Applying Machine Learning to Chinese Temporal Relation Resolution

Wenjie Li
Department of Computing
The Hong Kong Polytechnic University, Hong Kong
cswjli@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

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

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

Abstract

Temporal relation resolution involves extraction of temporal information explicitly or implicitly embedded in a language. This information is often inferred from a variety of interactive grammatical and lexical cues, especially in Chinese. For this purpose, inter-clause relations (temporal or otherwise) in a multiple-clause sentence play an important role. In this paper, a computational model based on machine learning and heterogeneous collaborative bootstrapping is proposed for analyzing temporal relations in a Chinese multiple-clause sentence. The model makes use of the fact that events are represented in different temporal structures. It takes into account the effects of linguistic features such as tense/aspect, temporal connectives, and discourse structures. A set of experiments has been conducted to investigate how linguistic features could affect temporal relation resolution.

1 Introduction

In language studies, temporal information describes changes and time of changes expressed in a language. Such information is critical in many typical natural language processing (NLP) applications, e.g. language generation and machine translation, etc. Modeling temporal aspects of an event in a written text is more complex than capturing time in a physical time-stamped system. Event time may be specified explicitly in a sentence, e.g. "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 left implicit, to be recovered by readers from context. For example, one may know that "修成立交桥以后，他们解决了该市的交通问题 (*after the street bridge had been built, they solved the traffic problem of the city*)", yet without knowing the exact time when the street bridge was built. As reported by Partee (Partee, 1984), the expression of relative temporal relations in which precise times are not stated is common in natural language. The objective of relative temporal relation resolution is to determine the type of relative relation embedded in a sentence.

In English, temporal expressions have been widely studied. Lascarides and Asher (Lascarides, Asher and Oberlander, 1992) suggested that temporal relations between two events followed from discourse structures. They investigated various contextual effects on five discourse relations (namely narration, elaboration, explanation, background and result) and then corresponded each of them to a kind of temporal relations. Hitzeman et al. (Hitzeman, Moens and Grover, 1995) described a method for analyzing temporal structure of a discourse by taking into account the effects of tense, aspect, temporal adverbials and rhetorical relations (e.g. causation and elaboration) on temporal ordering. They argued that rhetorical relations could be further constrained by event temporal classification. Later, Dorr and Gaasterland (Dorr and Gaasterland, 2002) developed a constraint-based approach to generate sentences, which reflect temporal relations, by making appropriate selections of tense, aspect and connecting words (e.g. before, after and when). Their works, however, are theoretical in nature and have not investigated computational aspects.

The pioneer work on Chinese temporal relation extraction was first reported by Li and Wong (Li and Wong, 2002). To discover temporal relations embedded in a sentence, they devised a set of simple rules to map the combined effects of temporal indicators, which are gathered from different grammatical categories, to their corresponding relations. However, their work did not focus on relative temporal relations. Given a sentence describing two temporally related events, Li and Wong only took the temporal position words (including *before*, *after* and *when*, which serve as temporal connectives) and the tense/aspect markers of the second event into consideration. The proposed rule-based approach

was simple; but it suffered from low coverage and was particularly ineffective when the interaction between the linguistic elements was unclear.

This paper studies how linguistic features in Chinese interact to influence relative relation resolution. For this purpose, statistics-based machine learning approaches are applied. The remainder of the paper is structured as follows: Section 2 summarizes the linguistic features, which must be taken into account in temporal relation resolution, and introduces how these features are expressed in Chinese. In Section 3, the proposed machine learning algorithms to identify temporal relations are outlined; furthermore, a heterogeneous collaborative bootstrapping technique for smoothing is presented. Experiments designed for studying the impact of different approaches and linguistic features are described in Section 4. Finally, Section 5 concludes the paper.

2 Modeling Temporal Relations

2.1 Temporal Relation Representations

As the importance of temporal information processing has become apparent, a variety of temporal systems have been introduced, attempting to accommodate the characteristics of relative temporal information. Among those who worked on temporal relation representations, many took the work of Reichenbach (Reichenbach, 1947) as a starting point, while some others based their works on Allen's (Allen, 1981).

Reichenbach proposed a point-based temporal theory. This was later enhanced by Bruce who defined seven relative temporal relations (Bruce. 1972). Given two durative events, the interval relations between them were modeled by the order between the greatest lower bounding points and least upper bounding points of the two events. In the other camp, instead of adopting time points, Allen took intervals as temporal primitives and introduced thirteen basic binary relations. In this interval-based theory, points are 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 Ma and Knight (Ma and Knight, 1994).

In natural language, events can either be punctual (e.g. 爆炸 (*explore*)) or durative (e.g. 盖楼 (*built a house*)) in nature. Thus Ma and Knight'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 (*after the street bridge had been built, they solved the traffic problem of the city*)" as an example, the relation held between building the bridge (i.e. an interval) and solving the problem (i.e. a point) is *BEFORE*.

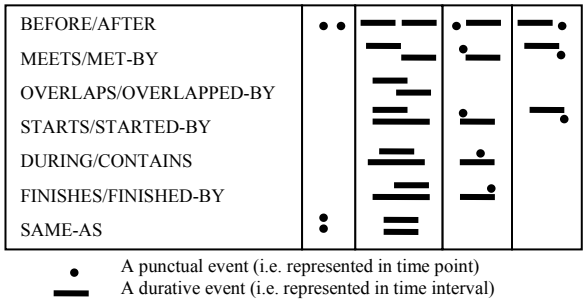

Figure 1. Thirteen temporal relations between points and intervals

2.2 Linguistic Features for Determining Relative Relations

Relative relations are generally determined by tense/aspect, connecting words (temporal or otherwise) and event classes.

Tense/Aspect in English is manifested by verb inflections. But such morphological variations are inapplicable to Chinese verbs; instead, they are conveyed lexically (Li and Wong, 2002). In other words, tense and aspect in Chinese are expressed using a combination of time words, auxiliaries, temporal position words, adverbs and prepositions, and particular verbs.

Temporal Connectives in English primarily involve conjunctions, e.g. *after*, *before* and *when* (Dorr and Gaasterland, 2002). They are key components in discourse structures. In Chinese, however, conjunctions, conjunctive adverbs, prepositions and position words are required to represent connectives. A few verbs which express cause and effect also imply a forward movement of event time. The words, which contribute to the tense/aspect and temporal connective expressions, are explicit in a sentence and generally known as *Temporal Indicators*.

Event Class is implicit in a sentence. Events can be classified according to their inherent temporal characteristics, such as the degree of telicity and/or atomicity (Li and Wong, 2002). The four widespread accepted temporal classes[1] are state, process, punctual event and developing event. Based on their classes, events interact with the tense/aspect of verbs to define the temporal relations between two events.

Temporal indicators and event classes are together referred to as *Linguistic Features* (see Table 1). For example, linguistic features are underlined in the sentence "(因为)修成立交桥(以后)，他们解决了该市的交通问题 *after*/*because* the street bridge *had been built* (i.e. a *developing event*), they *solved* the traffic problem of the city (i.e. a *punctual event*)".

[1] Temporal classification refers to aspectual classification.

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

Table 1 shows the mapping between a temporal indicator and its effects. Notice that the mapping is not one-to-one. For example, adverbs affect tense/aspect as well as discourse structure. For another example, tense/aspect can be affected by auxiliary words, trend verbs, etc. This shows that classification of temporal indicators based on part-of-speech (POS) information alone cannot determine relative temporal relations.

3 Machine Learning Approaches for Relative Relation Resolution

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

Relative relation resolution can be modeled as a relation classification task. We model the thirteen relative temporal relations (see Figure 1) as the *classes* to be decided by a classifier. The resolution process is to assign an event pair (i.e. the two events under concern)[2] to one class according to their linguistic features. For this purpose, we train two classifiers, a *Probabilistic Decision Tree Classifier* (PDT) and a *Naïve Bayesian Classifier* (NBC). We then combine the results by the *Collaborative Bootstrapping* (CB) technique which is used to mediate the sparse data problem arose due to the limited number of training cases.

3.1 Probabilistic Decision Tree (PDT)

Due to two domain-specific characteristics, we encounter some difficulties in classification. (a) Unknown values are common, for many events are modified by less than three linguistic features. (b) Both training and testing data are noisy. For this reason, it is impossible to obtain a tree which can completely classify all training examples. To overcome this predicament, we aim to obtain more adjusted probability distributions of event pairs over their possible classes. Therefore, a probabilistic decision tree approach is preferred over conventional decision tree approaches (e.g. C4.5, ID3). We adopt a non-incremental supervised learning algorithm in TDIDT (Top Down Induction of Decision Trees) family. It constructs a tree top-down and the process is guided by distributional information learned from examples (Quinlan, 1993).

3.1.1 Parameter Estimation

Based on probabilities, each object in the PDT approach can belong to a number of classes. These probabilities could be estimated from training cases with Maximum Likelihood Estimation (MLE). Let l be the decision sequence, z the object and c the class. The probability of z belonging to c is:

$$p(c \mid z) = \sum_{l} p(l, c \mid z) \approx \sum_{l} p(c \mid l) p(l \mid z) \quad (1)$$

let $l = B_1 B_2 ... B_n$, by MLE we have:

$$p(c \mid l) \approx p(c \mid B_n) = \frac{f(c, B_n)}{f(B_n)} \quad (2)$$

$f(c, B_n)$ is the count of the items whose leaf nodes are B_n and belonging to class c. And

[2] It is an *object* in machine learning algorithms.

$$p(l \mid z) = p(B_1 \mid z)p(B_2 \mid B_1, z)p(B_3 \mid B_1, B_2, z) \\ ...p(B_n \mid B_{n-1}...B_1, z) \quad (3)$$

where

$$p(B_m \mid B_{m-1}B_{m-2}...B_1, z) = \frac{p(B_m B_{m-1}B_{m-2}...B_1 \mid z)}{p(B_{m-1}B_{m-2}...B_1 \mid z)}$$

$$= \frac{f(B_m B_{m-1}B_{m-2}...B_1 \mid z)}{f(B_{m-1}B_{m-2}...B_1 \mid z)}, (m = 2,3,...,n).$$

An object might traverse more than one decision path if it has unknown attribute values. $f(B_m B_{m-1}B_{m-2}...B_1 \mid z)$ is the count of the item z, which owns the decision paths from B_1 to B_m.

3.1.2 Classification Attributes

Objects are classified into classes based on their *attributes*. In the context of temporal relation resolution, how to categorize linguistic features into classification attributes is a major design issue. We extract all temporal indicators surrounding an event. Assume m and n are the anterior and posterior window size. They represent the numbers of the indicators *BEFORE* and *AFTER* respectively. Consider the most extreme case where an event consists of at most 4 temporal indicators before and 2 after. We set m and n to 4 and 2 initially. Experiments show that learning performance drops when m>4 and n>2 and there is only very little difference otherwise (i.e. when m≤4 and n≤2).

In addition to temporal indicators alone, the position of the punctuation mark separating the two clauses describing the events and the classes of the events are also useful classification attributes. We will outline why this is so in Section 4.1. Altogether, the following 15 attributes are used to train the PDT and NBC classifiers:

$$TI_{e_1}^{l_4}, TI_{e_1}^{l_3}, TI_{e_1}^{l_2}, TI_{e_1}^{l_1}, class(e_1), TI_{e_1}^{r_1}, TI_{e_1}^{r_2},$$

$$wi / wo\ punc, TI_{e_2}^{l_4}, TI_{e_2}^{l_3}, TI_{e_2}^{l_2}, TI_{e_2}^{l_1}, class(e_2), TI_{e_2}^{r_1}, TI_{e_2}^{r_2}$$

l_i (i=1,2,3,4) and r_j (j=1,2) are the ith indictor before and the jth indicator after the event e_k (k=1,2). Given a sentence, for example, 先/TI_d 有/E0 了/TI_u 马车/n ，/w 才/TI_d 修/E2 了/TI_u 驿道/n 。/w, the attribute vector could be represented as: [0, 0, 0, 先, E0, 了, 0, 1, 0, 0, 0, 才, E2, 了, 0].

3.1.3 Attribute Selection Function

Many similar attribute selection functions were used to construct a decision tree (Marquez, 2000). These included information gain and information gain ratio (Quinlan, 1993), χ^2 Test and Symmetrical Tau (Zhou and Dillon, 1991). We adopt the one proposed by Lopez de Mantaraz (Mantaras, 1991) for it shows more stable performance than Quinlan's information gain ratio in our experiments. Compared with Quinlan's information gain ratio, Lopez's distance-based measurement is unbiased towards the attributes with a large number of values and is capable of generating smaller trees with no loss of accuracy (Marquez, Padro and Rodriguez, 2000). This characteristic makes it an ideal choice for our work, where most attributes have more than 200 values.

3.2 Naïve Bayesian Classifier (NBC)

NBC assumes independence among features. Given the class label c, NBC learns from training data the conditional probability of each attribute A_i (see Section 3.1.2). Classification is then performed by applying Bayes rule to compute the probability of c given the particular instance of $A_1,...,A_n$, and then predicting the class with the highest posterior probability ratio.

$$c^* = \arg\max_c score(c \mid A_1, A_2, A_3, ..., A_n) \quad (4)$$

$$score(c \mid A_1, A_2, A_3, ..., A_n) = \frac{p(c \mid A_1, A_2, A_3, ..., A_n)}{p(\bar{c} \mid A_1, A_2, A_3, ..., A_n)} \quad (5)$$

Apply Bayesian rule to (5), we have:

$$score(c \mid A_1, A_2, A_3, ..., A_n) = \frac{p(c \mid A_1, A_2, A_3, ..., A_n)}{p(\bar{c} \mid A_1, A_2, A_3, ..., A_n)}$$

$$= \frac{p(A_1, A_2, A_3, ..., A_n \mid c)p(c)}{p(A_1, A_2, A_3, ..., A_n \mid \bar{c})p(\bar{c})} \approx \frac{\prod_{i=1}^{n} p(A_i \mid c)p(c)}{\prod_{i=1}^{n} p(A_i \mid \bar{c})p(\bar{c})} \quad (6)$$

$p(A_i \mid c)$ and $p(A_i \mid \bar{c})$ are estimated by MLE from training data with Dirichlet Smoothing method:

$$p(A_i \mid c) = \frac{c(A_i, c) + u}{\sum_{j=1}^{n} c(A_j, c) + u \times n} \quad (7)$$

$$p(A_i \mid \bar{c}) = \frac{c(A_i, \bar{c}) + u}{\sum_{j=1}^{n} c(A_j, \bar{c}) + u \times n} \quad (8)$$

3.3 Collaborative Bootstrapping (CB)

PDT and NB are both supervised learning approach. Thus, the training processes require many labeled cases. Recent results (Blum and Mitchell, 1998; Collins, 1999) have suggested that unlabeled data could also be used effectively to reduce the amount of labeled data by taking advantage of collaborative bootstrapping (CB) techniques. In previous works, CB trained two homogeneous classifiers based on different independent feature spaces. However, this approach is not applicable to our work since only a few temporal indicators occur in each case. Therefore, we develop an alternative CB algorithm, i.e. to train two different classifiers based on the same feature spaces. PDT (a non-linear classifier) and NBC (a linear classifier) are under consideration. This is inspired by Blum and Mitchell's theory that two collaborative classifiers should be conditionally

independent so that each classifier can make its own contribution (Blum and Mitchell, 1998). The learning steps are outlined in Figure 2.

Inputs:	A collection of the labeled cases and unlabeled cases is prepared. The labeled cases are separated into three parts, training cases, test cases and held-out cases.
Loop:	While the breaking criteria is not satisfied
1	Build the PDT and NBC classifiers using training cases
2	Use PDT and NBC to classify the unlabeled cases, and exchange with the selected cases which have higher *Classification Confidence* (i.e. the uncertainty is less than a threshold).
3	Evaluate the PDT and NBC classifiers with the held-out cases. If the error rate increases or its reduction is below a threshold break the loop; else go to step 1.
Output:	Use the optimal classifier to label the test cases

Figure 2. Collaborative bootstrapping algorithm

3.4 Classification Confidence Measurement

Classification confidence is the metric used to measure the correctness of each labeled case automatically (see Step 2 in Figure 2). The desirable metric should satisfy two principles:

- It should be able to measure the uncertainty/ certainty of the output of the classifiers; and
- It should be easy to calculate.

We adopt entropy, i.e. an information theory based criterion, for this purpose. Let x be the classified object, and $C = \{c_1, c_2, c_3, ..., c_n\}$ the set of output. x is classified as c_i with the probability $p(c_i \mid x)$ $i = 1,2,3,..,n$. The entropy of the output is then calculated as:

$$e(C \mid x) = -\sum_{i=1}^{n} p(c_i \mid x) \log p(c_i \mid x) \quad (9)$$

Once $p(c_i \mid x)$ is known, the entropy can be determined. These parameters can be easily determined in PDT, as each incoming case is classified into each class with a probability. However, the incoming cases in NBC are grouped into one class which is assigned the highest score. We then have to estimate $p(c_i \mid x)$ from those scores. Without loss of generality, the probability is estimated as:

$$p(c_i \mid x) = \frac{score(c_i \mid x)}{\sum_{j=1}^{n} score(c_j \mid x)} \quad (10)$$

where $score(c_i \mid x)$ is the ranking score of x belonging to c_i.

4 Experiment Setup and Evaluation

Several experiments have been designed to evaluate the proposed learning approaches and to reveal the impact of linguistic features on learning performance. 700 sentences are extracted from Ta Kong Pao (a local Hong Kong Chinese newspaper) financial version. 600 cases are labeled manually and 100 left unlabeled. Among those labeled, 400 are used as training data, 100 as test data and the rest as held-out data.

4.1 Use of Linguistic Features As Classification Attributes

The impact of a temporal indicator is determined by its position in a sentence. In PDT and NBC, we consider an indicator located in four 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. 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 separating their roles. We introduce two methods, namely NA and SAP to check if the ambiguity affects the two learning approaches.

N(atural) **O**(rder): the temporal indicators between the two events are extracted and compared according to their occurrence in the sentences regardless which event they modify.

S(eparate) **A**(uxiliary) and **P**(osition) words: we try to resolve the above ambiguity with the grammatical features of the indicators. In this method, we assume that an indicator modifies the first event if it is an auxiliary word (e.g. 了), a trend verb (e.g. 起来) or a position word (e.g. 前); otherwise it modifies the second event.

Temporal indicators are either tense/aspect or connectives (see Section 2.2). Intuitively, it seems that classification could be better achieved if connective features are isolated from tense/ aspect features, allowing like to be compared with like. Methods SC1 and SC2 are designed based on this assumption. Table 2 shows the effect the different classification methods.

SC1 (Separate **C**onnecting words **1)**: it separates conjunctions and verbs relating to causality from others. They are assumed to contribute to discourse structure (intra- or inter-sentence structure), and the others contribute to the tense/aspect expressions for each individual event. They are built into 2 separate attributes, one for each event.

SC2 (Separate **C**onnecting words **2):** it is the same as SC1 except that it combines the connecting word pairs (i.e. as a single pattern) into one attribute.

EC (**E**vent **C**lass): it takes event classes into consideration.

Method	Accuracy	
	PDT	NBC
NO	82.00%	81.00%
SAP	82.20%	81.50%
SAP +SC1	80.20%	78.00%
SAP +SC2	81.70%	79.20%
SAP +EC	85.70%	82.25%

Table 2. Effect of encoding linguistic features in the different ways

4.2 Impact of Individual Features

From linguistic perspectives, 13 features (see Table 1) are useful for relative relation resolution. To examine the impact of each individual feature, we feed a single linguistic feature to the PDT learning algorithm one at a time and study the accuracy of the resultant classifier. The experimental results are given in Table 3. It shows that event classes have greatest accuracy, followed by conjunctions in the second place, and adverbs in the third.

Feature	Accuracy	Feature	Accuracy
PT	50.5%	VA	56.5%
VS	54%	C	**62%**
VC	54%	U	51.5%
TR	50.5%	T	57.2%
P	52.2 %	D	**61.7%**
PS	58.7%	EC	**68.2%**
VS	51.2%	None	50.5%

Table 3. Impact of individual linguistic features

4.3 Discussions

Analysis of the results in Tables 2 and 3 reveals some linguistic insights:

1. In a situation where temporal indicators appear between two events and there is no punctuation mark separating them, POS information help reduce the ambiguity. Compared with NO, SAP shows a slight improvement from 82% to 82.2%. But the improvement seems trivial and is not as good as our prediction. This might due to the small percent of such cases in the corpus.
2. Separating conjunctions and verbs relating to causality from others is ineffective. This reveals the complexity of Chinese in connecting expressions. It is because other words (such as adverbs, proposition and position words) also serve such a function. Meanwhile, experiments based on SC1 and SC2 suggest that the connecting expressions generally involve more than one word or phrase. Although the words in a connecting expression are separated in a sentence, the action is indeed interactive. It would be more useful to regard them as one attribute.
3. The effect of event classification is striking. Taking this feature into account, the accuracies of both PDT and NB improved significantly. As a matter of fact, different event classes may introduce different relations even if they are constrained by the same temporal indicators.

4.4 Collaborative Bootstrapping

Table 4 presents the evaluation results of the four different classification approaches. DM is the default model, which classifies all incoming cases as the most likely class. It is used as evaluation baseline. Compare with DM, PDT and NBC show improvement in accuracy (i.e. above 60% improvement). And CB in turn outperforms PDT and NBC. This proves that using unlabeled data to boost the performance of the two classifiers is effective.

Approach	Accuracy	
	Close test	Open test
DM	50.50%	55.00%
NBC	82.25%	72.00%
PDT	85.70%	74.00%
CB	88.70%	78.00%

Table 4. Evaluation of NBC, PDT and CB approaches

5 Conclusions

Relative temporal relation resolution received growing attentions in recent years. It is important for many natural language processing applications, such as information extraction and machine translation. This topic, however, has not been well studied, especially in Chinese. In this paper, we propose a model for relative temporal relation resolution in Chinese. Our model combines linguistic knowledge and machine learning approaches. Two learning approaches, namely probabilistic decision tree (PDT) and naive Bayesian classifier (NBC) and 13 linguistic features are employed. Due to the limited labeled cases, we also propose a collaborative bootstrapping technique to improve learning performance. The experimental results show that our approaches are encouraging. To our knowledge, this is the first attempt of collaborative bootstrapping, which involves two heterogeneous classifiers, in NLP application. This lays down the main contribution of our research.

In this pilot work, temporal indicators are selected based on linguistic knowledge. It is time-consuming and could be error-prone. This suggests two directions for future studies. We will try to automate or at least semi-automate feature selection process. An-

other future work worth investigating is temporal indicator clustering. There are two methods we could investigate, i.e. clustering the recognized indicators which occur in training corpus according to co-occurrence information or grouping them into two semantic roles, one related to tense/aspect expressions and the other to connecting expressions between two events.

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., 1981. An Interval-based Represent Action of Temporal Knowledge. In *Proceedings of 7th International Joint Conference on Artificial Intelligence*, pages 221-226. Los Altos, CA.

Blum, A. and Mitchell T., 1998. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory,* Madison, Wisconsin, pages 92-100

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

Collins M. and Singer Y, 1999. Unsupervised Models for Named Entity Classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 189-196. University of Maryland.

Dorr B. and Gaasterland T., 2002. Constraints on the Generation of Tense, Aspect, and Connecting Words from Temporal Expressions. (submitted to *JAIR*)

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.

Lascarides A., Asher N. and Oberlander J., 1992. Inferring Discourse Relations in Context. In *Proceedings of the 30th Meeting of the Association for Computational Linguistics*, pages 1-8, Newark, Del.

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.

Ma J. and Knight B., 1994. A General Temporal Theory. *The Computer Journal*, 37(2):114- 123.

Màntaras L., 1991. A Distance-based Attribute Selection Measure for Decision Tree Induction. *Machine Learning*, 6(1): 81–92.

Màrquez L., Padró L. and Rodríguez H., 2000. A Machine Learning Approach to POS Tagging. *Machine Learning,* 39(1):59-91. Kluwer Academic Publishers.

Partee, B., 1984. Nominal and Temporal Anaphora. *Linguistics and Philosophy*, 7(3):287-324.

Quinlan J., 1993. *C4.5 Programs for Machine Learning*. Morgan Kauman Press.

Reichenbach H., 1947. *Elements of Symbolic Logic.* Berkeley CA, University of California Press.

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

Wiebe, J.M., O'Hara, T.P., Ohrstrom-Sandgren, T. and McKeever, K.J, 1998. An Empirical Approach to Temporal Reference Resolution. *Journal of Artificial Intelligence Research*, 9:247-293.

Wong F., Li W., Yuan C., etc., 2002. Temporal Representation and Classification in Chinese. *International Journal of Computer Processing of Oriental Languages*, 15(2):211-230.

Yarowsky D., 1994. Decision Lists for Lexical Ambiguity Resolution: Application to the Accent Restoration in Spanish and French. In *Proceeding of the 32rd Annual Meeting of ACL*, San Francisco, CA.

Zhou X., Dillon T., 1991. A Statistical-heuristic Feature Selection Criterion for Decision Tree Induction. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 13(8): 834-841.