Predicting Morphological Types of Chinese Bi-Character Words by Machine Learning Approaches

Ting-Hao Huang, Lun-Wei Ku, Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan (R.O.C)
r96944003@csie.ntu.edu.tw; lwku@nlg.csie.ntu.edu.tw; hhchen@ntu.edu.tw

Abstract

This paper presented an overview of Chinese bi-character words' morphological types, and proposed a set of features for machine learning approaches to predict these types based on composite characters' information. First, eight morphological types were defined, and 6,500 Chinese bi-character words were annotated with these types. After pre-processing, 6,178 words were selected to construct a corpus named Reduced Set. We analyzed Reduced Set and conducted the inter-annotator agreement test. The average kappa value of 0.67 indicates a substantial agreement. Second, Bi-character words' morphological types are considered strongly related with the composite characters' parts of speech in this paper, so we proposed a set of features which can simply be extracted from dictionaries to indicate the characters' "tendency" of parts of speech. Finally, we used these features and adopted three machine learning algorithms, SVM, CRF, and Naïve Bayes, to predict the morphological types. On the average, the best algorithm CRF achieved 75% of the annotators' performance.

1. Introduction

Most Chinese characters have multiple meanings, multiple POSes, and even multiple pronunciations. Almost all Chinese characters are morphemes. They can not only be used to construct words, but also can be single-character words themselves. Moreover, the meanings of Chinese words composed of the morphemic characters are functions of the senses of the composing characters. Analysis of morphological structures is indispensable for understanding the meaning of Chinese words, and can be employed to many applications such as opinion mining (Ku, *et al*., 2009).

Several related studies (Tseng and Chen, 2002; Tseng, *et al.* 2005; Lu, *et al.* 2008) were proposed in the recent years. Tseng's work mainly focused on unknown words longer than two characters, and Lu considered two-character words as the smallest indivisible unit and ignored any morphological structures inside them. Neither of them processed bi-character words in their experiments, but a large proportion of Chinese words are bi-character. Table 1 shows that more than half of words in Chinese Treebank 5.1[1] are bi-character, that is, the morphological structures of over 50% words can not be determined by the previous researches even the methods for predicting morphological structures for words longer than two characters have been developed and the single-character words have no morphological structures.

In this paper, we present the development of a corpus with the annotation of the morphological types, and propose morphological type classifiers for bi-character words.

2. Research Objective

We have two main objectives: first, to develop a corpus of the morphological types of bi-character words; second, to implement a practical morphological type classifier for Chinese bi-character words.

word length (#character)	1	2	3	4	≧5
term freq	166,207	223,786	30,394	5,564	4,002
%	38.66%	52.05%	7.07%	1.29%	0.93%

Table 1: The term frequency of words of different lengths in Chinese Tree Bank 5.1[2]

On the aspect of developing corpus, we try to explore the native speakers' preference for understanding the morphological structures of Chinese bi-characters. So the corpus was annotated in majority rule. On the aspect of classifier, we attempt to explore the possibility of only using information of "characters": parts of speech of "words" or the context are not utilized. Chinese character is a much smaller fixed close set, which is much easier to handle than words, a character-based classifier can avoid lack of information when dealing with unknown or rare words.

3. Morphological Types

This paper investigates the relation between characters in Chinese bi-character words. We basically follow the morphological types proposed by linguists (Cheng and Tian, 1992) to develop our experimental corpus. The description is as follows:

(1) **Parallel (並列, 聯合)**: Two morphemic characters play coordinate roles in a word. For example, "財富" [cai2 fu4, money-wealth], "打罵" [da3 ma4, punish-blame], "男女" [nan2 nu3, male-female]. Note that the reduplication words (e.g., "人人", [ren2 ren2, people-people,

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T01U01

[2] Punctuation marks and non-Chinese words are not included. And some sentence with only one word like "完" [wan2, the notification of the articles' end] are not included, either.

Corpus	#word		Parallel	Substantive-Modifier	Subjective-Predicate	Verb-Object	Verb-Complement	Confirmation	Negation	Others	Segmentation Error
Original Set	6,500	#	1,514	2,935	85	826	704	43	11	269	113
		%	23.29	45.15	1.31	12.71	10.83	0.66	0.17	4.14	1.74
Reduced Set	6,187	#	1,433	2,927	85	824	704	0	0	214	0
		%	23.16	47.31	1.37	13.32	11.38	0	0	3.46	0

Table 2: Summary of the Original Set and Reduced Set

everybody], or "謝謝", [xie4 xie4, thanks-thanks, thanks]) are always of this type.

(2) **Substantive-Modifier (修飾，偏正)**: The modified character follows the modifying character. For example, "低級" [di1 ji2, low-level] and "痛哭" [tong4 ku1, bitterly-cry]. Note that the noun-noun compounds, such as "衣櫃" [yi1 gui4, cloth-cabinet, wardrobe] which contains two morphemic characters "衣" [yi1, cloth] and "櫃" [gui4, cabinet], also belong to this type because "櫃" is modified by "衣."

(3) **Subjective-Predicate (主謂)**: The second morphemic character is an expresser and the first is described. The structure is like a subject-verb sentence condensed in one word. For example, "心疼" [xin1 teng2, heart-hurt] and "氣虛" [qi4 xu1, spirit-weak].

(4) **Verb-Object (動賓，述賓)**: The first morphemic character is usually a verb which governs the second character. It makes this word similar to a verb with its object. For example, "失控" [shi1 kong4, lose-control] and "免職" [mian3 zhi2, dismiss-job].

(5) **Verb-Complement (動補，述補)**: The first morphemic character is usually a verb but sometimes is an adjective, and the second character explains the first one from different aspects. For example, "看清" [kan4 qing1, look-clearly] and "擊潰" [ji2 kui4, hit-crash].

Note that the Verb-Complement type is in post-modification form, while the Substantive-Modifier type is in pre-modification form. In modern Chinese, bi-character words in post-modification form are much rarer than those in pre-modification form, and most of the former bi-character words are of the Verb-Complement type. Only a few exceptions (e.g. "人客" [ren2 ke4, people-visitor]) are not of the Verb-Complement type and will be classified into the "Others" type.

(6) **Negation (否定)**: The first morpheme is a negation character such as "非" [fei1, no], "不" [bu4, no], "否" [fou3, no], "無" [wu2, no].

(7) **Confirmation (肯定)**: The first morpheme is an affirmation character such as "有" [you3, do; have; be].

(8) **Others**: Those words do not belong to the above seven types are of this type, including all single morpheme words (e.g. Chinese "binding word" [連綿詞]), transliteration words (e.g. "披薩" [pi1 sa4, pizza]), affixation-built words (e.g. "阿媽" [a1 ma1, prefix-mother, grandmother], "牛仔" [niu2 zi3, bull-suffix, cowboy]), abbreviation, idiomatic word, and most function words (e.g. "而且" [er2 qie3, and], "如果" [ru2 guo3, if].)

4. Annotation

4.1 Data and Annotation Method

We randomly selected 6,500 distinct Chinese bi-character words from the segmented NTCIR CIRB040 corpus[3] as the "Original Set." Then we hired ten undergraduate students from Chinese literature department as annotators and one graduate student as the expert to label these 6,500 terms. Each annotator should select one of eight types defined in Section 3 for each word. Then the ground truth will be determined by the majority procedure shown in Figure 1.

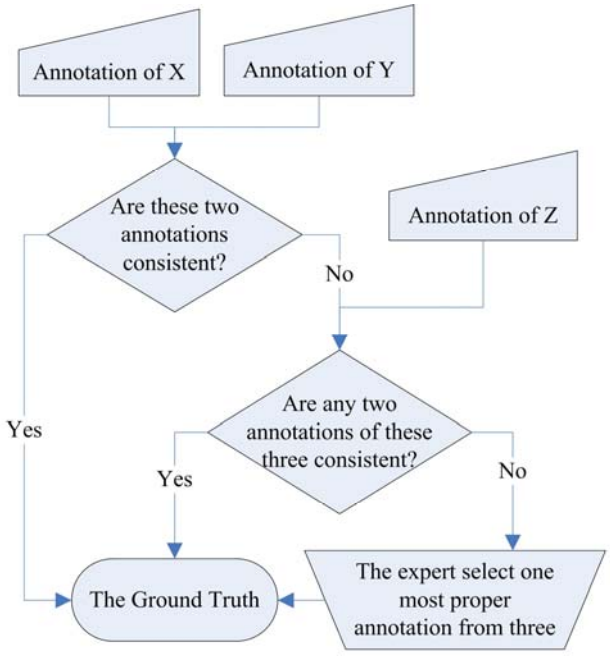

Figure 1: Annotation Procedure

[3] http://research.nii.ac.jp/ntcir/ntcir-ws6/data-en.html

annotator	K_a	accuracy	F-score					
			Parallel	Substantive-Modifier	Subjective-Predicate	Verb-Object	Verb-Complement	Others
A	0.73	0.83	0.76	0.90	0.36	0.81	0.80	0.22
B	0.73	0.82	0.71	0.88	0.50	0.90	0.85	0.22
C	0.66	0.76	0.77	0.83	0.40	0.82	0.78	0.34
D	0.84	0.89	0.84	0.93	0.40	0.93	0.83	0.64
E	0.70	0.79	0.83	0.83	0.31	0.88	0.86	0.33
F	0.83	0.85	0.78	0.90	0.62	0.90	0.83	0.64
average	0.75	0.82	0.78	0.88	0.43	0.87	0.83	0.40

Table 3: Performance of Annotators A-F

4.2 Preprocessing and Annotation Result

To focus on the prediction problem, those words which can be simply classified by their string pattern were filtered out from the Original Set. For instance, reduplication words are always of the Parallel type; words of the Negation and Confirmation types can be recognized by the appearance of characters like not or no (e.g. "非", "不", "否", "無",) and do or have (e.g. "有";) and the affixation-built words of Others type can be recognized by the appearance of affix characters (e.g. "阿" [a1, prefix], "仔" [zi3, suffix].) The resulting Reduced Set contains 6,187 words. Details of the Original Set and Reduced Set are shown in Table 2.

4.3 Agreement Test and Inter-annotator Ambiguity

To evaluate the reliability of the annotated corpus, an agreement test is performed. We randomly selected 340 words from the Reduced Set and asked six annotators to label them. A total of 15 (C_2^6) kappa values were calculated. These values range from 0.61 to 0.79 (0.67 on average), which indicates substantial agreement. Next, we calculated the kappa value (K_a) between annotators and the ground truth; the F-score, and the accuracy of each annotator to the ground truth. Results are shown in Table 3. The average performance of all annotators shows the degree of challenging for each type and will be compared to the performance of automatic classification in Section 6.

We also analyzed the disagreement between annotators, and found that the ambiguity is mainly caused by the high polysemy of Chinese characters. For instance, "文物" [wen2 wu4, cultural or historical relics] belongs to the Substantive-Modifier type when "文" is interpreted as "cultural" and "物" as "object;" however, "文" can also be interpreted as "calligraphy" and "物" can be "utensil," which are just the two major contents of Chinese cultural relics. In this case, "文物" belongs to the Parallel type. Another example is "跑開" [pao3 kai1, run-away]. It is of the Verb-Complement type when interpreting "跑" as "run" and "開" as "away". But "開" can also be interpreted as "leave," in this case, "跑開" belongs to the Parallel type due to its verb-verb structure.

Some other disagreement is caused by the ambiguity of word-building method. For instance, some words for an abstract concept such as "自由" [zi4 you2, self-from, freedom] or "文化" [wen2 hua4, culture] have confusing morphological structures, even Chinese native speakers can not determine the sense of every character clearly under such circumstances.

Note that the ambiguity of Chinese bi-character words' morphological types did not cause word sense ambiguity in almost all cases. The inter-annotator disagreement shows the different interpreting and understanding methods of Chinese characters for the same word. For instance, in "跑開" [pao3 kai1, run-away,] no matter "開" is interpreted as "away" or "leave", the referent of the whole word is the same.

5. Classification

5.1 Methodology

Three machine learning algorithms, CRF, SVM, and Naïve Bayes, are adopted. For every Chinese character C with the current pronunciation P_c, we extract a feature vector $F(C, P_c)$. Let a bi-character word be C_1C_2 with pronunciation $P_{c1}P_{c2}$. In SVM and Naïve Bayes, the feature vector is $[\ F(C_1, P_{c1}), F(C_2, P_{c2})\ ]$ and the class label is the morphological type. In the sequential labeling algorithm CRF, we consider C_1C_2 as a short "sentence" of length 2, where C_1 and C_2 are two individual "words" in the "sentence" with feature vectors $F(C_1, P_{c1})$ and $F(C_2, P_{c2})$, respectively. Then we use CRF to predict the labels of C_1 and C_2 similar to the POS tagging. If C_1C_2 is of the Substantive-Modifier type, then the label of C_1 is "*Substantive_Modifier_Prefix*" and C_2 is "*Substantive_Modifier_Suffix*," and so on. After training and labeling, the label combination with the highest probability will be selected as the prediction result.

5.2 Features

The main concept of designing $F(C, P_c)$ is that the POS of composite characters are strongly related to the words' morphological type. For instance, "Adj. + N." is likely to be the Substantive-Modifier type and "V. + N." is likely to be the Verb-Object type. Note that the POS we mentioned here is not of the characters that are single-character words in sentences but of the morphemic characters in

1. 好　部首　女　部首外筆畫　3　總筆畫　6

注音一式　ㄏㄠˇ

漢語拼音　h ǎ o　注音二式　h ǎ u

形

❶美、善，理想的。如：「好東西」、「好風景」、「花好月圓」、「好人好事」。唐・韋莊・菩薩蠻・人人盡說江南好詞：「人人盡說江南好，遊人只合江南老。」

❷友愛的。如：「好朋友」、「好同學」。

❸完整的、沒壞的。如：「完好如初」、「修好了。」

動

❶相善、彼此親愛。如：「友好」。唐・高適・贈別晉三處士詩：「知己從來不易知，慕君為人與君好。」紅樓夢・第二十七回：「誰和我好，我就和誰好。」

❷痊癒。如：「病好了！」警世通言・卷十六・小夫人金錢贈年少：「孩兒感些風寒，這幾日身子不快，來不得。傳語員外得知，一好便來。」

副

❶很、非常。表示程度深。如：「好久」、「好冷」、「好笨」、「好厲害」。

❷完成、完畢。如：「交待的工作做好了。」、「稿子寫好了。」儒林外史・第四十三回：「鄧梳好了椎髻，穿好了苗錦。」

❸容易。如：「這事好辦。」、「這問題好解決。」、「這小孩好帶。」

❹以便、便於。如：「快準備行李，好早點上路。」、「請告訴我你的住處，我好去找你。」唐・杜甫・聞官軍收河南河北詩：「白日放歌須縱酒，青春作伴好還鄉。」

❺可以、應該。如：「只好如此」、「正好試試」。官場現形記・第五十一回：「习邁彭屈指一算，後任明天好到，便約張太太三天回音。」

❻置於某些動詞之前，表效果佳。如：「好看」、「好玩」、「好吃」、「好笑」。

❼置於數量詞或時間詞之前，表示多或久的意思。如：「好些個」、「好幾處」、「好半天」、「好一會兒」。

嘆

❶表示稱讚或允許。如：「好！就這麼辦。」京本通俗小說・碾玉觀音：「郡王道：『好！正合我意。』」

❷表示責備或不滿意的語氣。如：「好！這下子事情愈來愈棘手了。」

ㄏㄠˇ　h à o（04802）

Figure 2: *The Revised Chinese Dictionary* Sample (take "好" [hao3] for instance)

POS	#sense	# Example Words									
		2-char		3-char			4-char			≧5	
		Prefix	Suffix	Prefix	Suffix	Else	Prefix	Suffix	Else		
Adjective	**3**	**0**	**0**	**4**[4]	**0**	0	**1**[5]	**0**	2[6]	0	→ V_{ADJ} = (0, 0, 4, 0, 1, 0)
Noun	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	0	0	→ V_{N} = (0, 0, 0, 0, 0, 0)
Verb	**2**	**0**	**1**[7]	**0**	**0**	0	**0**	**0**	0	0	→ V_{V} = (0, 1, 0, 0, 0, 0)
Adverb	**7**	**7**[8]	**0**	**4**[9]	**0**	0	**1**[10]	**0**	2[11]	0	→ V_{ADV} = (7, 0, 4, 0, 1, 0)
Auxiliary	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	0	0	→ V_{AUX} = (0, 0, 0, 0, 0, 0)
Conjunction	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	0	0	→ V_{CONJ} = (0, 0, 0, 0, 0, 0)
Pronoun	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	0	0	→ V_{PRON} = (0, 0, 0, 0, 0, 0)
Preposition	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	0	0	→ V_{PREP} = (0, 0, 0, 0, 0, 0)
Interjection	**2**	**0**	**0**	**0**	**0**	0	**0**	**0**	0	0	→ V_{INT} = (0, 0, 0, 0, 0, 0)

好 hao3

→ V_{POS} = (3, 0, 2, 7, 0, 0, 0, 0, 2)

Figure 3: Feature Extraction Method (take "好" [hao3] for instance)

multi-character words. However, it is difficult to recognize the actual POSes of morphemic characters within words. Even the morpheme dictionaries provide only the list of possible POSes of characters, the sense for each POS, and few example words for each sense. Exact POSes of morphemic characters are not available. Therefore, base on the intuitional assumption of the positive correlation between the numbers of senses and the "tendency" of POSes, we used the numbers of senses under all POSes in morpheme dictionaries as our features.

The POSes of morphemic characters sometimes depend on their positions in words. Take "戲" [xi4, drama] for instance, "戲" is usually used as a noun meaning "drama" at the end of a word (e.g. "看戲" [kan4 xi4, watch-drama]), but is used as a modifier (usually an adjective or an adverb) meaning "dramatic" as the prefix character (e.g. "戲弄" [xi4 nong4, dramatically-tease, to make a fool of], or "戲子" [xi4 zi3, dramatic-man, actor].)

We also extracted another set of features to represent

[4] {好東西, 好風景, 好朋友, 好同學}
[5] {好人好事}
[6] {花好月圓, 完好如初}
[7] {友好}
[8] {好久, 好冷, 好笨, 好看, 好玩, 好吃, 好笑}
[9] {好厲害, 好些個, 好幾處, 好半天}
[10] {好一會兒}
[11] {只好如此, 正好試試}

	SVM				CRF				Naïve Bayes				Annotators' average performance
	P	*R*	*F*	A_F	*P*	*R*	*F*	A_F	*P*	*R*	*F*	A_F	
Parallel	0.52	0.16	0.25	0.32	0.59	0.51	0.55	0.71	0.54	0.31	0.4	0.51	0.78
Substantive-Modifier	0.54	0.95	0.69	0.78	0.73	0.81	0.77	0.88	0.8	0.56	0.66	0.75	0.88
Subjective-Predicate	-	0	-	-	0.36	0.3	0.33	0.77	0.15	0.77	0.25	0.58	0.43
Verb-Object	0.66	0.2	0.31	0.36	0.6	0.56	0.58	0.67	0.53	0.66	0.59	0.68	0.87
Verb-Complement	0.78	0.4	0.53	0.64	0.77	0.79	0.78	0.94	0.47	0.84	0.6	0.72	0.83
Others	-	0	-	-	0.31	0.17	0.22	0.55	0.11	0.29	0.16	0.40	0.4
average	-	-	-	-	0.56	0.52	0.54	0.75	0.43	0.57	0.44	0.61	0.7

Table 4: Experiment Results

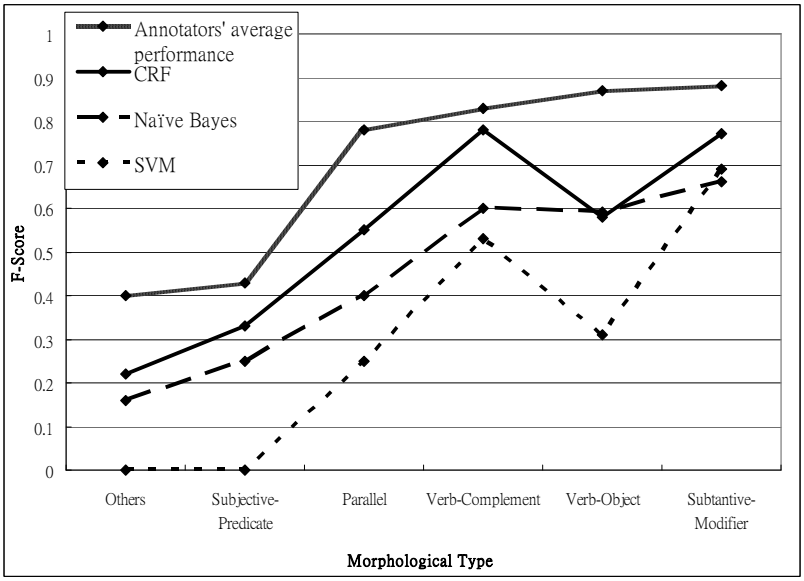

Figure 4: Classifiers' performance

position information from morpheme dictionaries: For all the 2, 3, 4-character example words of *C* (with the current pronunciation P_c,) we calculate the number of occurrences when *C* is the prefix/suffix character for each POS.

In this work, *The Revised Chinese Dictionary* (Ministry of Education, Taiwan, 1994) was selected as the morpheme dictionary. This dictionary defined nine POS for morphemes and used the character and pronunciation together as the primary key, which is a suitable resource for our feature extraction. Figure 3 takes "好" [hao3] for example to show how we extracted all features from the morpheme dictionary (Figure 2 is the original web page.)

For the character *C* with the current pronunciation P_c, we first extracted the *pronunciation feature vector* $f_p(C, P_c)$. In Figure 3, the value of f_p(好, *hao3)* is calculated by (1):

$$f_p(\text{好, hao3}) = (V_{POS}, V_{ADJ}, V_N, V_V, V_{ADV}, V_{AUX}, V_{CONJ}, V_{PRON}, V_{PREP}, V_{INT}) \quad (1)$$

However, the current pronunciations of composite characters cannot always be given. To prevent the zero vectors appear when the pronunciations are unknown, the unpronounced feature set must be added. We designed the *unpronounced feature vector* $f_{up}(C)$ as the vectors' sum of all $f_p(C, P_i)$ where P_i is one pronunciation of *C*. For instance, "好" has two pronunciations, "hao3" and "hao4", so that f_{up}(好) is the vector sum of f_p(好, *hao3)* and f_p(好, *hao4)*. Besides, the POS are also related to the current tones of characters in some cases. In the case of "好", "好" is usually used as a modifier which means "good, well" when its pronunciation is "hao3," but is treated as a verb which means "like" when its pronunciation is "hao4." We defined $t_c(P_c)$ as the current tone for P_c and used it as our feature. Finally, $F(C, P_c)$ can be calculated by (2):

$$F(C, P_c) = (f_p(C, P_c), \ f_{up}(C), \ t_c(P_c)),$$
$$f_{up}(C) = \sum_{i}^{\text{all } P_i \text{ of } C} f_p(C, P_i) \quad (2)$$

In the case of "好" [hao3, good] in "好人" [hao3 ren2, good-person], F(好, *hao3)* can be given by (3):

$$F(好, hao3) = (\ f_p(好, hao3),\ f_p(好, hao3)+ f_p(好, hao4),\ 3\ ) \quad (3)$$

6. Experiments

In the experiments, we adopted *LIBSVM* (Chang and Lin, 2001) for the SVM classification, *CRF++* (Kudo, 2006) for the CRF classification and *Rainbow* (McCallum, 1996) for the Naïve Bayes classification. All the evaluations were performed with four-fold cross-validation on the Reduced Set (6,187 words). The average recall (*R*), the average precision (*P*), and the macro-average F-score (*F*) were calculated. Results are shown in Table 4 and Figure 4. Besides these measures, we also calculated the ratio of classifiers' F-scores to the annotators' average performance (the achievement of F-score, A_F) for each type. This ratio shows how much the classifier can achieve annotators' performance. On the average, the best classifier CRF achieved 75% of the annotators' average performance.

7. Discussions

Except the words of the Subjective-Predicate and Others types which are obviously smaller in number (refer to Table 2), the other four types have similar degree of challenging in prediction due to the similar annotators' performance (refer to Table 3). However, our classifiers performed differently for different types. They performed better for words of Substantive-Modifier and Verb-Complement types, while worse for Verb-Object and Parallel types.

After analysis, we found that a linguistic phenomenon called "conversion (轉品)" has large effects. In Chinese, linguistic units may change from their usual POSes into the rarer ones in some specific situations. The conversions of POS from verbs or nouns to pre-modifiers (mostly adjectives or adverbs) are common seen in bi-character words. Take "跑" [pao3, run] as an example. When followed by the noun "步" [bu4, pace] as a verb, they together form a Verb-Object word. Instead, in the Substantive-Modifier word "跑車" [pao3 che1, running-car, sports car,] its POS converts to a pre-modifier (adjective). Another example is "書" [shu1, book]. When followed by the noun "報" [bao4, newspaper] as its usual POS noun, they form a Parallel word; when followed by another noun "桌" [zhuo1, table] as a pre-modifier (adjective), they form the Substantive-Modifier word "書桌" [book-table].

To identify this linguistic phenomenon, semantic information is necessary. Our features represent only POSes and positions of characters, so the classifiers did not performed well for such kinds of words except those of the Verb-Complement type. Conversion usually happens in the situation of pre-modification while Verb-Complement is a word-building method in the post-modification form.

Theoretically, this phenomenon would damage the classifiers' performance in the Substantive-Modifier type as well. However, this type had a big advantage in quantity, and the fact that Substantive-Modifier words accounted for nearly half of the corpus suggested that the pre-modification is a strong word-building principal for Chinese bi-character words, which would be much easier to learn. It might explain why the classifiers still performed the best for words of the Substantive-Modifier type.

On the aspect of the classifiers, CRF outperforms SVM and Naïve Bayes. In our work, CRF predicted the morphological types from the viewpoint of composite characters while SVM and Naïve Bayes from the viewpoint of words. The result shows that characters are more useful than words when predicting the morphological structures.

Generally speaking, the character-based method we proposed had pros and cons. It was a simple and effective approach using only the information of Chinese characters which is a fixed close set and easy to maintain. Without using any information of words or their context, the character-based method can also avoid the data sparseness when processing the unknown or rare words because the character set is a close set. However, there is a problem of character-based method: the information quantity of every Chinese character is too large to clarify in many cases: most ambiguity, such as the inter-annotators ambiguity mentioned in Section 4.3, or the "conversion" phenomenon, is basically caused by it.

There are two possible directions for solving this problem: up to the word level, or down to the characters' sense level. On the aspect of words, add more information of words could help to identify "conversion" characters. For instance, knowing "跑步" [run-pace] is a verb and "跑車" [running-car, sports car] is a noun might help us to determine the POSes of "跑" in those words. On the aspect of characters' senses, if we understand the concept relations (e.g. from WordNet or other concept nets) between "書" [book], "報" [newspaper], and "桌" [table], it might be easier to tell "書報" [book- newspaper] is a Parallel word and "書桌" [book-table] is a Substantive-Modifier word.

8. Conclusion and Future Work

The two main contributions of this paper are the construction of corpus and the prediction of the morphological type of words. In terms of corpus annotation, we followed a the proposed classification scheme to label 6,500 Chinese two-character words and gave a detailed analysis of the annotation results. In terms of morphological type prediction, we proposed a set of features and experimented with three machine learning algorithms. The classification performance achieved 75% of the annotators' average performance. We also applied the morphological type classifiers to practical systems (Ku, *et al.*, 2009). We used the classifier to provide additional information for opinion extraction and improved the performance significantly.

In the future, we will consider the uses of semantic information to deal with the conversion problem, and introduce more external features such as POSes of words or the context information. Besides, applications of the

morphological types in other NLP problems will be investigated.

9. Acknowledgements

Ting-Hao Huang would like to thank Miss Hui-Yun Song for her valuable help, and Overseas Compatriot Affairs Commission, R.O.C. (Taiwan) for the unreserved support during his substitute military service term.

10. References

C.C. Chang and C.J. Lin. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Cheng, X.-H. and Tian, X.-L. (1992). Modern Chinese. Bookman Books Ltd.

Huang, T.-H. (2009). Automatic Extraction of Intra- and Inter- Word Syntactic Structure for Chinese Opinion Analysis. Master's Thesis. National Taiwan University, R.O.C.(Taiwan).

Ku, L.-W., Huang T.-H., and Chen, H.-H. (2009). Using Morphological and Syntactic Structures for Chinese Opinion Analysis. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1260–1269.

Kudo T. (2005). CRF++: Yet another CRF toolkit. Software available at http://crfpp.sourceforge.net.

Lu, J., M. Asahara and Y. Matsumoto (2008). Analyzing Chinese Synthetic Words with Tree-based Information and a Survey on Chinese Morphologically Derived Words. Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, pages 53–60.

McCallum, Andrew Kachites (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. Software available at http://www.cs.cmu.edu/~mccallum/bow.

The Revised Chinese dictionary. (1994). Online Version. Taipei: Ministry of Education, Taiwan. Available at http://dict.revised.moe.edu.tw.

Tseng, H.-H. and K.-J. Chen (2002). Design of Chinese Morphological Analyzer. Proceedings of SIGHAN Workshop on Chinese Language Processing, pages 49-55.

Tseng, H.-H., D. Jurafsky, and C. Manning (2005). Morphological features help POS tagging of unknown words across language varieties. In Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, pages 32–39.