Combining Linguistic Features with Weighted Bayesian Classifier for Temporal Reference Processing

Guihong Cao
Department of Computing
The Hong Kong Polytechnic University, Hong Kong
csghcao@comp.polyu.edu.hk

Kam-Fai Wong
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong
kfwong@se.cuhk.edu.hk

Wenjie Li
Department of Computing
The Hong Kong Polytechnic University, Hong Kong
cswjli@comp.polyu.edu.hk

Chunfa Yuan
Department of Computer Science and Technology
Tsinghua University, Beijing, China.
cfyuan@tsinghua.edu.cn

Abstract

Temporal reference is an issue of determining how events relate to one another. Determining temporal relations relies on the combination of the information, which is explicit or implicit in a language. This paper reports a computational model for determining temporal relations in Chinese. The model takes into account the effects of linguistic features, such as tense/aspect, temporal connectives, and discourse structures, and makes use of the fact that events are represented in different temporal structures. A machine learning approach, Weighted Bayesian Classifier, is developed to map their combined effects to the corresponding relations. An empirical study is conducted to investigate different combination methods, including lexical-based, grammatical-based, and role-based methods. When used in combination, the weights of the features may not be equal. Incorporating with an optimization algorithm, the weights are fine tuned and the improvement is remarkable.

1 Introduction

Temporal information describes changes and time of the changes. In a language, the time of an event may be specified explicitly, for example "他们在1997年解决了该市的交通问题（*They solved the traffic problem of the city in* 1997)"; or it may be related to the time of another event, for example "修成立交桥以后, 他们解决了该市的交通问题 (*They solved the traffic problem of the city after the street bridge had been built*". Temporal reference describes how events relate to one another, which is essential to natural language processing (NLP). Its major applications cover syntactic structural disambiguation (Brent, 1990), information extraction and question answering (Li, 2002), language generation and machine translation (Dorr, 2002).

Many researchers have attempted to characterize the nature of temporal reference in a discourse. Identifying temporal relations[1] between two events depends on a combination of information resources. This information is provided by explicit tense and aspect markers, implicit event classes or discourse structures. It has been used to explain semantics of temporal expressions (Moens, 1988; Webber, 1988), to constrain possible temporal interpretations (Hitzeman, 1995; Sing, 1997), or to generate appropriate temporally conjoined clauses (Dorr, 2002).

The purpose of our work is to develop a computational model, which automatically determines temporal relations in Chinese. While temporal reference interpretation in English has been well studied, Chinese has been rarely discussed. In our study, thirteen related features are identified from linguistic perspective. How to combine these features and how to map their combined effects to the corresponding relations are the critical issues to be addressed in this paper.

Previous work was limited in that they just constructed constraint or preference rules for some representative examples. These methods are ineffective for computing purpose, especially when a large number of the features are involved and the interaction among them is unclear. Therefore, a machine learning approach is applied and the empirical studies are carried out in our work.

The rest of this paper is organized as follows. Section 2 introduces temporal relation representations. Section 3 provides linguistic background of temporal reference and investigates linguistic features for determining temporal relations in Chinese. Section 4 explains the methods used to combine linguistic features with Bayesian Classifier. It is followed by a description of the optimization algorithm which is used for estimating feature weights in Section 5. Finally, Section 6 concludes the paper.

2 Representing Temporal Relations

With the growing interests to temporal information processing in NLP, a variety of temporal systems have been introduced to accommodate the characteristics of temporal information. In order to process temporal reference in a discourse, a formal represen-

[1] The relations under examined include both intra-sentence and inter-sentence relations.

tation of temporal relations is required. Among those who worked on representing or explaining temporal relations, some have taken the work of Reichenbach (Reichenbach, 1947) as a starting point, while others based their works on Allen's (Allen, 1983).

Reichenbach proposed a point-based temporal theory. Reichenbach's representation associated English tenses and aspects with three time points, namely event time (*E*), speech time (*S*) and reference time (*R*). The reference of E-R and R-S was either before (or after in reverse order) or simultaneous. This theory was later enhanced by Bruce who defined seven temporal relations (Bruce, 1972). Given two durative events, the interval relations between them were modeled by the order between the greatest lower bounding point and least upper bounding point of the two events. In the other camp, instead of adopting time points, Allen took intervals as temporal primitives to facilitate temporal reasoning and introduced thirteen basic relations. In this interval-based representation, points were relegated to a subsidiary status as "meeting places" of intervals. An extension to Allen's theory, which treated both points and intervals as primitives on an equal footing, was later investigated by Knight and Ma (Knight, 1994).

In natural languages, events described can be either punctual or durative in nature. A punctual event, e.g., 爆炸 (*explore*), occurs instantaneously. It takes time but does not last in a sense that it lacks of a process of change. It is adequate to represent a punctual event with a simple point structure. Whilst, a durative event, e.g., 盖楼 (*built a house*), is more complex and its accomplishment as a whole involves a process spreading in time. Representing a durative event requires an interval representation. For this reason, Knight and Ma's model is adopted in our work (see Figure 1). Taking the sentence "修成立交桥以后，他们解决了该市的交通问题 (*They solved the traffic problem of the city after the street bridge had been built*)" as an example, the relation held between building the bridge (i.e., an interval) and solving the problem (i.e., a point) is BEFORE.

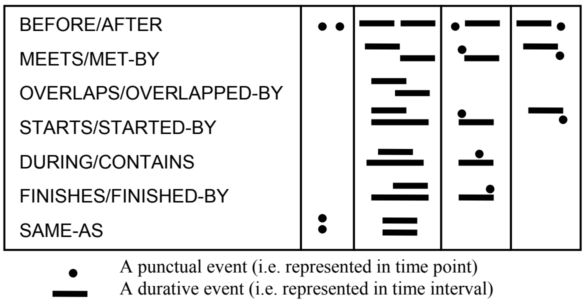

Figure 1 13 relations represented with points and intervals

3 Linguistic Background of Temporal Reference in a Discourse

3.1 Literature Review

There were a number of theories in the literature about how temporal relations between events can be determined in English. Most of the researches on temporal reference were based on Reichenbach's notion of tense/aspect structure, which was known as Basic Tense Structure (BTS). As for relating two events adjoined by a temporal/causal connective, Hornstein (Hornstein, 1990) proposed a neo-Reichenbach structure which organized the BTSs into a Complex Tense Structure (CTS). It has been argued that all sentences containing a matrix and an adjunct clause were subject to linguistic constraints on tense structure regardless of the lexical words included in the sentence. Generally, constraints were used to support syntactic disambiguation (Brent, 1990) or to generate acceptable sentences (Dorr, 2002).

In a given CTS, a past perfect clause should precede the event described by a simple past clause. However, the order of two events in CTS does not necessarily correspond to the order imposed by the interpretation of the connective (Dorr, 2002). Temporal/casual connective, such as "after", "before" or "because", can supply explicit information about the temporal ordering of events. Passonneau (Passonneau, 1988), Brent (Brent, 1990 and Sing (Sing, 1997) determined intra-sentential relations by accounting for temporal or causal connectives. Dorr and Gaasterland (Dorr, 2002), on the other hand, studied how to generate the sentences which reflect event temporal relations by selecting proper connecting words. However, temporal connectives can be ambiguous. For instance, a "when" clause permits many possible temporal relations.

Several researchers have developed the models that incorporated aspectual types (such as those distinct from states, processes and events) to interpret temporal relations between clauses connected with "when". Moens and Steedmen (Moens, 1988) developed a tripartite structure of events[2], and emphasized it was the notion of causation and consequence that played a central role in defining temporal relations of events. Webber (Webber, 1988) improved upon the above work by specifying rules for how events are related to one another in a discourse and Sing and Sing defined semantic constraints through which events can be related (Sing, 1997). The importance of aspectual information in retrieving proper aspects and connectives for sentence generation was also recognized by Dorr and Gaasterland (Dorr, 2002).

Some literature claimed that discourse structures suggested temporal relations. Lascarides and Asher (Lascarides, 1991) investigated various contextual effects on rhetorical relations (such as narration, elaboration, explanation, background and result). They corresponded each of the discourse relations to a kind of temporal relation. Later, Hitzeman (Hitzeman, 1995) described a method for analyzing temporal structure of a discourse by taking into account the effects of tense, aspect, temporal adverbials and rhe-

[2] The structure comprises a culmination, an associated preparatory process and a consequence state.

torical relations. A hierarchy of rhetorical and temporal relations was adopted so that they could mutually constrain each other.

To summarize, the interpretation of temporal relations draws on the combination of various information resources, including explicit tense/aspect and connectives (temporal or otherwise), temporal classes implicit in events, or rhetorical relations hidden in a discourse. This conclusion, although drawn from the studies of English, provides the common understanding on what information is required for determining temporal relations across languages.

3.2 Linguistic Features for Determining Temporal Relations in Chinese

Thirteen related linguistic features are recognized for determining Chinese temporal relations in this paper (See Table 1). The selected features are scattered in various grammatical categories due to the unique nature of language, but they fall into the following three groups.

(1) Tense/aspect in English is manifested by verb inflections. But such morphological variations are inapplicable to Chinese verbs. Instead, they are conveyed lexically. In other words, tense and aspect in Chinese are expressed using a combination of, for example, time words, auxiliaries, temporal position words, adverbs and prepositions, and particular verbs. They are known as Tense/Aspect Markers.

(2) Temporal Connectives in English primarily involve conjunctions, such as "after" and "before", which are the key components in discourse structures. In Chinese, however, conjunctions, conjunctive adverbs, prepositions and position words, or their combinations are required to represent connectives. A few verbs that express cause/effect imply a temporal relation. They are also regarded as a feature relating to discourse structure[3]. The words which contribute to the tense/aspect and temporal connective expressions are explicit in a sentence and generally known as Temporal Indicators.

(3) Event Classes are implicit in a sentence. Events can be classified according to their inherent temporal characteristics, such as the degree of telicity and atomicity. The four widespread accepted temporal classes are state, process, punctual event and developing event (Li, 2002). Based on their classes, events interact with the tense/aspect of verbs to determine the temporal relations between two events.

Temporal indicators and event classes are both referred to as Linguistic Features. Table 1 shows the association between a temporal indicator and its effects. Note that the association is not one-to-one. For example, adverbs affect tense/aspect (e.g. 正, *being*) as well as discourse structure (e.g. 边, *at the same time*). For another example, tense/aspect can be jointly affected by auxiliary words (e.g. 过, *were/was*), trend verbs (起来, *begin to*), and so on. Obviously, it is not a simple task to map the combined effects of the thirteen linguistic features to the corresponding relations. Therefore, a machine learning approach is proposed, which investigates how these features contribute to the task and how they should be combined.

4 Combining Linguistic Features with Machine Learning Approach

Previous efforts in corpus-based NLP have incorporated machine learning methods to coordinate multiple linguistic features, for example, in accent restoration (Yarowsky, 1994) and event classification (Siegel, 1998).

Temporal relation determination can be modeled as a relation classification task. We formulate the thirteen temporal relations (see Figure 1) as the classes to be decided by a classifier. The classification process is to assign an event pair to one class according to their linguistic features. There existed numerous classification algorithms based upon supervised learning principle. One of the most effective classifiers is Bayesian Classifier, introduced by Duda and Hart (Duda, 1973) and analyzed in more detail by Langley and Thompson (Langley, 1992). Its predictive performance is competitive with state-of-the-

Linguistic Feature	Symbol	POS Tag	Effect	Example
With/Without punctuations	PT	Not Applicable	Not Applicable	Not Applicable
Speech verbs	VS	TI_vs	Tense	报告, 表示, 称
Trend verbs	TR	TI_tr	Aspect	起来, 下去
Preposition words	P	TI_p	Discourse Structure/Aspect	当, 到, 继
Position words	PS	TI_f	Discourse Structure	底, 后, 开始
Verbs with verb objects	VV	TI_vv	Tense/Aspect	继续, 进行, 续
Verbs expressing wish/hope	VA	TI_va	Tense	必须, 会, 可
Verbs related to causality	VC	TI_vc	Discourse Structure	导致, 致使, 引起
Conjunctive words	C	TI_c	Discourse Structure	并, 并且, 不过
Auxiliary words	U	TI_u	Aspect	着, 了, 过
Time words	T	TI_t	Tense	过去, 今后, 今年
Adverbs	D	TI_d	Tense/Aspect/Discourse Structure	便, 并, 并未, 不
Event class	EC	E0/E1/E2/E3	Event Classification	State, Punctual Event, Developing Event, Process

Table 1 Linguistic features: eleven temporal indicators and one event class

[3] The casual conjunctions such as "*because*" are included in this group.

art classifiers, such as C4.5 and SVM (Friedman, 1997).

4.1 Bayesian Classifier

Given the class c, Bayesian Classifier learns from training data the conditional probability of each attribute. Classification is performed by applying Bayes rule to compute the posterior probability of c given a particular instance x, and then predicting the class with the highest posterior probability ratio. Let $x=[e_1,e_2,t_1,t_2,...,t_n]$, $e_1,e_2 \in E$ are the two event classes and $t_1,t_2,...,t_n \in T$ are the temporal indicators (i.e. the words). E is the set of event classes. T is the set of temporal indicators. Then x is classified as:

$$c^* = \arg\max_c \log\left(\frac{P(c\,|\,e_1,e_2,t_1,t_2,...,t_n)}{P(\bar{c}\,|\,e_1,e_2,t_1,t_2,...,t_n)}\right) \qquad \text{(E1)}$$

where $\bar{c}$ denotes the classes different from c. Assuming event classes are independent of temporal indicators given c, we have:

$$\begin{aligned}&\log\left(\frac{P(c\,|\,e_1,e_2,t_1,t_2,...,t_n)}{P(\bar{c}\,|\,e_1,e_2,t_1,t_2,...,t_n)}\right)\\ &=\log\left(\frac{P(e_1,e_2,t_1,t_2,...,t_n\,|\,c)P(c)}{P(e_1,e_2,t_1,t_2,...,t_n\,|\,\bar{c})P(\bar{c})}\right) \qquad \text{(E2)}\\ &=\log\left(\frac{P(c)}{P(\bar{c})}\right)+\log\left(\frac{P(e_1,e_2\,|\,c)}{P(e_1,e_2\,|\,\bar{c})}\right)+\log\left(\frac{P(t_1,t_2,...,t_n\,|\,c)}{P(t_1,t_2,...,t_n\,|\,\bar{c})}\right)\end{aligned}$$

Assuming temporal indicators are independent of each other, we have

$$\frac{P(t_1,t_2,...,t_n\,|\,c)}{P(t_1,t_2,...,t_n\,|\,\bar{c})}=\prod_{i=1}^{n}\frac{P(t_i\,|\,c)}{P(t_i\,|\,\bar{c})}, \quad (i=1,2,...n) \qquad \text{(E3)}$$

A Naïve Bayesian Classifier assumes strict independence among all attributes. However, this assumption is not satisfactory in the context of temporal relation determination. For example, if the relation between e_1 and e_2 is SAME_AS, e_1 and e_2 have to be identical. We release the independence assumption for e_1 and e_2, and decompose the second part of (E2) as:

$$\frac{P(e_1,e_2\,|\,c)}{P(e_1,e_2\,|\,\bar{c})}=\frac{P(e_1\,|\,c)P(e_2\,|\,e_1,c)}{P(e_1\,|\,\bar{c})P(e_2\,|\,e_1,\bar{c})} \qquad \text{(E4)}$$

Estimation of $p(e_2\,|\,e_1,c)$ is motivated by Absolute Discounting N-Gram language model (Goodman, 2001):

$$P(e_2\,|\,e_1,c)=\begin{cases}\dfrac{C(e_2,e_1,c)-D}{C(e_1,c)} & \text{if } C(e_2,e_1,c)>0\\ \alpha(e_1,c)P(e_2|c) & \text{if } C(e_2,e_1,c)=0\end{cases} \qquad \text{(E5)}$$

here D is the discount factor and is set to 0.5 experimentally. From the fact that $\sum_{e_2} P(e_2\,|\,e_1,c)=1$, we get:

$$\alpha(e_1,c)=\frac{1-\sum_{e_2|C(e_2,e_1,c)>0}P(e_2\,|\,e_1,c)}{1-\sum_{e_2|C(e_2,e_1,c)>0}P(e_2\,|\,c)} \qquad \text{(E6)}$$

$p(t_i\,|\,c)$ and $p(e_i\,|\,c)$ are estimated by MLE with Dirichlet Smoothing method:

$$P(t_i\,|\,c)=\frac{C(t_i,c)+u}{\sum_{t_i\in T}C(t_i,c)+u\,|\,T\,|} \quad (i=1,2,...n) \qquad \text{(E7)}$$

$$P(e_i\,|\,c)=\frac{C(e_i,c)+u}{\sum_{e_i\in E}C(e_i,c)+u\,|\,E\,|} \quad (i=1,2) \qquad \text{(E8)}$$

where u (=0.5) is the smoothing factor. Then, $p(t_i\,|\,\bar{c})$, $p(e_i\,|\,\bar{c})$ and $P(e_2\,|\,e_1,\bar{c})$ can be estimated with (E5) - (E8) by substituting c with $\bar{c}$.

4.2 Estimating $P(t_1,t_2,...t_n\,|\,c)$ with Lexical-POS Information

The effects of a temporal indicator are constrained by its positions in a sentence. For instance, the conjunctive word 因为 (*because*) may represent the different relations when it occurs before or after the first event. Therefore, in estimating $p(t_1,t_2,...t_n\,|\,c)$, we consider an indicator located in three positions: (1) BEFORE the first event; (2) AFTER the first event and BEFORE the second and it modifies the first event; (3) the same as (2) but it modifies the second event; and (4) AFTER the second event. Note that cases (2) and (3) are ambiguous. The positions of the temporal indicators are the same. But it is uncertain whether these indicators modify the first or the second event if there is no punctuation (such as comma, period, exclamation or question mark) separating their roles. The ambiguity is resolved by using POS information. We assume that an indicator modifies the first event if it is an auxiliary word, a trend word or a position word; otherwise it modifies the second.

Thus, we rewrite $P(t_1,t_2,...t_n\,|\,c)$ as $P(t_{11},...,t_{1n_1}$, $t_{21},...,t_{2n_2},t_{31},...,t_{3n_3},t_{41},...t_{4n_4}\,|\,c)$, where n_j is the total number of the temporal indicators occurring in the position j. $j=1,2,3,4$ represents the four positions and $\sum_{j=1}^{4}n_j=n$. Assuming t_{ji} are independent of each other, then $\prod_{i=1}^{n}P(t_i\,|\,c)$ in (E3) is revised as $\prod_{j=1}^{4}\prod_{i=1}^{n_j}P(t_{ji}\,|\,c)$. Accordingly, (E7) is revised as:

$$P(t_{ji}\,|\,c)=\frac{C(t_{ji},c)+u}{\sum_{t_{ji}\in T}C(t_{ji},c)+u\,|\,T\,|} \qquad \text{(E7')}$$

$(j=1,2,3,4$ and $i=1,2,...n_j)$

In addition to taking positions into account, we further classify the temporal indicators into two groups according to their grammatical categories or semantic roles. The rationale of grouping will be demonstrated in Section 4.3.

4.3 Experimental Results

Several experiments have been designed to evaluate the proposed Bayesian Classifier in combining linguistic features for temporal relation determination and to reveal the impact of linguistic features on learning performance. 700 instances are extracted from Ta Kong Pao (a local Hong Kong Chinese newspaper) financial version. Among them, 500 are used as training data, and 200 as test data, which are partitioned equally into two sets. One is similar as training data in class distribution, while the other is quite different. 209 lexical words, gathered from linguistic books and corpus, are used as the temporal indicators and manually marked with the tags given in Table 1.

4.3.1 Impact of Individual Features

From linguistic perspective, the thirteen features (see Table 1) are useful for temporal relation determination. To examine the impact of each individual feature, we feed a single linguistic feature to the Bayesian Classifier learning algorithm one at a time and study the accuracy of the resultant classifier. The experimental results are given in Table 2. It shows that event classes have greatest accuracy, followed by conjunctions in the second place, and adverbs in the third in the close test. Since punctuation shows no contribution, we only use it as a syntactic feature to differentiate cases (2) and (3) mentioned in Section 4.2.

Feature	Accuracy			Feature	Accuracy		
	Close test	Open test 1	Open test 2		Close test	Open test 1	Open test 2
VS	53.4%	48%	30%	VA	57%	50%	37%
VC	56.6%	56%	49%	C	62.6%	52%	45%
TR	50.2%	46%	28%	U	51.8%	50%	32%
P	52.4%	49%	30%	T	57.2%	48%	32%
PS	59%	53%	38%	D	59.6%	55%	47%
VV	51%	49%	29%	EC	72.4%	69%	68%

Table 2 Impact of each individual linguistic feature

4.3.2 Features in Combination

We now use Bayesian Classifier introduced in Sections 4.1 and 4.2 to combine all the related temporal indicators and event classes, since none of the features can achieve a good result alone. The simplest way is to combine the features without distinction. The conditional probability $P(t_{ji}\,|\,c)$ is estimated by (E7'). This model is called Ungrouped Model (UG).

However, as illustrated in table 1, the temporal indicators play different roles in building temporal reference. It is not reasonable to treat them equally. We claim that the temporal indicators have two functions, i.e., representing the connections of the clauses, or representing the tense/aspect of the events. We identify them as connective words or tense/aspect markers and separate them into two groups. This allows features to be compared with those in the same group. Let $T=[T^1,T^2]$, where T^1 is the set of connective words and T^2 is the set of tense/aspect markers. We have $t_1^1,t_2^1,...,t_m^1\in T^1$ and $t_1^2,t_2^2,...,t_l^2\in T^2$, m and l are the number of the connective words and the tense/aspect markers in a sentence respectively. We assume that the occurrences of the two groups are independent. By taking both grouping and position features into account, we replace $\prod_{i=1}^{n}P(t_i\,|\,c)$ with $\prod_{k=1}^{2}\prod_{j=1}^{4}\prod_{i=1}^{n_j^k}P(t_{ji}^k\,|\,c)$, $k=1,2$ represents the two groups and $\sum_{k=1}^{2}n_j^k=n_j$. To build the grouping-based Bayesian Classifier, (E7') is modified as:

$$P(t_{ji}^k\,|\,c)=\frac{C(t_{ji}^k,c)+u}{\sum_{t_{ji}^k\in T^k}C(t_{ji}^k,c)+u\,|\,T^k\,|} \qquad \text{(E7'')}$$

($k=1,2$, $j=1,2,3,4$ and $i=1,2,...n_j$)

4.3.3 Grouping Features by Grammatical Categories or Semantic Roles

We partition temporal indicators into connective words and tense/aspect markers in two ways. One is simply based on their grammatical categories (i.e. POS information). It separates conjunctions (e.g., 然后, *after*; 因为, *because*) and verbs relating to causality (e.g., 导致, *cause*) from others. They are assumed to be connective words (i.e. $\in T^1$), while others are tense/aspect markers (i.e. $\in T^2$). This model is called Grammatical Function based Grouping Model (GFG).

Unfortunately, such a separation is ineffective. In comparison with UG, the performance of GFG decreases as shown in figure 2. This reveals the complexity of Chinese in connecting expressions. It arises from the fact that some other words, such as adverbs (e.g., 边…边, *meanwhile*), prepositions (e.g., 在, *at*) and position words (e.g., 之前, *before*), can also serve such a connecting function (see Table 1). Actually, the roles of the words falling into these grammatical categories are ambiguous. For instance, the adverb 才 can express an event happened in the past, e.g., "他才刚刚写完报告 (*He just finished the report*)". It can be also used in a connecting expression (such as 才…又…), e.g., "他才写完报告又去图书馆了 (*He went to the library after he had finished the report*)".

This finding suggests that temporal indicators should be divided into two groups according to their semantic roles rather than grammatical categories. Therefore we propose the third model, namely Semantic Role based Grouping Model (SRG), in which the indicators are manually re-marked as TI_j_pos or TI_at_pos[4].

Figure 2 shows the accuracies of four models (i.e. DM. UG, GFG and SRG) based on the three tests. Test 1 is the close test carried out on training data and tests 2 and 3 are open tests performed on different test data. DM (i.e., Default Model) assigns all incoming cases with the most likely class and it is used as evaluation baseline. In our case, it is SAME_AS, which holds 50.2% in training data. SRG model outperforms UG and GFG models. These results validate our previous assumption empirically.

[4] "j" and "at" are the tags representing connecting and tense/aspect roles respectively. "pos" is the POS tag of the temporal indicator TI.

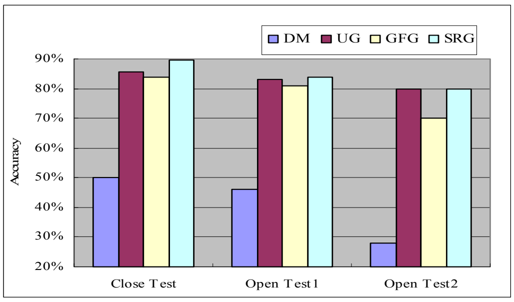

Figure 2 Comparing DM, UG, GFG and SRG models

4.3.4 Impact of Semantic Roles in SRG Model

When the temporal indicators are classified into two groups based on their semantic roles in SRG model, there are three types of linguistic features used in the Bayesian Classifier, i.e., tense/aspect markers, connective words and event classes. A set of experiments are conducted to investigate the impacts of each individual feature type and the impacts when they are used in combination (shown in Table 3). We find that the performance of methods 1 and 2 in the open tests drops dramatically compared with those in the close test. But the predictive strength of event classes in method 3 is surprisingly high. Two conclusions are thus drawn. Firstly, the models using tense/aspect markers and connective words are more likely to encounter over-fitting problem with insufficient training data. Secondly, different features have varied weights. We then incorporate an optimization approach to adjust the weights of the three types of features, and propose an algorithm to tackle over-fitting problem in the next section.

Method	Semantic Groups	Close test	Open test 1	Open test 2
1	Tense/aspect markers	71%	58%	40%
2	Connective words	75%	65%	57%
3	Event classes	66.6%	69%	68%
4	1+2	84.8%	70%	56%
5	1+3	76.6%	72%	66%
6	2+3	82.4%	84%	81%
7	1+2+3	89.8%	84%	80%
8	Default	50.2%	46%	28%

Table 3: Impact of Semantic Role based Groups

5. Weighted Bayesian Classifier

Let λ_1, λ_2, λ_3 be the weights of event classes, connective words and tense/aspect markers respectively. Then the Weighted Bayesian Classifier is:

$$\log\left(\frac{P(c \mid e_1, e_2, t_1, t_2, ..., t_n)}{P(\bar{c} \mid e_1, e_2, t_1, t_2, ..., t_n)}\right)$$

$$= \log\left(\frac{P(c)}{P(\bar{c})}\right) + \lambda_1 \log\left(\frac{P(e_1, e_2 \mid c)}{P(e_1, e_2 \mid \bar{c})}\right) \qquad \text{(E9)}$$

$$+ \lambda_2 \log\left(\frac{P(t_1^1, t_2^1, ..., t_m^1 \mid c)}{P(t_1^1, t_2^1, ..., t_m^1 \mid \bar{c})}\right) + \lambda_3 \log\left(\frac{P(t_1^2, t_2^2, ..., t_l^2 \mid c)}{P(t_1^2, t_2^2, ..., t_l^2 \mid \bar{c})}\right)$$

In order to estimate the weights, we need a suitable optimization approach to search for the optimal value of $[\lambda_1, \lambda_2, \lambda_3]$ automatically.

5.1 Estimating Weights with Simulated Annealing Algorithm

Quite a lot optimization approaches are available to compute the optimal value of $[\lambda_1, \lambda_2, \lambda_3]$. Here, Simulated Annealing algorithm is employed to perform the task, which is a general and powerful optimization approach with excellent global convergence (Kirkpatrick, 1983). Figure 3 shows the procedure of searching for an optimal weight vector with the algorithm.

```
1. $k = 1$, $t_k = T(t_{k-1})$
2. Generates a random change from the current weight vector $v_i$. The updated weight vector is denoted by $v_j$. Then
   computes the increasement of the objective function, i.e. $\Delta = f(v^j) - f(v^i)$.
3  Accepts $v_j$ as an optimal vector and substitutes $v_i$ with the following accept rate:
   $P(v^i \to v^j) = \begin{cases} 1 & \text{if } \Delta > 0 \\ \exp(\frac{\Delta}{t_k}) & \text{if } \Delta < 0 \end{cases}$
4  If $k < L_k$, lets $k = k + 1$, goes to step 2.
5  Else if $t_k < T_f$, goes to step 1.
6  Else stops looping and outputs the current optimal weight vector.
```

Figure 3 Simulated Annealing algorithm

In Figure 3, Markov chain length $L_k = 20$; temperature update function $T(t) = 0.9 * t$; starting point $v^0 = [\lambda_1^0, \lambda_2^0, \lambda_3^0]$ =[1,1,1]; initial temperature $t_0 = 20$ and final temperature $t_f = 10^{-8}$. Note that the initial temperature is critical for a simulated annealing algorithm (Kirkpatrick, 1983). Its value should assure that the initial accept rate is greater than 90%.

5.2 *K*-fold Cross-Validation

The accuracy of the classifier is defined as the objective function of the Simulated Annealing algorithm illustrated in Figure 3. If it is evaluated with the accuracy over all training data, the Weighted Bayesian Classifier may trap into over-fitting problem and lower the performance due to insufficient data. To avoid this, we employ *K*-fold Cross-Validation technique. It partitions the original set of data into *K* parts. One part is selected arbitrarily as evaluating data and the other *K*-1 parts as training data. Then *K* accuracies on evaluating data are obtained after *K* iterations and their average is used as the objective function.

5.3 Experimental Results

Table 4 shows the result of the experiment which compares WSRG (Weighted SRG) with SRG. We use error reduction to evaluate the benefit from incorporating weight parameters into Bayesian Classifier. It is defined as:

$$\text{error reduction} = \frac{error\_rate_{SRG} - error\_rate_{WSRG}}{error\_rate_{SRG}}$$

Model	Error Rate		
	Close Test	Open Test1	Open Test2
SRG	10.2%	16%	20%
WSRG	12.4%%	11%	13%
Error Reduction	-21.57%	31.25%	35%

Table 4 Compare WSRG with SRG on error rates

The experimental results show that the Weighted Bayesian Classifier outperforms the Bayesian Classifier significantly in the two open tests and it tackles the over-fitting problem well. To test Simulated Annealing algorithm's global convergence, we randomly choose several initial values and they finally converge to a small area [7.2±0.09, 5.8±0.02, 3.0±0.02]. The empirical result demonstrates that the output of a Simulated Annealing algorithm is a global optimal weighting vector.

6 Conclusions

Temporal reference processing has received growing attentions in last decades. However this topic has not been well studied in Chinese. In this paper, we proposed a method to determine temporal relations in Chinese by employing linguistic knowledge and machine learning approaches. Thirteen related linguistic features were recognized and temporal indicators were further grouped with respect to grammatical functions or semantic roles. This allows features to be compared with those in the same group. To accommodate the fact that the different types of features support varied importance, we extended Naïve Bayesian Classifier to Weighted Bayesian Classifier and applied Simulated Annealing algorithm to optimize weight parameters. To avoid over-fitting problem, *K*-fold Cross-Validation technique was incorporated to evaluate the objective function of the optimization algorithm. Establishing the temporal relations between two events could be extended to provide a determination of the temporal relations among multiple events in a discourse. With such an extension, this temporal analysis approach could be incorporated into various NLP applications, such as question answering and machine translation.

Acknowledgements

The work presented in this paper is partially supported by Research Grants Council of Hong Kong (RGC reference number PolyU5085/02E) and CUHK Strategic Grant (account number 4410001).

References

Allen J., 1983. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, 26(11):832-843.

Brent M., 1990. A Simplified Theory of Tense Representations and Constraints on Their Composition, In *Proceedings of the 28th Annual Conference of the Association for Computational Linguistics*, pages 119-126. Pittsburgh.

Bruce B., 1972. A Model for Temporal References and its Application in Question-Answering Program. *Artificial Intelligence*, 3(1):1-25.

Dorr B. and Gaasterland T., 2002. Constraints on the Generation of Tense, Aspect, and Connecting Words from Temporal Expressions. submitted to *Journal of Artificial Intelligence Research*.

Duda, R. O. and P. E. Hart, 1973. *Pattern Classification and Scene Analysis*. New York.

Friedman N., Geiger D. and Goldszmidt M., 1997. Bayesian Network Classifiers. *Machine Learning* 29:131-163, Kluwer Academic Publisher.

Goodman J., 2001. *A Bit of Progress in Language Modeling*. Microsoft Research Technical Report MSR-TR-2001-72.

Hitzeman J., Moens M. and Grover C., 1995. Algorithms for Analyzing the Temporal Structure of Discourse. In *Proceedings of the 7th European Meeting of the Association for Computational Linguistics*, pages 253-260. Dublin, Ireland.

Hornstein N., 1990. *As Time Goes By*. MIT Press, Cambridge, MA.

Kirkpatrick, S., Gelatt C.D., and Vecchi M.P., 1983. Optimization by Simulated Annealing. *Science*, 220(4598): 671-680.

Knight B. and Ma J., 1997. Temporal Management Using Relative Time in Knowledge-based Process Control, *Engineering Applications of Artificial Intelligence*, 10(3):269-280.

Langley, P.W. and Thompson K., 1992. An Analysis of Bayesian Classifiers. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pages 223–228. San Jose, CA.

Lascarides A. and Asher N., 1991. Discourse Relations and Defensible Knowledge. In P*roceedings of the 29th Meeting of the Association for Computational Linguistics*, pages 55-62. Berkeley, USA.

Li W.J. and Wong K.F., 2002. A Word-based Approach for Modeling and Discovering Temporal Relations Embedded in Chinese Sentences, *ACM Transaction on Asian Language Processing*, 1(3):173-206.

Moens M. and Steedmen M., 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics*, 14(2):15-28.

Passonneau R., 1988. A Computational Model of the Semantics of Tense and Aspect. *Computational Linguistics*, 14(2):44-60.

Reichenbach H., 1947. *The Elements of Symbolic Logic*. The Free Press, New York.

Siegel E.V. and McKeown K.R., 2000. Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights. *Computational Linguistics*, 26(4):595-627.

Singh M. and Singh M., 1997. On the Temporal Structure of Events. In *Proceedings of AAAI-97 Workshop on Spatial and Temporal Reasoning*, pages 49-54. Providence, Rhode Island.

Webber B., 1988. Tense as Discourse Anaphor. *Computational Linguistics*, 14(2):61-73.

Yarowsky D., 1994. Decision Lists for Lexical Ambiguity Resolution: Application to the Accent Restoration in Spanish and French. In *Proceeding of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 88-95. San Francisco, CA.