time. On the contrary, we process every possible event and state including those in non-finite verb forms or even in nominal forms, and extract their “semantic tense” information with respect to the document creation time uniformly. Moreover, our annotation also provides event type and modality information that are proved to be useful for semantic tense inference and may also be helpful for other natural language processing tasks.

Reichart and Rappoport (2010) introduced a more general Tense Sense Disambiguation (TSD) task to annotate and disambiguate the semantic tense for English. They provide a fine-grained sense taxonomy for tense which includes underlying senses such as ”things that are always true”, ”general and repeated actions and habits”, ”plans, expectations and hopes”. These are similar distinctions of different event type and modality categories, only in many unstructured fine-grained senses. We include similar distinctions in our annotation, but in a more structured manner with three dimensions.

There have been two general approaches on annotating and automatically inferring tense on Chinese data. One approach is to annotate tense for Chinese verbs directly (Ye et al., 2006; Ye, 2007; Xue et al., 2008; Xue, 2008). The issue with direct annotation on Chinese data is that main-

[2] We used the CRF++ package that can be found here: crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar

[3] A semantic dictionary on Chinese built by Peking University: http://www.icl.pku.edu.cn/icl_groups/syntac-dictn/specification.htm

Dataset	Tense	Event	Modality	Average Agreement	Average Kappa
Weblog	80.4	76.7	85.9	81.0	72.7
Newswire	87.6	83.9	95.3	89.0	84.9
Broadcast	87.1	84.1	87.0	86.1	77.7

Table 2: Annotation agreement statistics on different datasets (%)

taining the inter-annotator consistency is proved to be very chanllenging, given the total lack of explicit surface cues on Chinese. (Xue et al., 2008) reported an inter-annotator agreement of 75%, a result that is comparable to that of our first round of annotation. The second approach is to map grammatical tense in English onto Chinese via word aligned parallel data (Liu et al., 2011). However, the syntactic forms of tense change over languages even with same underlying semantics. And this leads to inconsistent tense information on the target language.

6. Conclusion and Future Work

We describe a distant annotation approach for annotating the tense, event type and modality of events in Chinese text by annotating their English counterpart via a word-aligned parallel corpus. Annotation agreements indicate that this approach shows promise as an effective alternative to annotating the Chinese text directly. Preliminary results also show that gold event type and modality are powerful indicators of semantic tenses. Our next step is to complete the annotation and develop models to predict tense fully automatically.

Acknowledgements

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.

7. References

Li, X., Strassel, S., Grimes, S., Ismael, S., Maamouri, M., Bies, A., and Xue, N. (2012). Parallel Aligned Treebanks at LDC: New Challenges Interfacing Existing Infrastructures. In *Proceedings of LREC-2012*, Istanbul, Turkey.

Liu, F., Liu, F., and Liu, Y. (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.

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

Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., and Mani, I. (2005). The specification language TimeML. In Mani, I., Pustejovsky, J., and Gaizauskas, R., editors, *The Language of Time: a Reader*. Oxford University Press.

Reichart, R. and Rappoport, A. (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.

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

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic, June. Association for Computational Linguistics.

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.

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

Xue, N., Hua, Z., and Chen, K.-Y. (2008). Automatic inference of the temporal location of situations in chinese text.

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

Ye, Y., Fossum, V. L., and Abney, S. (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.

Ye, Y. (2007). *Automatica Tense and Aspect Translation between Chinese and English*. Ph.D. thesis, University of Michigan.