Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp. 53-94

Chinese Main Verb Identification: From Specification to Realization[1]

Bing-Gong Ding[*], Chang-Ning Huang[+] and De-Gen Huang[*]

Abstract

Main verb identification is the task of automatically identifying the predicate-verb in a sentence. It is useful for many applications in Chinese Natural Language Processing. Although most studies have focused on the model used to identify the main verb, the definition of the main verb should not be overlooked. In our specification design, we have found many complicated issues that still need to be resolved since they haven't been well discussed in previous works. Thus, the first novel aspect of our work is that we carefully design a specification for annotating the main verb and investigate various complicated cases. We hope this discussion will help to uncover the difficulties involved in this problem. Secondly, we present an approach to realizing main verb identification based on the use of chunk information, which leads to better results than the approach based on part-of-speech. Finally, based on careful observation of the studied corpus, we propose new local and contextual features for main verb identification. According to our specification, we annotate a corpus and then use a Support Vector Machine (SVM) to integrate all the features we propose. Our model, which was trained on our annotated corpus, achieved a promising F score of 92.8%. Furthermore, we show that main verb identification can improve the performance of the Chinese Sentence Breaker, one of the applications of main verb identification, by 2.4%.

Keywords: Chinese Main Verb Identification, Text Analysis, Natural Language Processing, SVM

---

[1] The work was done while the author was visiting Microsoft Research Asia.

[*] Department of Computer Science, Dalian University of Technology, 116023, China
Email: dbg_dlut@hotmail.com, dlhuangdg@263.net

[+] Microsoft Research Asia, Beijing, 100080, China
Email: cnhuang@msrchina.research.microsoft.com

1. Introduction

The main verb is the verb corresponding to the main predicate-verb in a sentence. Our task is to identify the main verb of the sentence, which is a critical problem in natural language processing areas. It is a prerequisite for diverse applications such as dependency parsing [Zhou 1999], sentence pattern identification [Luo 1995], Chinese sentence breaker, and so on.

Unlike western languages, Chinese grammar has little inflection information. Chinese verbs appear in the same form no matter whether they are used as nouns, adjectives, or adverbs. Below are some examples[2].

***Example* 1**

他 /r(ta1) 深 /d(shen1) 得 /v(de2) 学 生 /n(xue2sheng1) 的 /u(de) **喜 爱 /vn(xi3ai4)** 。/ww

*(He is deeply **loved** by his students.)*

***Example* 2**

乡 镇 企 业 /n(xiang1zhen4qi3ye4) 都 /d(dou1) 很 /d(hen3) **盛 行 /v(sheng5xing2)** 。/ww

*(The Township Enterprises are very **popular.**)*

***Example* 3**

毫 不 /d(hao2bu4) **放 松 /v(fang4song1)** 地 /u(de) 继 续 /v(ji4xu4) 推 进 /v(tui1jin4) 党风/n(dang3feng1) 廉政/n(lian2zheng4) 建设/vn(jian4she4) ，/ww

*(Never **relaxedly** advance the cultivation of party conduct and construction of a clean government.)*

In the Example 1 sentence, the word in bold, “喜爱” (love), is a verbal noun. In the Example 2 sentence, “盛行” (popular) is modified by “很” (very), so it functions as an adjective. In the Example 3 sentence, “放松” (relax) is followed by “地” (de)[3], so “放松” functions as an adverbial. Thus, if one wants to identify the main verb in a Chinese sentence, one faces a more

[2] If not specially pointed out, the following examples come from the PK corpus, which was released by the Institute of Computational Linguistics, Peking University, and is available at http://icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp. The corpus contains one month of data from *People's Daily* (January 1998). It has been both word segmented and part-of-speech tagged. “/r”, “/v” etc. are the part-of-speech tags. “/ww” denotes the end of the sentences.

[3] In Chinese, “地”(de) is used after an adjective or phrase to form an adverbial adjunct before the verb.

difficult task than in an English sentence since one cannot use morphological forms as clues. The second characteristic of Chinese verbs is that they have no specific syntactic function. Verbs can be used as subjects, predicates, objects, adverbial modifiers, or complements. So there are occasions when verbs are used consecutively. This can be shown by the following examples.

***Example* 4**

转移/v(zhuan3yi2) 不/d(bu4) **等于/v(deng3yu2)** **压缩/v(ya1suo1)** 。/ww

(Shifting does not mean compressing.)

***Example* 5**

大多数/m(da4duo1shu4) 人/n(ren2) 更/d(geng4) **反对/v(fan3dui4)** **提前/v(ti2qian2)** **举行/v(ju3xing2)** **大选/v[4](da4xuan3)** 。/ww

(Most people were opposed to holding the election ahead of time.)

***Example* 6**

恶化/v(e4hua4) 的/u(de) 经济/n(jing1ji4) **得到/v(de2dao4)** **改善/v(gai3shan4)** 。/ww

(The deteriorated economy was improved.)

In the above three sentences, the verb "转移" (shift) is used as the subject. Verbs like "等于" (mean), "反对" (oppose), and "得到" (get) are used as predicate-verbs. "压缩" (compress), "提前" (ahead of time), "举行" (hold), "大选" (election), and "改善" (improve) are used as objects. "恶化" (deteriorate) is used as an adjective modifier. Note that in Example 5, four verbs are used consecutively.

Therefore, the essence of the main verb identification problem is to identify the main verb among several verbs in a sentence that have no inflections at all, which is determined by the characteristics of Chinese grammar.

Although the lack of regular morphological tense markers renders main verb identification complicated, finding the main verb cannot be bypassed since it plays a central role in Chinese grammar [Lv 1980]. For example, suppose one is building a sentence pattern identification system. There are several reasons why we should identify the main verb first.

---

[4] In our corpus annotation, we tend to follow the annotation of the Peking Corpus and try to set aside part-of-speech annotation, which still needs discussion among researchers. For example, some researchers may argue that "大选(da4xuan3)" should be annotated as a noun. Since the original annotation of "大选" in the Peking corpus is "/v", we have not revised its part-of-speech tag to "/n".

- It has been shown that most sentences have verbs as predicates. So once the verb-predicate sentence pattern has been analyzed, almost all the sentence patterns can be analyzed [Lv *et al.* 1999]. Our investigation on the annotated corpus also produced the same results for Wu's assertion as reported in [Lv *et al*. 1999] (see section 3.3 for the reference).
- A sentence pattern identification system generally needs to identify the subject, object, adverbial modifier, and complement. All these syntactic parts are related to the main verb [Luo 1995].
- Many sentence patterns are embodied by a set of verbs. By identifying these main verbs first, we can classify the sentence patterns. For example, in the pivotal sentence, the main verbs tend to be "使" (shi3, have (sb. to do sth.)), "让" (rang4, let), "叫" (jiao4, ask), "请" (qing3, invite), "派" (pai4, send) etc. Another example is a sentence that has a clause as its object; in this case verbs such as "觉得" (jue4de2, feel), "希望" (xi1wang4, hope), "认为" (ren4wei2, think), "是" (shi4, be) are more likely to be main verbs.

The points mentioned above are particularly related to Chinese sentence pattern identification, but analogous arguments can easily be made for other applications. See for example, the discussion in [Zhou 1999] about the subject-verb and object-verb dependency relations and section 6 with regard to the Chinese Sentence Breaker.

Recently researchers have arrived at a consensus that large annotated corpora are useful for applying machine learning approaches to solve different NLP problems. When constructing a large corpus, such as the Penn TreeBank [Xia *et al.* 2000; Marcus *et al.* 1993] or Chinese chunking [Li *et al.* 2004], the design of the specification is the basis part of the work. With this idea in mind, we propose the use of main verb specification to cover the various linguistic phenomena and provide a mechanism to ensure that the inter-annotator consistency is as high as possible. The second motivation of our new specification is as follows: The definition problem involved in automatically identifying the main verb from the computational point of view has not been tackled in detail. To our knowledge, only Luo [1995] has studied a relatively simple definition. Since there has not been sufficient discussion of the specification of main verb, it is difficult to push the research of main verbs forward. Finally, while we were designing our specification, we found that there exist different complicated cases with respect to main verb definition (see section 3.2.3 for details). Thus, the first step in our work was to devolop a more clear definition of a main verb and tries to investigate its ambiguities. This was the real foundation of our work.

Previous studies focused on exploring different statistical and heuristic features in order to identify predicates. Heuristic rules [Luo 1995] and statistical methods like the Decision Tree [Sui and Yu 1998b] have been used to identify predicates. But they either use one of the methods or just use them separately [Gong *et al.* 2003]. We believe it is better to combine the

heuristic and statistical features together. In this paper, we treat the Main Verb Identification (MVI) task as a binary classification problem of determining whether the VP is the MVP or not. We define the main VP (MVP) as the VP chunk in which the head word is the main verb. Here, a verb chunk, VP, is composed of a head verb with its pre-modifier or the following verb-particles, which form a morphologically derived word sequence [Li *et al.* 2004]. The head word of the VP is the verb that dominates the VP. For example, if the main verb is "返回" (fan3hui2, return), then the chunk "连忙/d 返回/v" (lian2mang2 fan3hui2, immediately return) is the MVP, in which the head verb is "返回". We can have a one-to-one mapping between main verbs and MVPs. Therefore, identifying the main verb is equal to identifying the MVP with additional available chunking information. So in the following, "MVP Identification" and "Main Verb Identification" are interchangeable.

We employ one of the most successful machine learning approaches, the Support Vector Machine (SVM), as the classifier. Our method combines lexicalized knowledge with statistical information. We evaluated the performance of our MVI system on the PK corpus, which is an annotated test set. The MVP recall and precision rates reached 92.1% and 93.6% respectively. The main aspects of our research are as follows:

- We investigated in detail the distribution of simple sentence structure and main verbs. After that, we tried to develop our specification and conducted a pilot study on the complicated aspects of the main verb definition.
- Because shallow parsing provides useful information such as chunks and chunk type information, we propose conducting MVI on the results of chunking [Li *et al.* 2004]. Our experiments show the MVI performance based on chunking is better than that of part-of-speech.
- We propose new features based on careful observations of the training corpus. The features are divided into two categories, local and contextual features. Among them, VP position, VP length, Probability of head verbs being MVPs, and Anti-patterns are all new features that we propose. Although they are simple, they work well in MVI.

The rest of the paper is structured as follows. Section 2 presents related works. Section 3 describes in detail the specification of the main verb and how the MVI is handled by our approach. Section 4 gives experimental results. Section 5 presents error analysis and discussion. Section 6 presents an application of main verb identification, the Chinese Sentence Breaker. Finally, we draw conclusions and suggest future work in section 7.

2. Related Works

The Chinese language is a typical SVO sequence language, in which 'V' is the main verb in the sentence [Jin and Bai 2003]. The problem of main verb identification has been studied

extensively by Chinese linguists for a long while [Lv 1980; Ding *et al.* 1961; Zhang 1982; Huang 1987; Fan 1995; Liu *et al.* 2002]. Since the definition of the main verb is related to different verb-predicate sentence patterns, linguists usually describe the different kinds of main verbs in the context of verb-predicate sentence patterns.

[Liu *et al.* 2002] divided verb-predicate sentences into five types: predicates that (1) take no object, (2) take a single object, (3) take double objects, (4) include an adverbial modifier, or (5) include complements. Based on this classification scheme, seven specific verb-predicate sentence patterns were also proposed and discussed individually, including "是" (shi4, to be) sentences, "有" (you3, have) sentences, series-verb sentences, pivotal sentences, existential sentences, "把" (ba3) sentences, and "被" (bei4, passive voice) sentences[5].

[Fan 1995] introduced a verb-predicate sentence pattern framework that includes seven subdivided sentence patterns, which overlap with Liu's classification. For example, SV (Subject-Verb), SVO (Subject-Verb-Object), SZV (Subject-Adverbial -Verb), and SVB (Subject-Verb-Complement) patterns in Fan's framework are similar to (1), (2), (4), and (5) in Liu's work. Other sentence patterns include SVL (Subject-Coordination), SCT (Subject-Series-Verb), and SVD (Subject-Duplicate-Verb). Detailed information can be found in his book. The reader should be aware that an SVL like "他一边走一边说" (He talks while walking) or "我们爱祖国爱人民" (We love both our motherland and our people) is equivalent to a series-verb instead of a sentence with verb-coordination in our definition (see section 3.1 for details). Fan's and Liu's works differ in that Fan tries to incorporate more sentence patterns into a single framework. For example, Fan further subdivides SZV into eight specific verb-predicate sentence patterns, like "被" (bei4, passive voice) sentences, "使" (shi3, let) sentences, "从" (cong2, from) sentences, etc. Fan also further subdivides SVB into seven constructions, like the "verb-resultative construction," "verb-得 construction," etc.

A particular feature of Huang's work [1987] is the examples he provides from real texts. The sentence patterns listed in his work are similar to those in [Liu *et al.* 2002].

[Zhang 1982] divided verb-predicate sentences into eleven types: verb sentences, verb-object phrase sentences, verb-compliment phrase sentences, modifier-verb phrase sentences, series-verb phrase sentences, pivot-verb phrase sentences, series-verb combined with pivot-verb phrase sentences, "把" sentences, "被" sentences, the negative form of verb-predicate sentences and the interrogative form of verb-predicate sentences. Similar to [Liu *et al.* 2002] and [Fan 1995], Zhang regards the adverbial-modifier as the basis for subdividing the verb-predicate sentence pattern. However, the author in [Lv 1980] did not use this kind of basis for classification. In addition, unlike [Lv 1980] and [Liu *et al.* 2002], Zhang

[5] "是"(shi4, to be), "有"(you3, have), "把"(ba3, ba), and "被"(bei4, passive voice) sentences are Chinese sentences which contain the above words.

uses the whole verb-object phrase, the verb-complement phrase, and the modifier-verb phrase to subdivide the verb-predicate. However, in our specification, longer phrases or whole phrases, such as whole verb-object phrases, are recursively defined. This categorization scheme cannot be used to subdivide our verb-predicate sentence pattern since a shallow parser cannot provide such information.

The findings in [Lv 1980] were the earliest and most widely ones accepted by other linguists. According to the different sentence structures, the author in [Lv 1980] introduced 13 types of verb-predicate sentence patterns. See Table 1.

Table 1. Verb-predicate sentence patterns in [Lv]

1. Transitive Verb Sentence
2. Intransitive Verb Sentence
3. Double Object Sentence
4. A sentence whose object is a verb
5. A sentence whose object is a clause
6. A sentence whose object is number
7. A sentence whose object is placed before the predicate
8. “把 (ba3)” Sentence
9. Passive Voice Sentence
10. Complement Sentence
11. Existential Sentence
12. Series Verb Sentence
13. Pivotal Sentence

Transitive Verb Sentence

	Subject	Adverbial modifier	**Verb**	Accusative Object	Non Accusative Ojbect	Auxiliary
A	你 她 你 她	从前 最近	**学过** **唱过** **会写** **吃**	英语	 女高音 这种笔 食堂	吗？ 吗？ 了
B	通县 这 晚上 这位同志	已经	**属于** **成为** **不如** **姓**	北京市 制度 早晨 李		

Figure 1. One example of a verb-predicate sentence pattern

For each type of the sentence, for example the Transitive Verb Sentence shown in Figure 1, the author of [Lv 1980] provides the predicate in the sentence pattern.

From the above discussion, we can conclude that when linguists describe and further subdivide verb-predicate sentences, an important basis of their work is the object of the predicate. For example, among the thirteen kinds of verb predicates in [Lv 1980], the first eight kinds of sentence patterns are subdivided according to the type of object. However, our work is different from theirs because we pay closer attention to main verb types in verb-predicate sentences than to object types. The reason for this is shown in the following example.

> [MVP 通知/v(tong1zhi1, inform)][NP 他们/r(ta1men2, them)][VP 准备/v(zhun3bei4, prepare for)][MP 三/m(san1, three) 天/n(tian1, days)] 的/u(de) [NP 干粮/n(gan1liang2, solid food)] 。/ww

See the above example cited from [Meng *et al.* 2003]. In this example, the sentence is explained as being a pivotal sentence like a) in [Meng *et al.* 2003]. Obviously the above sentence takes more than one parse, such as b), c), and d), if syntactic information only is available.

a) **Pivotal sentence**: [Piv-O 他们] [Piv-V 准备] 三 天 的 干粮
Note: "他们" is the pivotal object, which acts as both the object of "通知" and the subject of "准备".

b) **Series verb**: [Object 他们] [2nd-V 准备] 三 天 的 干粮
Note: "通知" and "准备" are two series verbs. "他们" acts as the object of "通知".

c) **Clause as object**: [Object 他们 准备 三 天 的 干粮]
Note: The whole clause "他们准备三天的干粮" acts as the object of "通知".

d) **Double objects**: [Obj1 他们] [Obj2 准备 三 天 的 干粮]
Note: "通知" takes double objects including "他们" and "准备三天的干粮".

Since it is hard to employ consistent annotation in such sentences and we prefer that our annotation be theoretically neutral, in our specification, we subdivide a verb-predicate sentence into four types, including simple verb-predicate sentences, series-verb sentences, pivotal sentences, and sentences with verb-coordination, instead of using the objects of their predicates.

The Chinese Penn TreeBank (CTB) is a large-scale bracketed corpus of hand-parsed sentences in Chinese [Xia *et al.* 2000; Xue and Xia 2000]. The annotation of the Chinese Penn Treebank is more complete because they annotate everything, whereas currently we only annotate verb predicates. Compared with the "Guideline for Bracketing in the Chinese Penn TreeBank" [Xue and Xia 2000], our specification is different in that the goal of the CTB is to

annotate linguistically-standard and non-controversial **parse trees**, while the goal of our MVP annotation is based on **chunking** which is relatively easily parseable. For this reason, the guideline of CTB is not entirely identical to our specification. Other differences are listed as follows.

- The annotation of CTB is based on sentences that end with periods, exclamation marks, or question marks. Our specification defines the main verbs of Chinese simple sentences (see section 3.2.1 for the reference).
- Since we only focus on the output of chunking instead of whole parsed trees as in CTB, the MVP in our specification is a verb chunk with the main verb, while the predicate in CTB may be a whole phrase. For example, in CTB, we have the following:

  (IP (NP-PN-SBJ (NR 张三 zhang1san1, Zhangsan))
  (VP (VV 应该 ying1gai1, should)
  (VP (VV 参加 can1jia1, join)
  (NP-OBJ (NN 会议 hui4yi4, meeting)))))

  "In the above example, the lowest level VP (VP 参加会议) is the predicate," whereas based on our parsed chunk results, [VP 应该/v 参加/v] is annotated as an MVP in this sentence according to our specification.
- In CTB, "…a VP is always a predicate, -PRD is assumed….." However, in our specification, we only tag the main verb, that is, the verb corresponding to the main predicate-verb in the sentence. This annotation scheme is consistent to the sentence analysis methodology of Chinese linguists [Lv 1980].
- CTB also tags non-verbal predicates, such as ADJP/NP etc. In our specification, we don't consider this case since our focus is verb-predicate sentences.

Linguists provide a grammatical view of Chinese sentences by analyzing them. Identifying the main verb automatically is a task faced by many computational linguists. Most of their works have focused on the identification process instead of on the definition of the main verb. Previous works on MVI can be grouped into three categories: heuristic methods [Luo 1995; Sui and Yu 1998a]; statistical methods [Chen and Shi 1997; Sui and Yu 1998b]; first heuristic and then statistical methods [Gong *et al.* 2003].

Heuristic methods were introduced in the early stage of MVI research. Some proposed approaches depend on linguists' knowledge; for example, Luo [1995] used hand-crafted rules to identify predicates. The rules are related to auxiliary words, such as "的(de)" or "得(de)",

or to numerical or temporal words. Other approaches employ a bilingual corpus to extract rules, for example, Sui and Yu's [1998a] method. However a bilingual corpus is not always available.

Statistic methods were proposed in [Chen and Shi 1997] and [Sui and Yu 1998b]. Both of these works are based on verb sub-categorization information. But their categorization frameworks are different. Chen and Shi's work [1997] uses only part-of-speech information to decide on the main verb. Sui and Yu [1998b] use not only sub-categorized part-of-speech information but also lexicalized context information, such as "的". Both static and the context features are integrated into a decision tree model.

[Gong *et al.* 2003] first used rules to filter quasi-predicates. The features used include the part-of-speech of the quasi-predicate, the contextual part-of-speech, and the contextual words like "的". Then each feature's weight is calculated from training data. The combined weights are used to determine the predicates in the sentences.

The works noted above except that in [Chen and Shi 1997] presume that the sentence boundary has been given. All of them detect predicates in simple sentences. However, they have a deficiency in that in real text, the sentence boundaries are not provided naturally. Another difference is that the above works identify verb predicates, nominal predicates and adjective predicates. In our work, we focus on verb-predicate since both previous [Lv *et al.* 1999] and our own observations show that the sentences with verb-predicates make up the most part in corpus.

Another point is that some of the above works use correct verb sub-categorization information as input [Chen and Shi 1997; Sui and Yu 1998b]. They do not provide main verb identification evaluation results, where verb sub-categorization needs to be done automatically as a preprocessing step performed on raw text. Although the task of verb sub-categorization has long been studied in the Chinese community, the performance achieved has not been satisfactory. Thus in our work, we make use of more reliable knowledge; for example, we will provide a closed set of specific verbs whose objects can include multiple clauses, rather than sub-categorization information in general.

Finally, it is difficult to compare our results with the results of related works because the test corpora used may be quite different and there are also some differences in the definitions of the main verb. Thus, we hope that our introduction of a clear specification and corpus for main verb identification will enable future researchers to compare their results with ours.

3. Our Solution

3.1 Motivation for Developing Another Type of Specification

One reason for designing a specification is to ensure consistency of the corpus. In the "guideline of bracketing the Chinese", Xue and Xia [2000] explain this issue as follows:

"Without doubt, consistency is one of the most important considerations in designing the corpus. . . .Many things can be done to ensure consistency, one of them is to make sure that the guidelines are clear, specific and consistent. . . .We also try to ensure that the guidelines cover all the possible structures that are likely to occur in the corpus. . . ."

The above description indicates that a clear and wide coverage specification will ensure consistency of the annotated corpus. However, such a specification is not available publicly for main verb identification. To our knowledge, Luo [1995] was the first and the only one to propose a relatively simple definition. There are several deficiencies, however, in his specification. First, the definition is based on verb sub-categorization, which has been long criticized by linguistic community. Secondly, some parts of the definition are relatively simple and unclear. For example, "the verbs that have the subcategorized part-of-speech vgo or vgs etc. will be main verb in general cases; the verbs that have the part-of-speech vgn or vgv etc. will be main verbs in some cases or the modifiers of predicate-verb in other cases." But the author does not explain in which cases this assertion is true. Finally the proposed verb analysis using rules of exclusion does not cover some commonly used sentence patterns, such as series-verb sentences or verb-coordination sentences.

Thus, we propose another type of specification with the following characteristics.

- In order to ensure that the most important syntactic relations are covered, we base our main verb definition on various verb-predicate sentences.
- For specific purposes, our definition makes use of more reliable knowledge, such as a closed set of certain verbs whose objects can include multiple clauses rather than sub-categorization information in general.
- To deal with ambiguous syntactic constructions, we adopt a scheme in which we preserve the basic information and make the structures easily converted to structures following other annotation scheme. A similar scheme was used in [Xia *et a.l* 2000] and [Lai and Huang 2000].
- A lot of different complicated cases are studied, and the findings help make the specification's description clearer.

3.2 Design Specification

In this paper, we propose to define the main verb based on a simple sentence structure for the following reasons.

- A simple sentence is a sentence with only one predicate, and in our definition each predicate includes only one main verb if any. This guarantees that the main verb will have a unique operational definition.
- Chinese linguists have provided simple sentence structures in details, which have less disagreement between them. Since main verbs are related to simple sentence structures, we suppose there will be less disagreement in main verb definition with the help of simple sentence structures.

Because our annotation is based on a simple sentence, we firstly define the simple sentence and then the predicate, especially the predicate-verb if one exists, of each simple sentence. Then, we discuss in detail on the complicated aspects of our spec design and corpus annotation. This discussion will help to uncover the difficult point of the main verb identification.

3.2.1 Sentence Definition

Chinese sentences are of two types: simple sentences and complex sentences. The boxes above the dashed line in Figure 2 show the widely accepted sentence pattern classification [Lv 1980; Ding *et al.* 1961]. In our specification, since we pay more attention to main verb types in verb-predicate sentences instead of object types, we subdivide verb-predicate sentences into four types as shown below the dashed line in Figure 2.

Definition 1:

A simple sentence is a sentence with only one predicate-verb.

The predicate of a simple sentence can be a verb, an adjective, a noun, or a subject-predicate in Chinese [Liu *et al.* 2002]. Accordingly, simple sentences are categorized as verb-predicate, adjective-predicate, noun-predicate and subject-predicate sentences, respectively. Here, a subject-predicate sentence has a subject-predicate phrase as its predicate. For example, in the sentence “他(ta1) 肚子(du4zi1) 疼(teng2)” (He has a stomach-ache), “肚子疼” is a subject-predicate phrase acting as the predicate, while “他” is the subject of the sentence. In our specification, we only focus on simple sentences with verb-predicates.

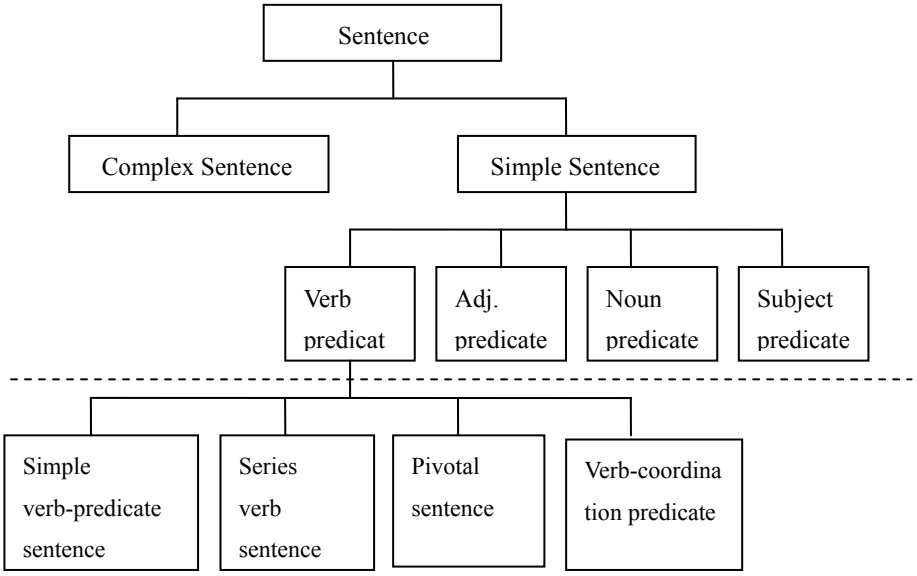

Figure 2. Sentence pattern classification

Definition 2:

A complex sentence is made up of two or more simple sentences. The simple sentences in one complex sentence can not be included each other.

Definition 3:

In a complex sentence，each *sub-sentence* is a sentence, which can be either a complex sentence or a simple sentence.

Another related topic that should be introduced is punctuation at the end of sentence. In general, "。|？|！|；" are punctuation used at the end of a sentence in Chinese. Sometimes "，|：|——|…" can also be seen as the end of a sentence if it has the main verb. See example 31 in section 6.

3.2.2 Main Verb Definition

Definition 4:

The main verb is the predicate-verb, if one exists, in a simple sentence. It corresponds to a tensed verb in English.

In this paper, we will only discuss the main verb in a verb-predicate sentence. Each verb-predicate sentence contains only one main verb, which is the predicate-verb of the sentence. Verb-predicate sentences can be classified into four types shown in Figure 2. Some examples of verb-predicate sentences are shown below.

***Example* 7** (simple verb-predicate sentence)

[NP 张/nr(zhang1) 晓伟/nr(xiao3wei3)] **[VP 坚决/ad(jian1jue2) 不/d(bu4) 收/v(shou1)]** 。/ww[6]

*(Zhang xiaowei **resolutely refused** to accept.)*

***Example* 8** (pivotal sentence)

[VP 必须/d(bi4xu1) 先/d(xian1) 请/v(qing3)] [NP 外国/n(wai4guo2) 专家/n(zhuan1jia1)] [VP 运行/v(yun4xing2)] [VP 管理/v(guan3li3)] 。/ww

*([One] **must first invite** the foreign expert to run and manage [it].)*

***Example* 9** (series-verb sentence)

[NP 张/nr(zhang1) 晓伟/nr(xiao3wei3)] **[VP 连忙/d(lian2mang2) 返回/v(fan3hui2)]** [NP 大/a(da4) 水潭/n(shui3tan2) 边/n(bian1)] [VP 去/v(qu4) 找/v(zhao3)] ，/ww

*(Zhang xiaowei **immediately returned to** the big puddle to search for [it])*

***Example* 10** (sentences with verb-coordination predicate)

[NP 交通/n(jiao1tong1) 肇事/vn(zhao4shi4)] [SP 后/f(hou4)] ，/w [NP 肇事/vn(zhao4shi4) 司机/n(si1ji1)] **[VP 伪造/v(wei3zao4)]** 、/w [VP 破坏/v(po4huai4)] [SP 现场/s(xian4chang3)] 。/ww

*(After the traffic accident, the trouble-making driver **falsified** and destroyed the scene.)*

In the above examples, the main verb in each sentence has been underlined. Without doubt, in simple verb-predicate sentences, the main verb is the predicate verb. In a series-verb sentence, a pivotal sentence, or a sentence with a verb-coordination predicate, the first predicate-verb of that construction is defined as the main verb of the sentence.

A serious concern with main verb definitions is the treatment of different syntactic constructions in different researchers' works. For instance, there is another point of view that both of the verbs in a verb-coordination sentence can be main verbs. However since there exist different levels of verb coordination, such as word level, phrasal level, and even clause level coordination [Xue and Xia 2000], we adopt a scheme similar to that used in [Lai and Huang 2000; Xia *et al.* 2000]. What we do is label the first verb as the main verb, preserve the VP information, and leave deeper analysis of verb-coordination for future work. From another point of view, it is easier to convert our annotation to other specifications with the preserved

[6] Chunks tags here are annotated according to our chunk spec [Li *et al.* 2004].

information.

3.2.3 Complicated Cases in Main Verb Annotation

Sentences in running text are complicated. To maintain inter-annotator consistency during corpus annotation, we not only perform cross-validation but also examine the phenomena that appear in our corpus annotation. This helps us to understand the problem of main verb identification. In the following, we classify the complicated cases into six types.

1) Verbs in a non verb-predicate sentence

Verbs or verb chunks (VPs) in a non verb-predicate sentence, whose predicates are formed by an adjective, nominal, or subject-predicate phrase, should not be treated as main verbs or MVPs. See Example 11 below.

***Example* 11**

[NP 我/r(wo3)] **[VP 吃/v(chi1)]** 的/u(de) [NP 邱县/ns(qiu1xian4) 饭/n(fan4)] ，/ww **[VP 喝/v(he1)]** 的/u(de) [NP 邱县/ns(qiu1xian4) 水/n(shui3)] ，/ww **[VP 当/v(dang1)]** 的/u(de) [NP 邱县/ns(qiu1xian4) 官/n(guan1)] ，/ww

(What I ate [was] Qiuxian's meal. What I drank [was] Qiuxian's water. What I worked as [was] a Qiuxian's officer.)

Note: These three sentences are sentences with predicates that are formed by subject-predicate phrases. All of them share the same subject, "[NP 我/r]" (I). "[VP 吃/v] 的/u [NP 邱县/ns 饭/n]" (what I ate) is a subject-predicate phase, in which the 的-structure "[VP 吃/v] 的/u" (ate + de) acts as a nominal subject, while [NP 邱县/ns 饭/n] (Qiuxian's meal) is a nominal-predicate. Thus, no main verbs can be found in these three sentences.

***Example* 12**

[NP 人们/n(ren2men2)] [VP 生活/v(sheng1huo2)] [ADJP 很/d(hen3) 苦/a(ku3)] 。/ww

(People's lives are very bitter.)

Note: This is a subject-predicate sentence in which the subject-predicate phrase [VP 生活/v] [ADJP 很/d 苦/a] acts as the predicate of the sentence. Thus, "生活"(life) should not be tagged as an MVP.

In example 12, annotators tend to tag "生活" (life) as a MVP because they incorrectly analyze

verb-predicate sentences and subject-predicate sentences with VPs.

2) Auxiliary Verbs

Auxiliary verbs are a special subdivision of verbs. Typically, they are placed before a verb, e.g., "会跳舞" (hui4tiao4wu3, be able to dance). In our specification, there is a closed set of 26 auxiliary verbs, including 能 (neng2, can), 会 (hui4, be able to), 可以 (ke3yi3, may), 应该 (ying1gai1, should) etc. However, these auxiliary verbs in the PK corpus share the same part of speech tag: "v".

As for the question of whether the auxiliary verbs can be used as main verbs, there is disagreement among Chinese linguists. Some suppose that auxiliary verbs can be treated as predicate verbs [Zhu 1982] while others propose that auxiliary verbs have the same syntactic functions of adverbial modifiers [Hong 1980]. Thus, we propose that auxiliary verbs should be annotated on a case by case basis.

● Auxiliary verb in a VP chunk

In our chunk specification [Li *et al.* 2004], we treat an auxiliary verb as a pre-modifier of an adjoining main verb. See in Example 13, the annotation of MVPs is not affected since the auxiliary verb is chunked with the main verb.

Example 13

[NP 欧盟/j(ou1meng2) 国家/n(guo2jia1)] **[MVP 也/d(ye3) 不/d(bu2) 会/v(hui4) 大力/d(da4li4) 干预/v(gan1yu4)]** 。/ww

(The countries of the European Union will not intervene energetically, either.)

In the above example, the main verb is "干预" (intervene), while the preceding auxiliary verb "会" is treated as a pre-modifier of "干预".

● Auxiliary verb outside a VP chunk

An auxiliary verb can be a single chunk of a VP that is separated from its modifying VP by a following prepositional phrase, noun phrase. Or the auxiliary verb is followed by VP coordination. In this case, we annotate the VP of the auxiliary verb as a MVP. Perhaps some will argue that the main verb can be a verb followed an auxiliary verb. In our annotation scheme, we want to annotate the sentences consistently. For example, in the sentence "[NP 价格/n(jia4ge2)] **[MVP 要/v(yao4)]** [ADJP 低/a(di1)] [MP 一些/m(yi4xie1)] ，/ww" (*The price **is** a little lower.),* there are no other verbs in the sentence, and the verb "要" is a MVP. Thus, there is no need to decide whether the verb "要" is a common verb or an auxiliary verb. From another point of view, if some researchers prefer to treat an auxiliary verb as a

Non-MVP, it is easy to convert our annotation in order to accommodate their specification. Some examples are listed as follows.

***Example* 14**

[NP 国家/n(guo2jia1)] 的/u(de) [NP 事/n(shi4)] **[MVP 要/v(yao4)]** [NP 大家/r(da4jia1)] [VP 关心/v(guan1xin1)] ，/ww

*(The businesses of the country **need** people's attention.)*

Note: In this example，there is a NP instead of a PP following the auxiliary verb "要".

***Example* 15**

[MVP 能够/v(neng2gou4)] [PP 把/p(ba3)] [NP 一般/a(yi4ban1) 号召/vn(hao4zhao1)] [PP 与/p(yu3)] [NP 个别/a(ge4bie2) 指导/vn(zhi3dao3)] [VP 结合/v(jie2he2) 起来/v(qi3lai2)] ，/ww

*([One] **is able to** combine the general calling with an individual guide.)*

***Example* 16**

[MVP 应该/v(ying1gai1)] [ADVP 坚决/ad(jian1jue2)] [VP 反对/v(fan3dui4)] 和/c(he2) [VP 制止/v(zhi4zhi3)] 。/ww

*([One] **should** firmly oppose and prevent [it].)*

Note: In the above sentence, the auxiliary verb "应该" (should) modifies a verb coordination phrase "[VP 反对/v] 和/c [VP 制止/v]" (oppose and prevent).

3) "PP+XP+VP" sequences

In real text, there are a lot of prepositional sequences like "[PP 从/p(cong2, from)] +… + [VP 起步/v(qi3bu4, beginning)]", "[PP 从/p(cong2, from)] + … +[VP 看/v(kan4, watch)]", "[PP 按/p(an1, according to)] + … +[VP 计算/v(ji4suan4, calculate)]", "[PP 以/p(yi3, according to)] + …+[VP 为由/v(wei2you2, excuse)]". We call these sequences PP+XP+VP sequences. One issue to be considered is whether the VP in the sequence is the object of the preposition (PP).

There is a limited number of cases where PP can include the following VP as a part of its object. See Example 17 in [Liu *et al.* 2002]. In this case, we do not annotate the VP as a MVP since the VP acts as the head of a verb phrase, which in turn acts as the object of the PP. The prepositions that can have a verb (or VP) or a clause as their object are also summarized in a

closed set, including “为了”(wei4le, for), “随着”(shui2zhe, with), “关于”(guan1yu2, about) etc.

***Example* 17**

[PP 关于/p(guan1yu2)] [NP 怎么样/r(zen3me1yang4)] [VP 学好/v(xue2hao3)] [NP 汉语/nz(han4yu3)] ，/w [NP 阿里/ns(a1li3)] **[MVP 谈/v(tan2) 了/u(le)]** [ADJP 很/d(hen3) 多/a(duo1)] 。/ww

*(Ali **talked** a lot about how to learn Chinese well.)*

Note: “学好汉语” (learn Chinese well) is a verb phrase in the object of the preposition “关于”. Thus “学好” (learn) should not be tagged as the main verb of the sentence.

However, in most situations, we cannot include the VP in the object of the PP. Nor can the VP be treated as the MVP since it is more likely to be parenthesis[7] in Chinese. See Example 18 below.

***Example* 18**

[PP 按/p(an1)] [NP 可比/vn(ke3bi3) 口径/n(kou3jing4)] **[VP 计算/v(ji4suan4)]** ，/w [TP 去年/t(qu4nian2)] "/w [NP 两/m(liang3) 税/n(shui4)] "/w [MVP 实际/ad(shi2ji4) 完成/v(wan2cheng2)] [MP ４０８３亿/m(yi4) 元/q(yuan2)] ，/ww

(Calculated from constant requirements, “two taxes” actually are collected 408,300 million yuan last year.)

Like the above example, we summerized 14 similar structures like [PP 按/p(an1, according to)]+XP+[VP 计算/v(ji4suan4, calculate)], [PP 从(cong2, from)]+XP+[VP 看/v(kan4, watch)] etc. VPs in these structures are not treated as MVPs.

Otherwise, in a PP+XP+VP sequence, VPs can be viewed as MVPs if those verbs are verbs whose objects can include multiple clauses. See Example 19 in [Liu *et al.* 2002].

[7] Parenthesis is a grammatical phenomenon in Chinese grammar. For example, 据了解, 据介绍, 我看, 我说 are all examples of parenthesis. In our spec, we should not tag a VP like “了解” or “介绍” as a MVP in these parentheses.

***Example* 19**

[PP 从/p(cong2)] [NP 孩子/n(hai2zi1)] [SP 嘴里/s(zui3li3)] **[MVP 知道/v(zhi1dao4)]** ，/w [NP 他/r(ta1)] [NP 姐姐/n(jie3jie3)] [VP 是/v(shi4)] [NP 个/q(ge4) 转业军人/n(zhuan3ye4jun1ren2)] 。/ww

(From the child's mouth , [we] ***know that*** *his elder sister is a former member of the military who has transferred to civilian work.)*

Note: Although the VP "知道" (know) follows the preposition "从" (from), "知道"(know) is a verb whose object can include multiple clauses. Thus, "知道" (know) should be treated as the MVP of the sentence. The following clause "他姐姐是个转业军人" (his elder sister is a former member of the military who has transferred to civilian work.) is the object of "知道" (know).

4) Verb "有"

"有"(have) can be used as a MVP in the following three sentence patterns: a "有-sentences", which has the basic possession sense, e.g., 我有一本书 (I have a book), series-verb sentences, and pivotal sentences [Liu *et al.* 2002]. In most of the above cases, "有" is annotated as the main verb. However, some "有" sentences should not be treated as series-verb or pivotal sentences, nor should "有" be treated as the predicate verb in these sentences. See example 20.

***Example* 20**

[VP 有/v(you3)] [MP 一/m(yi2) 次/q(ci4)] [NP 灵感/n(ling2gan3)] **[MVP 来/v(lai2) 了/y(le)]** ，/ww

(Once upon a time, the inspiration came.)

***Example* 21**

[VP 有/v(you3)] [NP 风险/n(feng1xian3)] [NP 我/r(wo3)] **[VP 来/v(lai2)担/v(dan1)]**。/ww

(I will take the risk.)

Note: This is a sentence with a predicate of a subject-predicate phrase, where the verb-object phrase "有/v 风险/n" (risk) is the subject of the sentence.

5) Verb "是"

Ambiguity is encountered in "是" (is) sentences when verbs are in the subjects of "是". If the VPs are inside the subject of the "是-sentence", we cannot annotate such VPs as MVPs no

matter whether there is punctuation like "，" immediately before "是" or not. See example 22.

***Example* 22**

[NP 买家/n(mai3jia1)] **[VP 不/d(bu2) 怕/v(pa4)]** [NP 赝品/n(yan4pin3)] ，/w **[MVP 也/d(ye3) 是/v(shi4)]** [PP 为了/p(wei4le)] [MP 一个/m(yi2ge4)] [NP "/w 钱/n(qian2) "/w 字/n(zi4)] 。/ww

*(It **is also** for the reason of "money" that the buyer is not afraid of forgeries.)*

Note: Although we find the punctuation "，" before "是", the whole clause, "[NP 买家/n] [VP 不/d 怕/v] [NP 赝品/n]" (the buyer is not afraid of forgeries), acts as the subject of the "是-sentence". Thus, the VP "不/d 怕/v" (is not afraid of) inside it should not be tagged as a MVP.

6) Multiple clauses in a subject or object

We should note that there are many long sentences in texts whose subjects or objects include multiple clauses. These clauses are similar to English ones, and the verbs are nearly a closed set. It includes, for example, "觉得" (feel), "希望" (hope), "认为" (think), and "以为" (suppose) which are listed in our specification. The problem with annotating this kind of sentence stems from the ambiguous subject or object boundaries. See example 23.

***Example* 23**

[NP 张三/nr(zhang1san1)] **[VP 承认/v(cheng2ren4)]** [NP 李四/nr(li3si4)] **[VP 是/v(shi4)]** [MP 一个/m(yi2ge4)] [ADJP 重要/a(zhong4yao4)] 的/u(de) [NP 谈判/vn(tan2pan4) 因素/n(yin1su4)] ，/ww

This sentence has two readings.

1) [VP 承认/v] (admit) is the main verb, and the following clause [NP 李四/ns][VP 是/v]…[NP 谈判/vn 因素/n] (Li is the negotiation factor) is the object of [VP 承认/v]. An English translation of this sentence is "*Zhang admitted that Li is an important negotiation factor.*"
2) [VP 是/v] (is) is the main verb, and the clause [NP 张三/nr] [VP 承认/v] [NP 李四/nr] (Zhang admit Li) is the subject. The English gloss of this sentence is "*[The fact] that Zhang admitted Li is an important negotiation factor*."

Example 23 shows ambiguity with respect to the subject boundary. Example 24 below shows

ambiguity with respect to the object boundary.

***Example* 24a**

[NP 中/j(zhong1)] 、/w [NP 俄/j(e2)] 、/w [NP 法/j(fa3) 等/u(deng3) 国/n(guo2)] [VP 认为/v(ren4wei2)] [VP 可以/v(ke3yi3) 结束/v(jie2shu4)] [PP 对/p(dui4)] [NP 伊拉克/ns(yi1la1ke4)] 的/u(de) [VP 核查/v(he2cha2)] ，/w [NP 美国/ns(mei3guo2)] [VP 则/d(ze2) 坚决/ad(jian1jue2) 反对/v(fan3dui4)] 。/ww

This sentence also has two readings.

1) Both of the clauses following [VP 认为/v] (think) are its objects. In this case, the sentence can be translated as "*Countries such as Chinese, Russia and France thought that the investigation on Iraq could be finished, and [they also thought] that the United States firmly opposed it*."
2) Only the clause immediately following [VP 认为/v] (think) is its object. The next sentence is an independent one. In this case, the sentence can be translated as "*Countries such as Chinese, Russia and France thought that the investigation on Iraq can be finished. [However], the United States firmly opposed it.*"

The two readings of Example 23 are reasonable. But only the reading 2) of Example 24a is reasonable according to the context. However, for a computer, it is hard to make decision here since 1) in Example 24 is also a reasonable parsing candidate if the computer does not have the additional knowledge. For these ambiguities, we apply an annotation scheme similar to that in CTB [Xue and Xia 2000]. If the syntactic ambiguity can be resolved with the knowledge of the context, then we annotate the correct reading. The proposed annotation of Example 23 is based on the context. The proposed annotation of Example 24a is as follows:

***Example* 24b**

[NP 中/j] 、/w [NP 俄/j] 、/w [NP 法/j 等/u 国/n] **[MVP 认为/v]** [VP 可以/v 结束/v] [PP 对/p] [NP 伊拉克/ns] 的/u [VP 核查/v] ，/ww [NP 美国/ns] **[MVP 则/d 坚决/ad 反对/v]** 。/ww

(Countries such as Chinese, Russia and France thought that the investigation on Iraq could be finished. [However], the United States was firmly opposed to [it].)

In the above example, if there is no punctuation immediately after the predicate-verb, the predicate-verb is annotated as a MVP, and the first sentence will end after the punctuation following the first clause. This means that the VP in the first sub-sentence should not be tagged as a MVP at all. The remaining sub-sentences will annotate their predicate-verbs as MVPs and are broken one by one. Also, if some linguists prefer the clause "[NP 美国/ns] [MVP 则/d 坚决/ad 反对/v]"(the United States firmly opposed) as the object of [VP 认为] (think), then they can carry out another task to identify this kind of object since none of the syntactic information of this sentence is lost.

3.2.4 Assignment of Descriptors

Three annotation descriptors are needed: "MVP", "/ww" and "#/ww". The chunk labels are pre-annotated before MVP annotation is performed. The combined label "MVP" indicates the main verb chunk of a sentence. "/ww" and "#/ww" stands for the end of a sentence, where "#/ww" is used to indicate that the sentence lacks of an ending punctuation.

3.3 MVP Statistic

Based on the main verb definition given above, we investigated the distribution of simple sentence types in the annotated PK corpus, which has a total of 100, 417 tokens[8]. The sentences in the corpus were manually annotated with the sentence end tag "/ww" defined above. We got 8, 389 sentences of this kind.

In Figure 3, we show the distribution of three sentence types, that is, sentences with MVPs, sentences without MVPs but with one or more VPs, and sentences without any VPs at all. Sentences with MVPs are given in Examples 7 to 10. Sentences without MVP but with VPs are ones like "[NP 人们/n(ren2men2)] [VP 生活/v(sheng1huo2] [ADJP 很/d(hen3) 苦/a(ku3)] 。"(People's lives are hard). Sentences without VPs are ones like "[NP 劳动/vn(lao2dong4) 经验/n(jing1yan4)] [ADJP 少/a(shao3)] 。" (Work experience is rare).

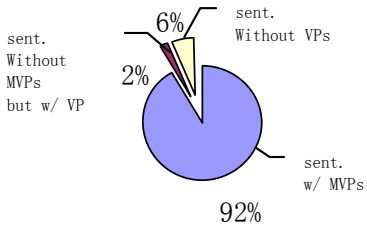

Figure 3. Distribution of sentences w/o MVPs

[8] Here *tokens* include words, punctuation mark in the entire corpus.

From the above figure, we can see sentences with MVPs make up most of the sentences, approximately 92%. This result agrees with Wu's assertion in [Lv *et al*. 1999]. Among these 92% sentences, we find that about 80% of the MVPs are the first VPs in the sentences.

Figure 4 shows the distribution of the remaining 8% of the sentences, totally 671 sentences without any MVPs. The non-predicate sentences are sentences like [NP 照片/n 人物/n] 的/u [NP 故事/n] #/ww. (The story of the people in pictures). These sentences come from the titles of texts or headlines of news reports.

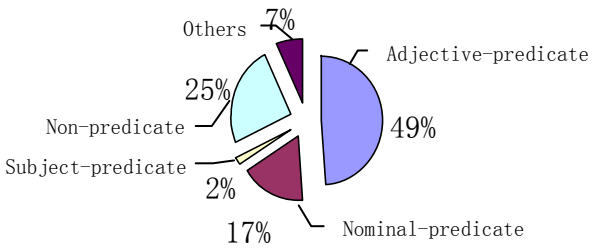

Figure 4. Distribution of sentences without MVPs

Since the MVP sentences amount for most of the sentences (i.e., 92% of all the sentences in the PK corpus), our study focused on identifying the verb predicates in the sentences. We will explore them in more detail below.

3.4 A Model for Chinese Main Verb Identification

Our aim is to conduct main verb identification on a binary classifier. For each VP, we determine whether it is an MVP or not. The Support Vector Machine (SVM) is one of the most successful binary classifiers. This method has been used in many domains of NLP, such as part-of-speech tagging [Nakagawa *et al.* 2001], Name Entity recognition [Isozaki and Kazawa 2002], Chunking [Li *et al.* 2004] and Text categorization [Joachims 2002]. To our knowledge, the use of SVM to identify Chinese main verbs has not been studied previously. Moreover, there are indications that the differences among various learning techniques tend to get smaller as the size of the training corpus increases [Banko and Brill 2001].

We follow the definition of SVM in [Vapnik 1995]. Suppose the training samples are $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where each x_i $(1 \le i \le N)$ represents an input vector defined on an n-dimensional space, and each dimension is a feature we define in the following sections. $y_i \in \{1, -1\}(1 \le i \le N)$ indicates whether it is MVP or not. The separating hyperplane is defined by

$$\vec{w} \cdot \vec{x} + b = 0 \quad \vec{w} \in R^n, b \in R .$$

SVM searches for the hyperplane that separates a set of training examples that contain two distinct classes with the maximum margin. We use SVM^{light} [Joachims 1999] as our implementation tool.

A processing cycle can arise here. Because most of the related works are based main verb identification in sentences with pre-determined sentence boundaries, sentence boundary labeling must be done before tagging. But if sentence boundary labeling is done before tagging, where does the predicate information come from? So instead of doing sentence boundary labeling beforehand, we first detect the predicate without using sentence boundary information. It is for this reason that we want to break the sentence into simple sentences that by definition require main verbs. This procedure is similar to the work in [Chen and Shi 1997]. Firstly, we break the sentence into process units. They are word sequence separated by punctuation marks, such as "。！？ ，", but we do not know if they are sentence ending labels or not. Secondly, our algorithm determines whether the VPs are MVPs in these units. If the value is negative, the VP is not a MVP and vice versa. Finally, if more than two MVPs are identified in a processing unit, we rank these MVPs according to the classifier's output (value of the decision function) and choose the one with the highest rank as the MVP. The chunk information is obtained from our chunking system.

Building an effective SVM classifier involves choosing good features. We break up the features used in our research into two categories, local and contextual. The first set of features is derived from the surface information of VPs. Since these features are based on chunks themselves, they are called local features. The second set of features is derived from the context information of VPs, while also incorporating some lexical knowledge and patterns. Thus, we call these features contextual features. Our model is based on the level of chunking because our experiments show that this is better than basing the model on parts of speech.

In the following sections, we will describe the feature set in detail.

3.4.1 Local Features

Local features are explored based on careful observation of the training corpus. All of them are new features we have proposed. Although they are simple, they work well in MVI since they represent the characteristics of the VPs themselves. Our model captures three local features: 1) the VP position, 2) the VP length 3) and the probability of head verbs being MVPs. Here, VP position and VP length are feature groups. Each feature group is made up of several binary features. This means for each VP, if one feature in the group is set to 1, other features in the same feature group are set to zero.

VP position is a feature group. Totally, there are six binary features in this group. This means that the phrasal position number of a VP appears in the process unit, which starts with 1. For example, if the VP is the first VP in the process unit, the value of the first feature is 1, and the other feature values are set to zero. If the position value of the VP in the process unit is larger than 5, then the value of the sixth feature is 1, and the other features are set to zero. Figure 5 shows the VP position distribution.

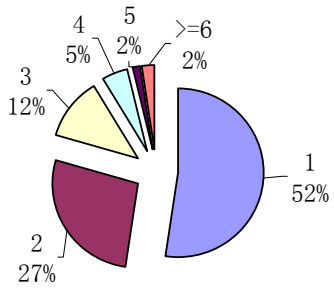

Figure 5. VP position distributions

In the figure, we show the distribution of up to six binary features because the percentage of VPs with position values of 6 or less is 98%.

Also based on our statistics for the training data, about 80% of the MVPs are the first VPs in the sentences. So we use this feature as the base-line feature (refer to section 4.2).

VP length is also a feature group. Totally there are six binary features, chosen based on our intuition that the longer a VP is, the more likely it is a MVP since it has more modifiers. VP length is measured in terms of the number of words in a VP. Thus, the i^{th} feature in the feature group stands for a VP with a length of i (starting from 1). The sixth feature means the VP has a length larger than 5. See the VP length distribution shown in Figure 6.

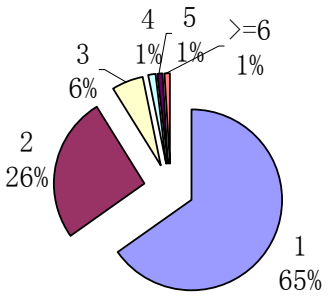

Figure 6. VP length distribution

In the above figure, we show the distribution of up to six binary features since the percentage of VPs with lengths smaller than 6 is about 99%. For example, the VP "坚决/ad 不/d 收/v" has a length of three. See in table 2, the third feature is set to 1, and other feature values are set to zero.

Table 2. VP length feature table

Feature Number	*1*	*2*	*3*	*4*	*5*	*6*
Feature Value	0	0	1	0	0	0

Probability of head verbs being MVPs is a real value feature. Our statistics show that some VPs are MVPs, like "是 (is)" and "认为 (think)". This feature is estimated beforehand as follows based on the training corpus.

$$\text{MVP\_P(x)}=\frac{\text{C(x in MCVP)}}{\text{C(x)}},$$

where C(x in MVP) is the number of occurrences of verb x as a MVP and C(x) is the total occurrences of verb x in the training data.

3.4.2 Contextual Features

So far, we have introduced features that are based on the characteristics of a VP itself. One problem with these features is that they only use the surface information of a VP, not its contextual information. Related works [Sui and Yu 1998b; Gong *et al.* 2003] have shown contextual features are helpful in MVI. Thus, we also incorporate contextual features into our model. One difference is that in our work, we added new features included in our specification, such as "PP+XP+VP" sequences, into our model. In addition, we integrate them into our SVM model instead of dealing with this problem in two steps as in [Gong *et al.* 2003].

Pattern features are one type of binary feature. The patterns we define include features like "的" (de) and "得" (de) that were also used in [Gong *et al.* 2003; Sui and Yu 1998b]. One difference is that we only consider "的" when it is next to a VP. In addition, we find that in about 92% of cases, the verb "是" followed by "的" is used as a MVP, so we treat this word differently from other verbs. These pattern features are very precise based on our statistics on training corpus.

Table 3. Pattern feature table

Pattern Features	VP+SP
	VP+NO_CHUNK_UNIT
	PP+VP
	"《" + VP + "》"
	"的" + VP
	"、" + VP

In Table 3, VP+SP means a SP chunk followed a VP chunk. NO_CHUNK_UNIT indicates the out-of-chunk units as defined in our chunk system, including "等" (etc.), "之/r" (zhi), "的"(de) etc. These pattern features indicate the contexts of MVPs. They share the same formulation shown below:

$$f^i(VP) = \begin{cases} 1, \text{if VP corresponds to a defined pattern} \\ 0, \text{otherwise} \end{cases}.$$

Anti-features include words and patterns in which VPs can not be used as MVPs. We define an anti-feature as a binary feature. If a VP meets this requirement, the f(VP) = 1; otherwise, f(VP) = 0. If a VP appears in an anti-pattern, it will be masked, and other features will not be added. Anti-features are mostly derived through our careful observation of the specification.

1) Lexical anti-feature to exclude MVP

As described in our specification, "据了解" (ju4liao2jie3, it is reported), "据介绍"(ju4jie2shao4, it is introduced), "我看" (wo3kan4, I see), "我说" (wo3shuo1, I say) are all examples of parenthesis. VPs like "了解" (report)" and "介绍" (introduce) in such contexts are not used as MVPs. In addition, based on the statistic of the training data, some words are typically not used as MVPs, like "新年伊始" (xin1nian2yi1shi3, the beginning of New Year), "解放思想" (jie3fang4si1xiang3, emancipate the mind), etc. Lexical anti-features of the above two types are set to 1. This kind of information is stored in a list of words explored in our specification.

2) Frame anti-feature to exclude MVP

The VPs in frame-like structures like [PP 在/p(at)]+...+[SP 上/s(above)] are not MVPs. Because of the right boundary of such a long prepositional phrase is hard to identify and to avoid ambiguities, our chunk system only finds the PP chunks of frame-like construction with explicit boundaries and length constraints, such as [PP 在/p (at) … 中/f (middle)] [Li *et al.* 2004]. We have to detect Non-MVPs in longer prepositional phrases. Statistics show that based on the current PP chunk tags, some right boundaries of longer prepositional phrases can be recognized.

Take the PP "在"(zai4) as an example. We collect all the SP chunks as its right boundary candidates in the training data. Among the resulting 111 SP chunks, only 1 SP chunk is not a right boundary. So the pattern [PP 在/p] + SP is very precise to form a longer PP. From another point of view, we use such kind of patterns to perform a rough PP boundary recognition. For example, [PP 当/p(dang1)] [NP 他们/r(ta1men2)] [VP 来到/v(lai2dao4)] [NP 另/r(ling4) 一个/m(yi1ge4) 风景点/n(feng1jing3dian3] [VP 要/v(yao4) 拍照/v(pai1zhao4)] [SP 时/Ng(shi2)] ，/w (When they come to the spots of interest to take photos) is a PP, and [VP 来到/v] (come) and [VP 要/v 拍照/v] (to take photos) are masked as Non-MVPs. Table 4 lists three types of anti-patterns. It should be noted that the frame structures are not limited to PPs. These structures are selected from the statistics of the training corpus.

Table 4. Frame anti-feature types

Frame Anti-pattern type	Examples
Only chunk type	[PP 在/p] + [SP */*] ;
	[VP 有/v] + [MP */*]
Only lexical type	[PP 当/p] + [SP 之际/Ng]
Both chunk and lexical types	[PP 将/p] + [NOCHUNK_WORD 的/u] [NP */*];
	[VP 找到/v] + [NP */*] [SP 时/Ng]

Here, the first chunk is the trigger chunk. That is, if we encounter such a chunk, we trigger the pattern matching module, and all the VP chunks are blinded. That is, we set f(VP) = 1 if the chunks match one of the patterns from MVP identification. See the example above where feature values of "来到" (lai2dao4, come) and "要/v 拍照/v" (yao4 pai1zhao4, to take photos) are both set to 1. Totally, we have 62 patterns. Among them, 52 patterns have PP trigger chunks. Ten patterns have VP trigger chunks. Similar patterns can be designed according to "有" sentences or "是" sentences in our specification.

In our implementation, we have a module that we use to convert the corpus into the proper input format for SVM^{light} [Joachims 1999]. Each of the above features corresponds to one dimension of the feature vector. In the next section, we will discuss the evaluation results.

4. Experiments

We evaluated our MVI approach using manually annotated data, which was a subset of the PK corpus. The PK corpus was released by the Institute of Computational Linguistics, Peking University. The corpus contains one month's data from the *People's Daily* (January 1998). This corpus has already been segmented and part-of-speech tagged. Its specification has been published in [Yu *et al.* 2002]. Totally, there are about 40 part-of-speech tags including noun (/n), verb(/v), adjective(/a), name entity tags, verbal noun (/vn) etc. The details of our MVI training and testing corpus are shown in Table 5.

Table 5. Training and testing corpus

Data	# of Chunks	# of Tokens	# of Simple Sent.	# of Whole Sent.	Ave. Simple Sent. Length	Ave. Whole Sent. Length
Train	72, 645	100, 417	8, 389	3, 784	11.97	26.54
Test	19, 468	26, 334	2, 456	1, 047	10.72	25.15

Here, *tokens* include words and punctuation marks in the entire corpus. *Chunk* marks are annotated according to our chunk spec [Li *et al.* 2004] with 11 chunk tags. *Simple sentences* are annotated as described in our spec. *Whole Sentences* are sentences that use "。！？" as sentence endings. As the table shows, following annotation of simple sentences, the average length of a simple sentence was less than two times the length of a whole sentence. Notice that we do not use sentence-ending label information in our algorithm.

4.1 Evaluation Metrics

The evaluation metrics used here are the traditional Precision, Recall and F Measure:

$$\text{Precision (P)} = \frac{\#\text{ of Correct MVPs in system output}}{\#\text{ of Total MVPs in system output}};$$

$$\text{Recall (R)} = \frac{\#\text{ of Correct MVPs in system output}}{\#\text{ of Total MVPs in answer}};$$

$$F = 2 * P * R/(P + R).$$

We also compared our evaluation metrics with the Sentence Accuracy Rate (SAR) used in related works:

$$\text{SAR} = \frac{\#\text{ of correct tagged verb - predicate sentences}}{\#\text{ of total verb - predicate sentences}}.$$

We propose using the Precision/Recall evaluation metrics for three reasons: Firstly, these evaluation metrics are more widely used than a single percent-correct score. Secondly, we don't deal with sentences whose predicates are adjective phrases or noun phrases. So if we include these sentences into the total number of sentences in our calculations, the performance will suffer and the result will not reflect the performance of identifying MVPs. Thirdly, in our approach, we do not discriminate verb-predicate sentences with other sentences. However, in order to show the soundness of our technical approach, we also provide the SAR, and we manually calculate the number of verb-predicate sentences.

4.2 Impact of Different Features on MVP Identification Results

We investigated the contributions of different features as shown in Table 6.

Table 6. Impact of different features

Model	**Precision**	**Recall**	**F-Measure**	**SAR**
Baseline (*VP Position*)	76.7	87.5	81.74	78.1
Baseline (*VP Position*) + **Other Local Features** (*VP Length; Probability of head verbs being MVPs*)	82.89	88.8	85.74	80
Baseline (*VP Position*) + **Other Local Features** (*VP Length; Probability of head verbs being MVPs*) + **One Contextual Feature** (*Pattern Features*)	90.23	89.66	89.94	85.1
Baseline (*VP Position*) + **Other Local Features** (*VP Length; Probability of head verbs being MVPs*) + **Contextual Features** (*Pattern Features; Anti-features*)	93.6	92.1	**92.8**	**88.6**

1) Local features improved the performance by 4%. One of the problems with local features is data sparsity, because the real value feature, that is, "Probability of head verbs being MVPs" is estimated from the whole training data. There are occasions when verbs in the testing corpus have not been encountered before. Thus, we will investigate the use of smoothing technology in future research.

2) Pattern features of the contextual type are very useful for MV identification and here increased the performance further by 4.2% from 85.74% to 89.94%. The lexicalized contextual features like "的", punctuation like "《》、" really helps to improve the performance.

3) Anti-features also contributed about 3% to the performance based on a comparison of the results obtained with and without anti-features. The reason is that anti-features can exclude VPs that have no chance of being MVPs.

We also provide the SAR results in the table. However, they are not comparable essentially because of different test data and amounts of data used in other works. The above results show that the SVM provides a flexible statistical framework for incorporating a wide variety of knowledge, including local and contextual features, for MVP identification.

After we tested the impact of different features on the performance of MVP identification, we wanted to know whether our annotated corpus was large enough to achieve acceptable performance. We used the best feature set according to the results of the above experiment. The performance achieved showed that the current training size had almost reached the saturation point.

4.3 Impact of Chunk Information on MVP Identification

Since we annotate MVPs based on chunk levels, we wanted to know how this shallow syntactic information affected the MVP identification. So we devised the following two experiments.

1. We firstly tested the performance of main verb identification based on POS, which does not include any shallow syntactic information. We stripped all the chunk tags in the corpus and used a simple rule to tag the predicate verb based on MVP chunks. That is, the headword of MVP chunk is the main verb of the sentence. For example, a) sentence is mapped to b) in the following.

a) [NP 公园/n(gong1yuan2)] **[MVP 时时/d(shi2shi2) 梦想/v(meng4xiang3) 着/u(zhe)]** [VP 有/v(you3)] [NP 条件/n(tiao2jian4)] [VP 繁育/v(fan2yu4) 出/v(chu1)] [NP 小√a(xiao3) 虎/n(hu3)] 。

b) 公园/n 时时/d 梦想/v_$ 着/u 有/v 条件/n 繁育/v 出/v 小√a 虎/n 。

(People in the park have always dreamed that it is possible to breed tigers)

Here /v_$ denote "梦想" (dream) is the main verb of sentence because it is the head verb of [MVP 时时/d 梦想/v_$ 着/u] (always dream of).

In this way, from the MVP training and testing corpus, we got the training and testing corpus with main verbs tagged. In our algorithm, we use the correct part-of-speech tags as input for main verb identification.

When we identify main verbs based on part-of-speech tags, all the features except for VP length are mapped to verb features. For example, the feature "VP position" is mapped to "verb position" an so on. The feature "Probability of verbs being MVPs" is revised to obtain the following formulation:

$$\mathrm{MV\_P}(\mathrm{X}) = \frac{\mathrm{C(x\ is\ main\ verb)}}{\mathrm{C}(\mathrm{x})},$$

where C(x is the main verb) is the number of occurrences of the verb x as the main verb, and C(x) is the total occurrences of verb x in the training data.

"Only lexical type" among the frame anti-pattern features shown in Table 4 is also modified to use part-of-speech tags without chunk tags. The others are not modified since they have general chunk information, such as [SP */*], which cannot be directly converted to POS.

2. We also stripped all the chunk tags in the corpus, but this time we used our chunk system [Li *et al.* 2004] to re-chunk the data based on part-of-speech tags. Our chunk system is built on HMM. TBL-based error correction is used to further improve chunking performance. The average chunk length was found to be about 1.38 tokens and the F measure of chunking reached 91.13%. Inevitably, our chunk system will incur errors. Based on this noisy data, we use the same feature set to identify the MVPs. In this experiment, we wanted to know how the chunk errors would affect the MVP identification results.

The experimental results for the above two cases are shown in the Table 7.

Table 7. Impact of chunk information

Model	Precision	Recall	F-Measure
POS	84.98	84.04	84.5
POS+Chunk1	88.56	89.27	88.9
POS+Chunk2	93.6	92.1	92.8

The *POS* model row shows the first set of experiment results discussed above, that is, the results of identifying main verbs without using any chunk information. The *POS+Chunk1* model row shows the second set of experimental results: identifying MVPs with noisy chunk

information. The *POS+Chunk2* model row shows the results of identifying MVP with correct chunk information.

As the above table shows, the model trained on part-of-speech tags had the worst performance. This is because that the model lacks both chunk length information and part of the frame anti-pattern information. For example, if VP chunk is available, the VP chunk length can be calculated. Thus, the observation that VPs are precise (more than 85%) to be MVPs when their lengths are longer than 4 can improve the performance of main verb identification. Further more, the model trained on part-of-speech tags tends to tag the first verb as the main verb.

We performed the error analysis on results of POS+Chunk1 model. We wanted to see how many errors resulted in chunking errors in the table 8 below.

Table 8. Error types for the POS+Chunk1 Model

Error Types	**Total Number**	**Caused By Chunk Errors**	**Caused By MVP Tag Errors**
Miss	242	80 (33.1%)	162 (66.9%)
False	260	88 (33.8%)	172 (66.2%)
All	502	168(33.5%)	334 (66.5%)

In the table, Chunk Error means errors are caused by the chunk output, such as over-combining, under-combining etc. MVP Tag Error means we have correct chunks but MVP tagging is incorrect. It can be seen that more than one-third of the errors are caused by the chunk errors. The POS+Chunk2 model had the best performance since it uses shallow syntactic information and no errors appear in chunks.

5. Error Analyses and Discussion

The errors appearing in test data fall into the following categories.

5.1 Ambiguity of VP in Subject

Disambiguating VPs in subjects and predicates is a difficult problem. Since main verb identification is not based on syntactic and semantic parsing, we can only find the surface features of sentences. Thus, while the current algorithm correctly handles Example 25 and Example 27, it fails to handle Example 26 and Example 28.

Example 25 can be handled because the VP length feature helps. However, in some cases, the VP length will not help. Example 26 is a typical sentence in which the MVP should be “提醒” (ti2xing3, remind). The whole phrase “[SP 街上/s(jie1shang4)]…[NP 爆竹声/n(bao4zhu2sheng1]” (The sound of firecrackers …in the street) acts as the subject of “提醒”

(remind). The double objects of the main verb are "我" (I) and "[TP １日/t] [VP 是/v] [TP 新年/t]"(January the first is a new year's day). Both of the VPs in the subject are longer than the VP "提醒" (remind). Although we can exclude the second VP as the MVP (the pattern feature "VP+的" helps), it is rather difficult to exclude the first VP simply based on surface information. What leads to more ambiguity is "是" (is) in the object which also has a large probability of being a MVP. From the above analysis, it is currently difficult for our algorithm to detect "提醒" (remind) as a MVP.

***Example* 25**

Correct:

[VP 没有/v(mei3you3)] [NP 这/r(zhe4) 点/q(dian3) 精神/n(jing1shen2)] **[MVP 就/d(jiu4) 不/d(bu2) 配/v(pei4)]** [NP 电力/n(dian4li4) 人/n(ren2)] [NP 这/r(zhe4)] [ADJP 光荣/a(guang1rong2)] 的/u(de) [NP 称号/n(cheng1hao4)] 。/ww

*(Without that spirit, you will **not deserve to** have the glorious title "electronic people".)*

This example can be handled because the VP length feature helps.

***Example* 26**

Correct:

[SP 街上/s(jie1shang4)] **[VP 不时/d(bu4shi2) 地/u(de) 响起/v(xiang3qi3)]** [MP 一阵阵/m(yi2zhen4zhen4)] [PP 在/p(zai4)] [NP 北京/ns(bei3jing1)] **[VP 已/d(yi3) 听/v(ting1) 不/d(bu4) 到/v(dao4)]** 的/u(de) [NP 爆竹声/n(bao4zhu2sheng1] **[MVP 提醒/v(ti2xing3)]** [NP 我/r(wo3)] [TP １日/t(ri4)] **[VP 是/v(shi4)]** [TP 新年/t(xin1nian2)] 。/ww

*(The sound of the firecracker in the street every now and then, which haven't been heard already in Beijing, **remind** me that January the first is a new year's day.)*

System Output:

[SP 街上/s] **[MVP 不时/d 地/u 响起/v]** [MP 一阵阵/m] [PP 在/p] [NP 北京/ns] **[VP 已/d 听/v 不/d 到/v]** 的/u [NP 爆竹声/n] **[VP 提醒/v]** [NP 我/r] [TP １日/t] **[VP 是/v]** [TP 新年/t] 。/ww

In Example 27 and Example 28, since there is not enough information to determine that "是" is not in the object of "承认", the algorithm fails to find that "是" is the main verb.

***Example* 27**

Correct:

[NP 各方/r(ge4fang1)] **[MVP 承认/v(cheng2ren4)]** [NP 波黑/ns(bo1hei1)] **[VP 是/v(shi4)]** [MP 一个/m(yi1ge4)] [ADJP 统一/a(tong3yi1)] 的/u(de) [NP 主权/n(zhu3quan2) 国家/n(guo2jia1)] ，/ww

*(Each side **admits** that Bosnia-Herzegovena is a unified, sovereign country.)*

This example can be handled because the VP position helps.

***Example* 28**

Correct:

[VP 承认/v(cheng2ren4)] [NP 错误/n(cuo4wu4)] **[MVP 是/v(shi4)]** [MP 一/m(yi1) 种/q(zhong3)] [NP 好/a(hao3) 习惯/n(xi2guan4)] 。/ww

(It is a kind of good habit to be able to acknowledge making mistakes.)

System Output:

[MVP 承认/v] [NP 错误/n] **[VP 是/v]** [MP 一/m 种/q] [NP 好/a 习惯/n] 。/ww

5.2 Long Adjective Modifier

In Chinese parsing, the left boundary of "的" is a typical ambiguity problem. This problem also arises in main verb identification. The algorithm falsely identifies VPs in adjective modifiers as MVPs. See the following examples.

Example 29

Correct:

[VP 积淀/v(ji1dian4)] [PP 在/p(zai4) 大众/n(da4zong4) 血液/n(xue4ye4) 中/f(zhong1)] 的/u(de) [NP 传统/n(chuan2tong3) 文化/n(wen2hua4) 基因/n(ji1yin1)] [ADVP 也/d(ye3)] [PP 在/p(zai4) 传承/v(chuan2cheng2) 中/f(zhong1)] **[MVP 发生/v(fa1sheng1)]** [NP 种种/q(zhong3zhong3) 变异/n(bian4yi4)] 。/ww

*(The genes of the traditional culture which have been settling in the blood of the masses **undergo** various mutations when passing on.)*

System Output:

[MVP 积淀/v] [PP 在/p 大众/n 血液/n 中/f] 的/u [NP 传统/n 文化/n 基因/n] [ADVP 也/d] [PP 在/p 传承/v 中/f] **[VP 发生/v]** [NP 种种/q 变异/n] 。/ww

Note: The main verb of the whole sentence should be "发生". The verb phrase "[VP 积淀/v] [PP 在/p 大众/n 血液/n 中/f]" acts as a pre-modifier of the head noun "[NP 传统/n 文化/n 基因/n]". Thus, "积淀" should not be identified as a MVP in the whole sentence.

***Example* 30**

Correct:

[PP 于/p(yu2)] [TP 7月/t(qi1yue4) 5日/t(wu3ri4)] **[MVP 作出/v(zuo4chu1)]** **[VP 确定/v(que4ding4)]** [NP 肇事人/n(zhao4shi4ren2)] [NP 张/nr(zhang1) 成聚/nr(cheng2ju4)] **[VP 负/v(fu4)]** [NP 事故/n(shi4gu4) 全部/m(quan2bu4) 责任/n(ze2ren4)] ，/w [NP 受害人/n(shou4hai4ren2)] [NP 张/nr(zhang1) 平/nr(ping2)] **[VP 不/d(bu2) 负/v(fu4)]** [NP 责任/n(ze2ren4)] 的/u(de) [NP 交通/n(jiao1tong1) 事故/n(shi4gu4) 责任/n(ze2ren4) 认定书/n(ren4ding4shu1)] 。/ww

(On July 5th, the officer wrote the Traffic Accident Responsibility Assertion Book, in which the traffic troublemaker, Zhang Chenju, takes all the responsibility while the victim, Zhangpin is not responsible.)

System Output:

[PP 于/p] [TP 7月/t 5日/t] **[MVP 作出/v]** **[VP 确定/v]** [NP 肇事人/n] [NP 张/nr 成聚/nr] **[VP 负/v]** [NP 事故/n 全部/m 责任/n] ，/ww [NP 受害人/n] [NP 张/nr 平/nr] **[MVP 不/d 负/v]** [NP 责任/n] 的/u [NP 交通/n 事故/n 责任/n 认定书/n] 。/ww

Note: The main verb of the whole sentence is "作出". The adjective modifier of NP "交通/n 事故/n 责任/n 认定书/n" (Traffic Accident Responsibility Assertion Book) consists of two sub-sentences, in which the VPs "负" (take) and "不负" (not take) act as main verbs. Thus, "负" (take) and "不负" (not

take) should not be annotated as MVPs in the whole sentence.

6. Application

Labeling sentence boundaries is a prerequisite for many natural language processing tasks, including information extraction, machine translation etc. However, as Yu and Zhu [2002] pointed out, the problem is that "We have discussed a lot of word segmentation problems. But limited work has been done on Chinese sentence segmentation and it is still a difficult problem for computers." Without predicate information, it is difficult to predict sentence boundaries. Thus, we first identify the main verb and then label the sentence boundaries. The tagged results of simple sentence boundary labeling are like the following examples.

***Example* 31**

[NP 母爱/n(mu3ai4)] , /w [MVP 作为/v(zuo4wei2)] [NP 人类/n(ren2lei4)] [MP 一/m(yi1) 种/q(zhong3)] [ADJP 崇高/a(chong2gao1)] 的/u(de) [NP 爱/vn(ai4)] ，/ww [MVP 是/v(shi4)] [MP 一/m(yi4) 棵/q(ke1)] [NP 人类/n(ren2lei4) 精神/n(jing1shen2) 大树/n(da4shu4)] ，/ww [NP 她/r(ta1)] [MVP 永久/d(yong3yuan3) 地/u(de) 枝繁叶茂/i(zhi1fan2ye4mao4)] 。/ww

(Mother's love is a kind of lofty love of mankind. It is a big tree of the human spirit. It will have a permanent foundation with luxuriant foliage and spreading branches.)

There are three simple sentences in this example. Our task is to use MVP information to break the sentence up into simple sentences. Here, when we refer to a "sentence," we mean a verb-predicate sentence. Since there are no MVPs in non- verb-predicate sentence, we cannot use the MVP information to break up these sentences.

We compared two sentence-breaking models. First, in the base line, we tagged all the commas as sentence ending punctuation if the sentences had at least one VP. Second, we tagged all the commas as sentence ending punctuation if the sentences had MVPs. This was an end-to-end evaluation because MVP identification was used as preprocessing step before sentence breaking.

The evaluation metrics we used were as follows:

$$\text{Precision} = \frac{\#\text{ of correct sentence} - \text{stoppunc.in system output}}{\#\text{ of sentence} - \text{stoppunc.in system output}};$$

$$\text{Recall} = \frac{\#\text{ of correct sentence} - \text{stoppunc.in system output}}{\#\text{ of sentence} - \text{stoppunc.in answer}};$$

$$F = 2 * P * R/(P + R).$$

Table 9. Performance of Chinese Sentence Breaker

Model	**Precision**	**Recall**	**F**
Baseline	86.74	94.57	90.48
Tag with MVP	94.22	91.34	92.76

From the above table, one can see that the simple sentence breaker improved the performance by about 2.4% with the help of MVP identification. Errors in the tagging of stop-punctuation were mostly caused by errors in the tagging of MVPs.

7. Conclusion and Future Work

Main verbs are useful for dependency parsing, sentence pattern identification, and Chinese sentence breaking. Chinese linguists have done research on predicate-verbs for a long time and provided a grammatical view of analyzing the Chinese sentence. However, automatically identifying main verbs is quite another problem. Most of the previous works by computational linguists have focused on the identification process instead of the definition of a main verb. In this paper, we have discussed in detail the whole process of automatically identifying Chinese main verbs from specification to realization.

The contributions of our work are as follows.

1) We have thoroughly investigated main verbs from both linguists' point of view and the computational point of view. Based on this investigation, we have presented our specification as well as a corpus annotation method. The advantage of our specification is that the main verbs of different verb-predicate sentences are included. More specific and reliable knowledge is applied in our main verb definition. Various complicated cases have been studied, and abundant examples from real text have been provided.
2) We have presented our results of identifying main verbs based on chunking levels. The experimental results show that the performance of our approach is better than that of the

approach based on part-of-speech tags. We have also proposed an end-to-end evaluation based on the use of a Chinese simple sentence breaker.

3) New local and contextual features investigated in our specification and statistics have been incorporated into our identification algorithm and used to achieve promising results.

In future work, we would like to find more effective features from lexical knowledge and solve the data sparse problem that is encountered in feature selection. We also are interested in developing more applications based on MVP information, such as an application for extracting the verb-subject or verb-object dependency relations.

Acknowledgements

We would like to thank Prof. Jianyun Nie, Dr. John Chen, Dr. Ming Zhou, and Dr. Jianfeng Gao for their valuable suggestions and for checking the English in this paper. We thank the anonymous reviewers of this article for their valuable comments and criticisms. We would also like to thank Juan Lin for her help with our specification design and corpus annotation.

References

Banko, M. and E. Brill, "Scaling to very very large corpora for natural language disambiguation," *Proceedings of the 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics,* 2001, pp. 26-33.

Chen, X.H. and D.Y. Shi, "To Mark Topic and Subject in Chinese Sentences," *Proceedings of 4th National Computational Linguistics,* Tsinghua University Press, 1997m, pp. 102-108.

Ding, S.S., S.X. Lv, R. Chen, D.X. Sun, X.C. Guan, J. Fu, S.Z. Huang and Z.W. Chen, "xiandai hanyu yufa jianghua," (Modern Chinese Grammar Talk), The Commercial Press, 1961.

Fan, X., "sange pingmian de yufa guan," (The Grammar View of Three Levels), Publishing house of Beijing Language Institute, 1995.

Gong, X.J., Z.S. Luo and W.H. Luo, "Recognizing the Predicate Head of Chinese Sentences," *Journal of Chinese Information Processing,* vol. 17, no. 2, 2003, pp. 7-13.

Hong, X.H., "hanyu cifa jufa chanyao," (The Brief Introduction to Chinese Word and Syntax), Jilin People's Press, 1980.

Huang, Z.K., "xiandai hanyu changyong jushi," (The Daily Sentence Pattern of Modern Chinese), People's Education Publishing House, 1987.

Isozaki, H. and H. Kazawa, "Efficient Support Vector Classifiers for Named Entity Recognition," P*roceedings of the 17th International Conference on Computational Linguistics,* Taipei, Taiwan, 2002, pp. 390-396.

Jin, L.X. and S.Z Bai, "The Characteristics of Modern Chinese Grammar and the Research Standards," *Chinese Language Learning,* no. 5, Oct, 2003, pp. 15-21.

Joachims, T., "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning," Schölkopf B., C. Burges and A. Smola (editor), MIT-Press, 1999, pp.169.

Lai, T. B. Y. and C.N. Huang, "Dependency-based Syntactic Analysis of Chinese and Annotation of Parsed Corpus," *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics,* Hong Kong, China, 1-8 October 2000, pp. 255-262.

Li, H.Q., C.N. Huang, J.F. Gao and X.Z. Fan, "Chinese Chunking with Another Type of Spec," *The Third SIGHAN Workshop on Chinese Language Processing,* Barcelona, July 24-26, 2004, pp. 41-48.

Li, H.Q., C.N. Huang, J.F. Gao and X.Z Fan, "The use of SVM for Chinese new word identification," *The First International Joint Conference on Natural Language Processing.* March 22-24, 2004, pp. 497-504.

Liu, Y.H., W.Y. Pan and W. Gu, "xiandai hanyu shiyong yufa," (Practical Chinese Grammar Book), The Commercial Press. 2002.

Liu, Q. and S.W. Yu, "Discussion on the Difficulties of Chinese-English Machine Translation," *International Conference on Chinese Information Processing,* January 1998, pp.507-514.

Luo, Z.S., C.J. Sun and C. Sun, "An Approach to the Recognition of Predicates in the Automatic Analysis of Chinese Sentence Patterns," *Proceedings of 3th National Computational Linguistics,* 1995, pp. 159-164.

Lv, S.X., "xiandai hanyu babaici," (Eight Hundred Words of Modern Chinese) The Commercial Press, 1980.

Lv, S.X. and Q.Z. Ma (editor), "yufa yanjiu rumen," (Elementary Study of Chinese Grammar). The Commercial Press. 1999.

Marcus, M., B. Santorini and M. Marcinkiewicz, "Building a Large Annotated Corpus of English: the Penn Treebank," *Computational Linguistics*, 19(2), 1993, pp. 313-330.

Meng, C., H.D. Zheng, Q.H. Meng and W.L. Cai, "hanyu dongci yongfa cidian," (The Chinese Verb Usage Dictionary). The Commercial Press. 2003.

Nakagawa, T., T. kudoh and Y. Matsumoto, "Unknown Word Guessing and Part-of-speech Tagging Using Support Vector Machines," P*roceedings of the 6th NLPRS,* 2001, pp. 325-331.

Sui, Z.F. and S.W. Yu, "The Research on Recognizing the Predicate Head of a Chinese Simple Sentence in EBMT," *Journal of Chinese Information Processing,* vol.12, no. 4, 1998a, pp. 39-46.

Sui, Z.F. and S.W. Yu, "The Acquisition and Application of the Knowledge for Recognizing the Predicate Head of a Chinese Simple Sentence," *Journal (Natural Sciences) Of Peking University,* vol. 34, no. 223, 1998b, pp. 221-230.

Vapnik, V. N., "The nature of Statistical Learning Theory," Springer, 1995.

Xia, F., M. Palmer, N.W. Xue, M.E. Okurowski, J. Kovarik, F.D. Chiou, S.Z. Huang, T. Kroch and M. Marcus, "Developing Guidelines and Ensuring Consistency for Chinese Text Annotation," *Proceedings of the second International Conference on Language Resources and Evaluation (LREC 2000),* Athens, Greece, 2000, pp. 3-10.

Xue, N.W. and F. Xia, "The Bracketing Guidelines for the Penn Chinese Treebank," http://www.cis.upenn.edu/~chinese/parseguide.3rd.ch.pdf. (3.0). Oct. 2000.

Yu, S.W. and X.F. Zhu, "Chinese Information Processing and Its Methodology," Applied Linguistics, May 2002, pp. 51-58.

Yu, S.W., H.M. Duan, X.F. Zhu and B. SWEN, "The Specification of Basic Processing of Contemporary Chinese Corpus," *Journal of Chinese Information Processing,* vol.16, issue 5, pp. 49-64 & issue 6, pp. 58-64, 2002.

Zhang, Z.G., "xiandai hanyu," (Modern Chinese,Volume Two), People's Education Publishing House, 1982.

Zhou, M., "J-Beijing Chinese-Japanese Machine Translation System," *1999 Joint Symposium on Computational Linguistics (JSCL-1999).* Tsinghua University Press, 1999, pp. 312-319.

Zhu, D.X., "yufa jiangyi,"(Grammar Tutorial), The Commercial Press, 1982.