Automatic inference of the temporal location of situations in Chinese text

Nianwen Xue
Center for Computational Language and Education Research
University of Colorado at Boulder
Colorado, U.S.A.
Nianwen.Xue@colorado.edu

Abstract

Chinese is a language that does not have morphological tense markers that provide explicit grammaticalization of the temporal location of situations (events or states). However, in many NLP applications such as Machine Translation, Information Extraction and Question Answering, it is desirable to make the temporal location of the situations explicit. We describe a machine learning framework where different sources of information can be combined to predict the temporal location of situations in Chinese text. Our experiments show that this approach significantly outperforms the most frequent tense baseline. More importantly, the high training accuracy shows promise that this challenging problem is solvable to a level where it can be used in practical NLP applications with more training data, better modeling techniques and more informative and generalizable features.

1 Introduction

In a language like English, tense is an explicit (and maybe imperfect) grammaticalization of the temporal location of situations, and such temporal location is either directly or indirectly defined in relation to the moment of speech. Chinese does not have grammaticalized tense in the sense that Chinese verbs are not morphologically marked for tense. This is not to say, however, that Chinese speakers do not attempt to convey the temporal location of situations when they speak or write, or that they cannot interpret the temporal location when they read Chinese text, or even that they have a different way of representing the temporal location of situations. In fact, there is evidence that the temporal location is represented in Chinese in exactly the same way as it is represented in English and most world languages: in relation to the moment of speech. One piece of evidence to support this claim is that Chinese temporal expressions like 今天 ("today"), 明天 ("tomorrow") and 昨天 ("yesterday") all assume a temporal deixis that is the moment of speech in relation to which all temporal locations are defined. Such temporal expressions, where they are present, give us a clear indication of the temporal location of the situations they are associated with. However, not all Chinese sentences have such temporal expressions associated with them. In fact, they occur only infrequently in Chinese text. It is thus theoretically interesting to ask, in the absence of grammatical tense and explicit temporal expressions, how do readers of a particular piece of text interpret the temporal location of situations?

There are a few linguistic devices in Chinese that provide obvious clues to the temporal location of situations, and one such linguistic device is aspect markers. Although Chinese does not have grammatical tense, it does have grammaticalized aspect in the form of aspect markers. These aspect markers often give some indication of the temporal location of an event. For example, Chinese has the perfective aspect marker 了 and 过, and they are often associated with the past. Progressive aspect marker 着, on the other hand, is often associated with the present. In addition to aspect, certain adverbs also provide clues to the temporal location of the situations they are as-

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 707–714,
Honolulu, October 2008.

sociated with. For example, 已 or 已经 (”already”), often indicates that the situation they are associated with has already occurred and is thus in the past. 在, another adverbial modifier, often indicates that the situation it modifies is in the present. However, such linguistic associations are imperfect, and they can only be viewed as tendencies rather than rules that one can use to deterministically infer the temporal location of a situation. For example, while 已 indeed indicates that the situation described in (1) is in the past, when it modifies a stative verb as it does in (1b), the situation is still in the present.

(1) a. 他 [已] 做完 该 项目 。
he already finish this project .

”He already finished the project.”

b. 中国 [已] 拥有 产生 世界级
China already has produce world-class
软件 的 基础 。
software DE foundation .

”China already has the foundation to produce world-class software.”

More importantly, only a small proportion of verb instances in any given text have such explicit temporal indicators and therefore they cannot be the whole story in the temporal interpretation of Chinese text. It is thus theoretically interesting to go beyond the obvious and investigate what additional information is relevant in determining the temporal location of a situation in Chinese.

Being able to infer the temporal location of a situation has many practical applications as well. For example, this information would be highly valuable to Machine Translation. To translate a language like Chinese into a language like English in which tense is grammatically marked with inflectional morphemes, an MT system will have to infer the necessary temporal information to determine the correct tense for verbs. Statistical MT systems, the currently dominant research paradigm, typically do not address this issue directly. As a result, when evaluated for tense, current MT systems often perform miserably. For example, when a simple sentence like “他/he 明天/tomorrow 返回/return 上海/Shanghai” is given to Google’s state-of-the-art Machine Translation system[1], it produces the output “He returned to Shanghai tomorrow”, instead of the correct “he will return to Shanghai tomorrow”. The past tense on the verb “returned” contradicts the temporal expression “tomorrow”. Determining the temporal location is also important for an Information Extraction task that extracts events so that the extracted events are put in a temporal context. Similarly, for Question Answering tasks, it is also important to know whether a situation has already happened or it is going to happen, for example.

In this paper, we are interested in investigating the kind of information that is relevant in inferring the temporal location of situations in Chinese text. We approach this problem by manually annotating each verb in a Chinese document with a “tense” tag that indicates the temporal location of the verb[2]. We then formulate the tense determination problem as a classification task where standard machine learning techniques can be applied. Figuring out what linguistic information contributes to the determination of the temporal location of a situation becomes a feature engineering problem of selecting features that help with the automatic classification. In Section 2, we present a linguistic annotation framework that annotates the temporal location of situations in Chinese text. In Section 3 we describe our setup for an automatic tense classification experiment and present our experimental results. In Section 4 we focus in on the features we have used in our experiment and attempt to provide a quantitative as well as intuitive explanation of the contribution of the individual features and speculate on what additional features could be useful. In Section 5 we discuss related work and Section 6 concludes the paper and discusses future work.

2 Annotation framework

It is impossible to define the temporal location without a reference point, a temporal deixis. As we have shown in Section 1, there is convincing evidence from the temporal adverbials like 昨天(“yesterday”), 今天(“today”) and 明天

[1] http://www.google.com/language_tools

[2] For simplicity, we use the term “tense” exchangeably with the temporal location of an event or situation, even though tense usually means grammatical tense while temporal location is a more abstract semantic notion.

（“tomorrow”) that Chinese, like most if not all languages of the world, use the moment of speech as this reference point. In written text, which is the primary source of data that we are dealing with, the temporal deixis is the document creation time. All situations are temporally related to this document creation time except in direct quotations, where the temporal location is relative to the moment of speech of the speaker who is quoted.

In addition to the moment of speech or document creation time in the case of written text, Reference Time and Situation Time are generally accepted as important to determining the temporal location since Reichenbach (1947) first proposed them. Situation Time is the time that a situation actually occurs while Reference time is the temporal perspective from which the speaker invites his audience to consider the situation. Reference Time does not necessarily overlap with Situation Time, as in the case of present perfective tense, where the situation happened in the past but the reader is invited to look at it from the present moment and focus on the state of completion of the situation. Reference Time is in our judgment too subtle to be annotated consistently and thus in our annotation scheme we only consider the relation between Situation Time and the document creation time when defining the temporal location of situations. Another key decision we made when formulating our annotation scheme is to define an abstract “tense” that do not necessarily model the actual tense system in any particular language that has grammatical tense. In a given language, the grammatical tense reflected in the morphological system may not have a one-to-one mapping between the grammatical tense and the temporal location of a situation. For example, in an English sentence like “He will call me after he gets here”, while his “getting here” happens at a time in the future, it is assigned the present tense because it is in a clause introduced by “after”. It makes more sense to ask the annotator, who is necessarily a native speaker of Chinese, to make a judgment of the temporal location of the situation defined in terms of the relation between the Situation Time and the moment of speech rather than by such language-specific idiosyncracies of another language.

Temporal locations that can be defined in terms of the relation between Situation Time and the moment of speech are considered to be *absolute tense*. In some cases, the temporal location of a situation cannot be directly defined in relation to the moment of speech. For example in (2), the temporal location of 有意 (“intend”) cannot be determined independently of that of 透露(“reveal”). The temporal location of 有意 is simultaneous with 透露. If the temporal location of 透露 is in the past, then the temporal location of 有意 is also in the past. If the temporal location of 透露 is in the future, then the temporal location of 有意 is also in the future. In this specific case, the situation denoted by the matrix verb 透露 is in the past. Therefore the situation denoted by 有意 is also located in the past.

(2) 他还 透露 俄罗斯有意 在今后十年
he also reveal Russia intend in next ten years
内 ,向伊朗提供 武器 .
within , to Iran provide weapons .

“He also revealed that Russia intended to provide weapons to Iran within the next ten years.”

Therefore in our Chinese “tense” annotation task, we annotate both *absolute* and *relative* tenses. We define three absolute tenses based on whether the situation time is anterior to (in the past), simultaneous with (in the present), or posterior to (in the future) document creation time. In addition to the absolute tenses, we also define one relative tense, future-in-past, which happens when a future situation is embedded in a past context. We do not assign a tense tag to modal verbs or verb particles. The set of tense tags are described in more detail below:

2.1 Present tense

A situation is assigned the present tense if it is true at an interval of time that includes the present moment. The present tense is compatible with states and activities. When non-stative situations are temporally located in the present, they either have an imperfective aspect or have a habitual or frequentive reading which makes them look like states, e.g.,

(3) 他常常 参加 户外 活动 。
he often attend outdoors activities .

“He often attends outdoors activities.”

2.2 Past tense

Situations that happen before the moment of speech (or the document creation time) are temporally located in the past as in (4):

(4) 中方 人员 及 侨胞
Chinese personnel and Chinese nationals
安全 撤离 乍得 。
safely withdraw from Chad .

"Chinese personnel and Chinese nationals safely withdrew from Chad."

2.3 Future tense

Situations that happen posterior to the moment of speech are temporally located in the future. Future situations are not simply the opposite of past situations. While past situations have already happened by definition, future situations by nature are characterized by uncertainty. That is, future situations may or may not happen. Therefore, future situations are often linked to possibilities, not just to situations that will definitely happen. A example of future tense is given in (5):

(5) 大会 明年 在新加坡 举行 。
conference next year in Singapore hold .

"The conference will be held in Singapore next year."

2.4 Future-in-past

The temporal interpretation of one situation is often bound by the temporal location of another situation. One common scenario in which this kind of dependency occurs is when the target situation, the situation we are interested in at the moment, is embedded in a reference situation as its complement. Just as the absolute "tense" represents a temporal relation between the situation time and the moment of speech or document creation time, the relative "tense" represents a relation between the temporal location of a situation and its reference situation. Although theoretically the target situation can be anterior to, simultaneous with, or posterior to the reference situation, we only have a special tense label when the target situation is posterior to the reference situation and the reference situation is located in the past. In this case the label for the target situation is future-in-past as illusrated in (6):

(6) 公司 员工 透露 《星际2》测试
company personnel reveal " Star 2 " trial
版 即将面世 。
version soon face the world .

"The company personnel revealed that 'Star 2' trial version would soon face the world."

2.5 No tense label

Modals and verb particles do not receive a tense label:

(7) 科索沃 独立 可能 引发 骚乱，
Kosovo independence may cause riot .
联合国人员 已 准备 撤离 。
UN personnel already prepare withdraw .

"Kosovo independence may cause riot. UN personnel have already prepared to leave."

The "situations" that we are interested in are expressed as clauses centered around a verb, and for the sake of convenience we mark the "tense" on the verb itself instead of the entire clause. However, when inferring the temporal location of a situation, we have to take into consideration the entire clause, because the arguments and modifiers of a verb are just as important as the verb itself when determining the temporal location of the situation. The annotation is performed on data selected from the Chinese Treebank (Xue et al., 2005), and more detailed descriptions and justifications for the annotation scheme is described in (Xue et al., 2008). Data selection is important for tense annotation because, unlike POS-tagging and syntactic annotation, which applies equally well to different genres of text, temporal annotation in more relevant in some genres than others. The data selection task is made easier by the fact that the Chinese Treebank is already annotated with POS tags and Penn Treebank-style syntactic structures. Therefore we were able to just select articles based on how many constituents in the article are annotated with the temporal function tag -TMP. We have annotated 42 articles in total, and all verbs in an article are assigned one of the five tags described above: present, past, future, future-in-past, and none.

3 Experimental results

The tense determination task is then a simple five-way classification task. Theoretically any standard machine learning algorithm can be applied to the task. For our purposes we used the Maximum Entropy algorithm implemented as part of the Mallet machine learning package (McCallum, 2002) for its competitive training time and performance tradeoff. There might be algorithms that could achieve higher classification accuracy, but our goal in this paper is not to pursue the absolute high performance. Rather, our purpose is to investigate what information when used as features is relevant to determining the temporal location of a situation in Chinese, so that these features can be used to design high performance practical systems in the future.

The annotation of 42 articles yielded 5709 verb instances, each of which is annotated with one of the five tense tags. For our automatic classification experiments, we randomly divided the data into a training set and a test set based on a 3-to-1 ratio, so that the training data has 4,250 instances while the test set has 1459 instances. As expected, the past tense is the most frequent tense in both the training and test data, although they vary quite a bit in the proportions of verbs that are labeled with the past tense. In the training data, 2145, or 50.5% of the verb instances are labeled with the past tense while in the test data, 911 or 62.4% of the verb instances are labeled with the past tense. The 62.4% can thus be used as a baseline when evaluating the automatic classification accuracy. This is a very high baseline given that the much smaller proportion of verbs that are assigned the past tense in the training data.

Instead of raw text, the input to the classification algorithm is parsed sentences from the Chinese Treebank that has the syntactic structure information as well as the part-of-speech tags. As we will show in the next section, information extracted from the parse tree as well as the part-of-speech tags prove to be very important in determining the temporal location of a situation. The reason for using "correct" parse trees in the Chinese Treebank is to factor out noises that are inevitable in the output of an automatic parser and evaluate the contribution of syntactic information in the "ideal" scenario. In a realistic setting, one of course has to use an automatic parser.

The results are presented in Table 1. The overall accuracy is 67.1%, exceeding the baseline of choosing the most frequent tense in the test, which is 62.4%. It is worth noting that the training accuracy is fairly high, 93%, and there is a steep drop-off from the training accuracy to the test accuracy although this is hardly unexpected given the relatively small training set. The high training accuracy nevertheless attests the relevance of the features we have chosen for the classification, which we will look at in greater detail in the next section.

tense	precision	recall	f-score
present	0.51	0.62	0.56
past	0.75	0.81	0.78
future	0.33	0.45	0.38
future-in-past	0.76	0.18	0.30
none	0.86	0.83	0.84
overall	0.93 (train), 0.671 (test)		

Table 1: Experimental results

4 What information is useful?

Our classification algorithm scans the verbs in a sentence one at a time, from left to right. Features are extracted from the context of the verb in the parse tree as well as from previous verbs the tense of which have already been examined. We view features for the classification algorithm as information that contributes to the determination of the temporal location of situations in the absence of morphological markers of tense. The features we used for the classification task can all be extracted from a parse tree and the POS information of a word. They are described below:

- Verb Itself: The character string of the verbs, e.g., 拥有("own"), 是("be"), etc.
- Verb POS: The part-of-speech tag of the verb, as defined in the Chinese Treebank. There are three POS tags for verbs, *VE* for existential verbs such as 有("have, exist"), *VC* for copula verbs like 是("be"), *VA* for stative verbs like 高("tall"), and *VV* for all other verbs.
- Position of verb in compound: If the target verb is part of a verb compound, the position

of the compound is used as a feature in combination with the compound type. The possible values for the position are *first* and *last*, and the compound type is one of the six defined in the Chinese Treebank: *VSB*, *VCD*, *VRD*, *VCP*, *VNV*, and *VPT*. An example feature might be "last+VRD".

- Governing verb and its tense: Chinese is an SVO language, and the governing verb, if there is one, is on the left and is higher up in the tree. Since we are scanning verbs in a sentence from left to right, the tense for the governing verb is available at the time we look at the target verb. So we are using the character string of the governing verb as well as its tense as features. In cases where there are multiple levels of embedding and multiple governing verbs, we select the closest governing verb.

- Left ADV: Adverbial modifiers of the target verb are generally on the left side of the verb, therefore we are only extracting adverbs on the left. We first locate the adverbial phrases and then find the head of the adverbial phrase and use character string of the head as feature.

- Left NT: NT is a POS in the Chinese Treebank for nominal expressions that are used as temporal modifiers of a verb. The procedure for extracting the NT modifers is similar to the procedure for finding adverbial modifiers, the only difference being that we are looking for NPs headed by nouns POS-tagged NT.

- Left PP: Like adverbial modifiers, PP modifiers are also generally left modifiers of a verb. If there is a PP modifier, the character string of the head preposition combined with the character string of the head noun of its NP complement is used as a feature, e.g., "在+期间" ("at+period").

- Left LC: Left localizer phrases. Localizers phrases are also called post-positions by some and they function similarly as left PP modifiers. If the target verb has a left localizer phrase modifier and the character string of its head is used as a feature, e.g., 以来("since").

- Left NN: This feature is intended to capture the head of the subject NP. The character string of the head of the NP is used as a feature.

- Aspect marker. Aspect markers are grammaticalizations of aspect and they immediately follow the verb. If the target verb is associated with an aspect marker, the character string of that aspect marker is used as a feature, e.g., "了".

- DER: DER is the POS tag for 得, a character which introduces a resultative construction when following a verb. When it occurs together with the target verb, it is used as a feature.

- Quantifier in object: When there is a quantifier in the NP object for the target verb, its character string is used as a feature.

- Quotation marks: Finally the quotation marks are used as a feature when they are used to quote the clause that contains the target verb.

We performed an ablation evaluation of the features to see how effective each feature type is. Basically, we took out each feature type, retrained the classifier and reran the classifier on the test data. The accuracy without each of the feature types are presented in Table 2. The features are ranked from the most effective to the least effective. Features that lead to the most drop-off when they are taken out of the classification algorithm are considered to be the most effective. As shown in Table 2, the most effective features are the governing verb and its tense, while the least effective features are the quantifiers in the object. Most of the features are lexicalized in that the character strings of words are used as features. When lexicalized features are used, features that appear in the training data do not necessarily appear in the test data and vice versa. This provides a partial explanation of the large discrepancy between the training and test accuracy. In order to reduce this discrepancy, one would have to use a larger training set, or make the features more generalized. Some of these features can in fact be generalized or normalized. For example, a temporal modifier such as the date "1987" can be reduced to something like "before the document creation time", and this is something that we will experiment with in

our future work. The training set used here is sufficient to show the efficacy of the features, but to improve the tense classification to a satisfactory level of accuracy, more training data need to be annotated.

Feature	accuracy (w/o)
Governing verb/tense	0.620
verb itself	0.635
Verb POS	0.656
Position verb in compound	0.656
left ADV	0.657
left NT	0.657
Quotation mark	0.657
left PP	0.663
left LC	0.664
Right DER	0.665
Aspect marker	0.665
left NN	0.665
Quantifier in object	0.669
overall	0.671 (test)

Table 2: Feature Performance

Features like adverbial, prepositional, localizer phrase modifiers and temporal noun modifiers provide explicit temporal information that is relevant in determining the temporal location. The role of the governing verb in determining the temporal location of a situation is also easy to understand. As we have shown in Section 2, when the target verb occurs in an embedded clause, its temporal location is necessarily affected by the temporal location of the governing verb of this embedded clause because the temporal location of the former is often defined in relation to that of the latter. Not surprisingly, the governing verb proves to be the most effective feature. Quotation marks in written text change the temporal deixis from the document creation time to the moment of speech of the quoted speaker, and the temporal location in quoted speech does not follow the same patterns as target verbs in embedded clauses. Aspect markers are tied closely to tense, even though the contributions they made are small due to their rare occurrences in text.

The relevance of other features are less obvious. The target verb itself and its POS made the most contribution other than the governing verb. It is important to understand why they are effective or useful at all. In a theoretic work on the temporal interpretation of verbs in languages like Chinese which lacks tense morphology, Smith and Erbaugh (2005) pointed out that there is a default interpretation for bounded and unbounded situations. Specifically, bounded situations are temporally located in the past by default while unbounded situations are located in the future. The default interpretation, by definition, can be overwritten when there is explicit evidence to the contrary. Recast in statistical terms, this means that bounded events have a tendency to be located in the past while unbounded events have a tendency to be located in the present, and this tendency can be quantified in a machine-learning framework. Boundedness has many surface manifestations that can be directly observed, and one of them is whether the verb is stative or dynamic. The target verb itself and its POS tag represents this information. Resultatives in the form of resultative verb compound and the DER construction, quantifiers in the object are other surface reflections of the abstract notion of boundedness. The fact that these features have contributed to the determination of the temporal location of situations to certain extent lends support to Smith's theoretical claim.

5 Related work

Inferring the temporal location is a difficult problem that is not yet very well understood. It has not been studied extensively in the context of Natural Language Processing. Olson et al (2000; 2001) realized the importance of using the aspectual information (both grammatical and lexical aspect) to infer tense in the context of a Chinese-English Machine Translation system. They encoded the aspectual information such as telicity as part of the Lexical Conceptual Structure and use it to heuristically infer tense when generating the English output. This rule-based approach is not very suited for modeling the temporal location information in Chinese. As they themselves noted, aspectual information can only be used as a tendency rather than a deterministic rule. We believe this problem can be better modeled in a machine learning framework where different sources of information, each one being imperfect, can be combined based on their effectiveness to provide a more reasonable overall prediction.

Ye (2007) did approach this problem with machine learning techniques. She used Chinese-English parallel data to manually map the tense information from English to Chinese and trained a Conditional Random Field classifier to make predictions about tense. She used only a limited number of surface cues such as temporal adverbials and aspect markers as features and did not attempt to model the lexical aspect information such as boundedness, which we believe would have helped her system performance. Her data appeared to have a much larger percentage of verb instances that have the past tense and thus her results are mostly incomparable with that of ours.

6 Conclusion and future work

We have defined the automatic inference of the temporal location of situations in Chinese text as a machine learning problem and demonstrated that a lot more information in the form of features contributes to the solution of this challenging problem than previously realized. The accuracy on the held-out test is a significant improvement over the baseline, the proportion of verbs assigned the most frequent tense (the past tense). Although there is a large drop-off from the training accuracy to the test accuracy due to the lexical nature of the features, the high training accuracy does show promise that this challenging problem is solvable with a larger training set, better modeling techniques and more refined features. In the future we will attempt to solve this problem along these lines and work toward a system that can be used in practical applications.

Acknowledgments

We would like to thank Hua Zhong and Kaiyun Chen for their efforts to annotate the data used in our experiments. Without their help this work would of course be impossible.

References

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mari Olson, David Traum, Carol Vaness Dykema, Amy Weinberg, and Ron Dolan. 2000. Telicity as a cue to temporal and discourse structure in Chinese-English Machine Translation. In *Proceedings of NAACL-ANLP 2000 Workshop on Applied interlinguas: practical applications of interlingual approaches to NLP*, pages 34–41, Seattle Washington.

Mari Olson, David Traum, Carol Vaness Dykema, and Amy Weinberg. 2001. Implicit cues for explicit generation: using telicity as a cue for tense structure in a Chinese to English MT system. In *Proceedings of Machine Translation Summit VIII*, Santiago de Compostela, Spain.

Hans Reichenbach. 1947. *Elements of Symbolic Logic*. The MacMillan Company, New York.

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

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, Zhong Hua, and Kai-Yun Chen. 2008. Annotating "tense" in a tenseless language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

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