An Algorithm for Situation Classification of Chinese Verbs

Xiaodan Zhu, Chunfa Yuan
State Key Laboratory for Intelligent Technology and System, Dept. of Computer Science & Technology, Tsinghua University, Beijing 100084, P.R.C.

K.F.Wong , Wenjie.Li
Dept. of System Engineering and Engineering Management, Chinese University of HongKong

Abstract

Temporal information analysis is very important for Chinese Information Process. Comparing with English, Chinese is quite different in temporal information expression. Based on the feature of Chinese a phase-based method is proposed to deal with Chinese temporal information. To this end, an algorithm is put forward to classify verbs into different situation types automatically. About 2981 verbs were tested. The result has shown that the algorithm is effective.

1.*Introduction

We are now launching a research project on Events Extraction from Chinese Financial News, which requires us to extract the related temporal information from news. Temporal expressions in Chinese form a complex system. We cannot fully understand the temporal information only by extracting the verbs, adverbs, auxiliary words and temporal phrases. Instead, more profound analysis is needed. In this paper, we first introduce the temporal system of Chinese, then we put forward a method in dealing with Chinese temporal information, in which situation types is very important. Therefore, an algorithm is rendered to classify verbs into several situation types.

* Supported by National Natural Science Foundation of China (69975008) and 973 project (G1998030507)

1.1 Temporal System of Chinese

Commonly, Chinese linguists [3][4] think that the temporal system of Chinese includes three parts: phase, tense and aspect. Each of these represents some profile of temporal expression (these definitions are a little different from linguistic theory of English).

(1) Phase. A sentence may describe a static state or an action; an action may be durative or instantaneous; a durative action may indicate a terminal or not. All of these are the research fields of phase. So, static vs. dynamic, durative vs. instantaneous, telic vs. non-telic are three pairs of phase features. Phase depends fully on meaning. According to phase features, we can classify the verbs into different situation types.

(2) Tense. Tense describes the relations between an event (E), reference time (R) and speaking time (S). First, taking S as the origin, we can get three relations between R and S: if R is before S, the sentence describes past; if R is the same time as S, it describes present; if R is after S, it describes future. This is called primary tense. Secondly, we can get three relations between E and S: If E is before R, we call it anterior; if E is the same time as R, we call it simple; otherwise we call it posterior. This is called secondary tense. Therefore, there are nine tenses including anterior past, anterior future, simple future, posterior present, etc.

(3) Aspect . Aspect reflects the way we observe an event. For the same event, there are many perspectives. We can take the event as atomic and not consider its inner structure, and call it perfective. We can consider it being in process, and call it imperfective. For imperfective, we can observe it at a position before it, at the beginning of it, in the middle of it, etc. Different perspectives lead to different expressions in the language.

Phase, tense and aspect are not independent even though they are different conceptions; each of them can influence and restrict the others, ultimately building up the complex temporal system of Chinese.

1.2 Phase-based Chinese temporal information analysis

Most languages express temporal information through phase, tense and aspect, however, for different languages, the relative importance of the three parts is different. A very important feature of English is that tense and aspect are expressed by variation of predicates. But for Chinese, predicates keep the same form no matter how the tense and aspect are different.

Therefore, in English, temporal information analysis mainly considers tense and aspect, as well as temporal adjective and time words and phrases. But in Chinese, tense and aspect of a sentence are not very clear, verbs do not vary in form with the change of tense and aspect. So we suggest basing temporal information analysis on phase. We mainly perceive the situation type of a sentence, then roughly acquire tense from adverbs and auxiliary words. After considering the temporal phrases, we can understand the temporal information of single event fully. Finally, according to the absolute temporal information of single event, we can get the temporal relation between two events. Phase-based temporal information analysis has been used in our research on Event Extraction from Financial News, in which the most important and fundamental problem is to acquire the situation types of a sentence.

1.3 Situation Classification of Chinese Verbs

In the West, research on situation has a long history. The earliest can be traced to the times of Aristotle. In resent years, Western researchers have published a large volume of papers, which present many points of view. The most important are Vendler(1967), Bache(1982), and Smith (1985) . They approximately classify the situation as four types: state, activity, accomplishment, and achievement.

Chinese researchers have also done considerable work, among which the most typical research were done by Chen[3] and Ma[5].

Ma[5] stated that the situation of a sentence is fully determined by the situation of the main verb of the sentence. He use three phases: static, durative, telic to classify verbs into four situational types V1,V2,V3,V4.

	Static	Durative	Telic
V1	+	+	+
V2	-	+	+
V3	-	+	-
V4	-	-	+

Table 1.1

Chen[3] stated that the situation of a sentence not only depends on the main verb of the sentence but also on other parts of the sentence. That is, although the main verb is the most important in determining a sentence situation, other parts such as adverbs also have effect. Chen's classification is more detailed.

NO.	Verb types	Instances
(1)	Attribute	是(be), 等于(equal)
(2)	Mental state	相信(believe), 抱歉(regret)
(3)	Position	站(stand), 坐(sit), 躺(lie)
(4)	Action and Mental Activity	跳(jump), 想(think), 猜(guess)
(5)	Verb-object Structure	读书(read (books)), 唱歌(sing (songs))
(6)	Change	变化(change), 成为(become)
(7)	Directional Action	跑来(run up), 爬上(climb on)
(8)	Instantaneous Change	死(die), 躺(lie), 断(snap)
(9)	Instantaneous Action	坐(sit), 站(stand)
(10)	Verb-verb or Verb-adjective	推倒(push down), 切碎(smash (into pieces))

Table 1.2

Phase feture / Situation types	Static	Dura-tive	Telic	verb types (table above)
State	+			(1) (2) (3)
Activity	-	+	-	(3) (4) (5)
Accomplishment	-	+	+	(3) (4)
Simple change	-	-	+	(6) (7)
Complex change	-	-	-	(8) (9) (10)

Table 1.3

From the tables above, we can find that some words(such as (3) and (4) in table 1.3) can belong to more than one category, so Chen use modifiers, auxiliary words and prepositions to eliminate the ambiguity.

	State			Acti-vity	Accom-plishment	Complex change	Simple change
	(1)	(2)	(3)				
很+V	-	+	-	-	-	-	-
V+着	-	(-)	+	+	+	-	-
在+V	-	(-)	(-)	+	+	+	-
V+(了)+TQP+了(act)	-	-	-	+	+	+	-
V+(了)+TQP+了(state)	-	+	+	-	+	+	+

TQP: Time Quantity Phrase, (-) : in most case, it is ↪

Table 1.4

2. Our Classification Algorithm for Verbs' Situation

2.1 Guiding Thoughts

(1) Our algorithm is for information processing

•Ma[5] uses three pairs of phase features in classifying, but from which we can not get an automatic classification algorithm for computers; the classification can only be done manually.

•In linguistics, telicity is a phase feature used in classifying. In table 1.1 the difference between category V2 and V3, in table 1.3, the difference between "activity" and "accomplishment", are attributed to telicity. But in information process, we need not distinguish whether an event is telic or not. For example,

Exp.1

他在吹笛子。 (He is playing the flute)

他在吹"梁祝"。(He is playing a song "Liangzhu")

Chen[3] thinks that in Exp. 1, the first sentence has the features: dynamicity, and durativity, and non-telicity; it belongs to "activity". The second sentence has the features dynamicity, durativity, and telicity, because in the second sentence, there is a default terminal----when the song "liangzhu" is over, the action "play" is over, so the sentence belongs to "accomplishment" instead of "activity". However we think such discrimination is useless for information extraction, because telicity is an ambiguous concept itself. What we need is to acquire the exact duration of the event. So if we knew the event is durative or not, and got the temporal phrases, we can know terminal time of the event. Besides, whether an event is telic or not can not be attributed to collocation and only can be done manually(as the exp 1 shows). For these reasons, we consider the two verbs in Exp. 1 belonging to the same situational type, that is, we do not use talicity as a phase feature to classifying verbs.

(2) Separate classification of the verb situation from classification of the sentence situation.

Chen[3] points that some verbs belong to more than one category, and gives a method to distinguish these cases. To make the ideal more clear, we use two steps to complete the sentence situation recognition. In this paper, we render an algorithm to classify verbs into different categories, which is the basis of another research----recognition of sentence situation, which will be discussed in future work.

2.2 Classification Method

We classify the verbs into five categories , Att(Attribute), Men(Mentality), Act (Activity) , Ins(Instantaneous) , Amb (Ambiguous).

Att: 是(be), 等于(equal), 包含(include), 合(accord with)
Men: 喜欢(like), 轻视(belittle), 爱(love), 满意 (be satisfied with)
Act: 画(draw), 嘚叨(gab), 喝(drink), 跑步(run)
Ins: 爆炸(explore), 熄灭(extinguish), 断(snap), 发现 (discovery)
Amb: 坐(sit), 站(stand), 躺(lie), 跪(kneel), 带(bring), 挂(hang), 戴(wear), 安装(install)

Amb(Ambiguous) include those words which describe different situations in different context. For example:

Exp. 2:
他挂照片挂在墙上 (they hung the picture on the wall.)
照片挂在墙上 (Picture is hanging on the wall.)

In Exp. 2, the two sentences have the same predicate "挂" (hang). In the first sentence, "挂" describes an instantaneous action, but the second sentence describes a state. In English, forms of these two predicates are different; while in Chinese, they are the same. For this reason, we consider it ambiguous and indistinguishable without context.

We have pointed out previously that phase depends only on meaning. However different situational types collocate with different words. So the essence of our algorithm is replace semantic judgement with collocational judgement. For example, "Men(Mentality)" can follow "很" (very). Verbs in the "Amb" category can followed by "介宾(preposition-object)" structures, etc. The following is the set of collocational features.

	Static verbs		Amb	Act	Ins
	Att	Men			
Verb+了	-	+	+	+	+
很+Verb	-	+	-	-	-
Verb+着	-	(-)	+	(+)	-
在+Verb	-	(-)	(-)	+	-
Verb+介宾	-	-	+	-	+

Table 2.1

2.3 Implementation of the algorithm

According to table 2.1, a classification algorithm was designed, and we use two resources to implement our algorithm: The *Contemporary Chinese Cihai* [11] (which we will refer to as the *Cihai* below) dictionary and the *Machine Tractable Dictionary of Contemporary Chinese Predicate Verbs* [12](which we will refer to as the *predicate dictionary* below). The *Cihai* dictionary includes 12,000 entries and 700,000 collocation instances. *predicate dictionary* includes about 3000 verbs with their semantic information, case relations and detailed collocation information. These two dictionaries both include some of the collocation information that the algorithm needs.

Considering the features of these two dictionaries, we adjust part of our algorithm:
(1) In predicate dictionary, there is a slot named "verb type", which includes "transitive verb", "intransitive verb", "attributive verb", "linking verb" etc. So, at the beginning of the algorithm, we judge if the verb is a "linking verb" ("是(be)", "等于(equal)",etc) or a "possessive verb" ("具有" (have)). If it is, we directly classify the verb as "att(attribute)" without further processing.
(2) The *predicate dictionary* provides the case relation of verbs, and their semantic ategories. We restrict the agent of verbs in the "Men(mentality)" to belong to one of:

"{人}"(people), "{人类}"(human), "{人群}"(multitude), "{集体}"(collectivity), "{生物}(creatures)", "{信念}(belief)", "{动物}(animal)" (3) Because *Cihai* includes collocation instances instead of collocation relations, we should consider synonyms. To be exact, when we determine whether a verb belongs to "Men(Mentality)" or not, we judge if it can follow "很"(very) and synonyms such as "极其", "十分", "分外", "尤其", "非常". However, some seldom seen instances were included. All these cause some errors.

The final algorithm is as follows:

if (a verb is labeled as "linking verb" or "possessive verb" in *predicate dictionary*)
then the verb belongs to "Att(Attribute)"
else if (the verb can follow "很"(very) and synonyms "极其", "十分", "分外", "尤其", "非常") and (its agent's semantic belongs to set1*)
  then the verb belongs to "Men(Mentality)"
  else if (it can follow "在") or (be followed by"着")
    then if (it can be followed by "preposition-object" structure)
      then the verb belongs to "Amb(Ambiguous)"
      else the verb belongs to "Act(Activity)"
    else if (it can be followed by"了")
      then the verb belongs to"Ins(Instantaneous)"
      else the verb belongs to" unknown"

*set1={human, multitude, collectivity, creatures, belief, animal }

3. Results and Analysis

3.1 Results

We use the algorithm above to classify the 2981 words in predicate dictionary, at the same time, we do the classification manually, Table 3.1 is the result:

	Att	Men	Amb	Ins	Act	Un-known	Totel
Human	20	112	500	662	1683	4*	2981
Algo-Rithm	20	111	537	691	1519	103	2981

*this 4 words are not verb.

Table 3.1

Table 2.2 shows the details:

by algo-rithm	Classifying by human						
	Att	Men	Amb	Ins	Act	Non-verb	Totel
Att	20	0	0	0	0	0	20
Men	0	99	1	1	10	0	111
Amb	0	0	473	9	55	0	537
Ins	0	0	3	637	51	0	691
Act	0	1	12	2	1504	0	1519
Un-known	0	12	11	13	63	4	103
Totel	20	112	500	662	1683	4	2981

Table 3.2

Table 3.3 shows precision and recall:

	Att	Men	Amb	Ins	Act	Average
Precision	100.0	89.9	88.1	92.2	99.0	93.8
Recall	100.0	88.4	94.6	96.2	89.4	93.7

Table 3.3

3.2 Analysis

We mainly analyze the errors:

(1) Failure of algorithm

Chinese is a very complex language. Replacing semantic judgment by using collocations has limitation itself. For example, in most cases, whether a verb is durative or not can be decided by whether it can be used in such structure "verb+'着'" (In most case, "着" represents an action in progress). But some instantaneous verbs such as "敲(knock)", can also be used in such structure to express a repeated action.

(2) Errors caused by the resources

(2.1) Collocation incompleteness in *Cihai*: for example, "反对(disagree)" can collocate with "很"(very), but this collocation is not included in *Cihai*.

(2.2)Errors caused by *predicate dictionary*: It is obvious that a certain proportion of dictionary errors is inevitable. For example, though "恳求"(beg) can follow "在" to represent the action is in progress, it is not included in the corresponding slot of predicate dictionary.

(2.3) The inconsistency between the two dictionaries: For example, "崇敬"(admire), "在乎"(mind), and "看不起"(belittle) are included in *predicate dictionary* but not in *Cihai*. Although "抱歉"(regret) is included in *Cihai* , it is taken as an adjective instead of a verb.

4. Conclusion

In this paper, we advance a phase-based method to analyze temporal information in Chinese. For this purpose, an algorithm is rendered to classify verbs into different situation types. Because a verb's situation depends on the meaning of the verb, the essence of our algorithm takes advantage of collocations to avoid semantics. The result shows the algorithm is successful. We also believe that if the errors caused by resources were eliminated, the result would be improved significantly.

Although the five categories are defined by us, they can describe basic situations of Chinese. The classification algorithm itself is independent of resources, so it can be applied to other resources (dictionaries) if these resources include sufficient collocation information. Furthermore, Discarding dictionaries and doing classification directly on large-scale real corpus, especially in certain domain, deserve the future research.

Our algorithm is very useful for the future analysis of sentence situation for Information Extraction system and for dictionary construction.

References

[1]Message Understanding Conference Website, http://www.muc.saic.com.

[2]Message Understanding Evaluation and Conference: Proceedings of 3rd-6th APRA Workshops, Morgan Kaufmann Publishers Inc., 1996.

[3]Chen ping. Discussion On Temporal System of Contemporary Chinese: China Chinese Vol.6,1998.

[4]Gong Qianyan. Phase, Tense and Aspect of Chinese, Commercial Press.

[5]Ma Qingzhu. Time Quantity Phrase and Categories of Verbs: China Chinese. Vol.2,1981.

[6]Hu Yushu & Fan Xiao. Research on Verbs, Henan Univ. Press, 1995

[7]Hwang C.H. & Schubert L. K. Interpreting Tense, Aspect and Time Adverbials: A Compositional, Unified Approach : Proceeding of 1st International Conference in Temporal Logic, Bonn, Germany, July 1994.

[8]Allen J.F. Towards a General Theory of Action and Time: Artificial Intelligence, 23,123-154.

[9]Allen J.F. & George, F.. Action and Events in Interval Temporal Logic: Journal of Logic and Computation, Special Issue on Actions and Processes, 1994.

[10]Ion Androutsopoulos , Graeme Ritche & Peter Thanisch . Time ,Tense and Aspect in Natural Language Database Interface, CMP-LG Mar 1998.

[11]Ni Wenjie . Contemporary Chinese Cihai, People China Press, 1994.

[12]Chen Qunxiu . Designing and implement of Machine Tractable Dictionary of Contemporary Chinese Predicate Verbs, Proceedings of ICCC96, Singapore, Jun, 1996.