Construction of a Chinese Idiom Knowledge Base and Its Applications

Lei Wang
Key Laboratory of Computational Linguistics of Ministry of Education
Department of English, Peking University
wangleics@pku.edu.cn

Shiwen Yu
Key Laboratory of Computational Linguistics of Ministry of Education,
Peking University
yusw@pku.edu.cn

Abstract

Idioms are not only interesting but also distinctive in a language for its continuity and metaphorical meaning in its context. This paper introduces the construction of a Chinese idiom knowledge base by the Institute of Computational Linguistics at Peking University and describes an experiment that aims at the automatic emotion classification of Chinese idioms. In the process, we expect to know more about how the constituents in a fossilized composition like an idiom function so as to affect its semantic or grammatical properties. As an important Chinese language resource, our idiom knowledge base will play a major role in applications such as linguistic research, teaching Chinese as a foreign language and even as a tool for preserving this non-material Chinese cultural and historical heritage.

1 Introduction

An idiom is a multi-word expression that has a figurative meaning that is comprehended in regard to a common use of that expression that is separate from the literal meaning or definition of the words of which it is made (McArthur, 1992). From a linguistic perspective, idioms are usually presumed to be figures of speech that are contradictory to the principle of compositionality. The words that construct an idiom no longer keep their original meaning or popular sense, while in the process of its formation it develops a specialized meaning as an entity whose sense is different from the literal meanings of the constituent elements.

Although an idiom is an expression not readily analyzable from its grammatical structure or from the meaning of its component words, it is the distinctive form or construction of a particular language that has a unique form or style characteristic only of that language. An idiom is also used, in most cases, with some intention of the writer or to express certain emotion or attitude. Thus in nature, idioms are exaggerative and descriptive and do not belong to the plain type.

Therefore, to classify idioms according to its emotional property or descriptive property is important for many practical applications. In recent years, emotion classification has become a very popular task in the area of Natural Language Processing (NLP), which tries to predict sentiment (opinion, emotion, etc.) from texts. Most research has focused on subjectivity (subjective/objective) or polarity (positive/neutral/negative) classification. The applications with this respect include human or machine translation, automatic text classification or Teaching Chinese as a Foreign Language (TCFL). For example, when a student learning Chinese as a foreign language encounters an idiom in his or her reading or conversation, for better understanding it is important for him or her to know whether the idiom is used to indicate an appreciative or derogatory sense which is very crucial to understand the attitude of the idiom user. Another example is that long articles about politics in newspapers often include a lot of idiom usage to boost their expressiveness and these idioms may carry emotional information. Obviously by knowing the emotional inclination we may easily obtain a clue about the general attitude of the particular medium. We may even be able to detect or monitor automatically the possible hostile attitude from certain electronic media which today provide so huge amount of information that seems hard for human processing on a daily basis.

The rest of this paper is organized as follows. Section 2 describes the construction of

Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), pages 10–17,
Beijing, August 2010

a Chinese Idiom Knowledge Base (CIKB) and introduces its several applications so far. Section 3 concludes the related work that serves as the basis of the building of CIKB and the emotion classification experiment introduced in this paper. Section 4 describes the classification method, feature settings, the process of emotion classification and the analysis of the result. Section 5 includes conclusions and our future work.

2 Chinese Idioms and Chinese Idiom Knowledge Base

Generally an idiom is a metaphor — a term requiring some background knowledge, contextual information, or cultural experience, mostly to use only within a particular language, where conversational parties must possess common cultural references. Therefore, idioms are not considered part of the language, but part of a nation's history, society or culture. As culture typically is localized, idioms often can only be understood within the same cultural background; nevertheless, this is not a definite rule because some idioms can overcome cultural barriers and easily be translated across languages, and their metaphoric meanings can still be deduced. Contrary to common knowledge that language is a living thing, idioms do not readily change as time passes. Some idioms gain and lose favor in popular literature or speeches, but they rarely have any actual shift in their constructs as long as they do not become extinct. In real life, people also have a natural tendency to over exaggerate what they mean or over describe what they see or hear sometimes and this gives birth to new idioms by accident.

Most Chinese idioms (成语: chéng[1] yǔ, literally meaning "set phrases") are derived from ancient literature, especially Chinese classics, and are widely used in written Chinese texts. Some idioms appear in spoken or vernacular Chinese. The majority of Chinese idioms consist of four characters, but some have fewer or more. The meaning of an idiom usually surpasses the sum of the meanings carried by the few characters, as Chinese idioms are often closely related with the fable, story or historical account from which they were originally born. As their constructs remain stable through history, Chinese idioms do not follow the usual lexical pattern and syntax of modern Chinese language which has been reformed many a time. They are instead highly compact and resemble more ancient Chinese language in many linguistic features.

Usually a Chinese idiom reflects the moral behind the story that it is derived. (Lo, 1997) For example, the idiom "破釜沉舟" (pò fǔ chén zhōu) literally means "smash the cauldrons and sink the boats." It was based on a historical story where General Xiang Yu in Qin Dynasty (221 B. C. – 207 B. C.) ordered his army to destroy all cooking utensils and boats after they crossed a river into the enemy's territory. He and his men won the battle for their "life or death" courage and "no-retreat" policy. Although there are similar phrases in English, such as "burning bridges" or "crossing the Rubicon", this particular idiom cannot be used in a losing scenario because the story behind it does not indicate a failure. Another typical example is the idiom "瓜田李下" (guā tián lǐ xià) which literally means "melon field, under the plum trees". Metaphorically it implies a suspicious situation. Derived from a verse called 《君子行》 (jūn zǐ xíng, meaning "A Gentleman's Journey") from Eastern Han Dynasty (A. D. 25 – A. D. 220), the idiom is originated from two lines of the poem "瓜田不纳履，李下不整冠" (guā tián bù nà lǚ, lǐ xià bù zhěng guān) which describe a code of conduct for a gentleman that says "Don't adjust your shoes in a melon field and don't tidy your hat under plum trees" in order to avoid suspicion of stealing. However, most Chinese idioms do not possess an allusion nature and are just a combination of morphemes that will give this set phrase phonetic, semantic or formal expressiveness. For example, the idiom "欢天喜地" (huān tiān xǐ dì, metaphorically meaning "be highly delighted") literally means "happy heaven and joyful earth"; or in the idiom "锒铛入狱" (láng dāng rù yù, meaning "be thrown into the jail"), the word "锒铛" is just the sound of a prisoner's fetters.

[1] The marks on the letters in a Pinyin are for the five tones of Chinese characters.

For the importance of idioms in Chinese language and culture, an idiom bank with about 6,790 entries were included in the most influential Chinese language knowledge base – the Grammatical Knowledge base of Contemporary Chinese (GKB) completed by the Institute of Computational Linguistics at Peking University (ICL), which has been working on language resources for over 20 years and building many knowledge bases on Chinese language. Based on that, the Chinese Idiom Knowledge Base (CIKB) had been constructed from the year 2004 to 2009 and collects more than 38, 000 idioms with more semantic and pragmatic properties added.

Basically the properties of each entry in CIKB can be classified into four categories: lexical, semantic, syntactic and pragmatic, each of which also includes several fields in its container -- the SQL database. Table 1 shows the details about the fields.

Categories	Properties
Lexical	idiom, Pinyin[2], full Pinyin[3], bianti[4], explanation, origin
Semantic	synonym, antonym, literal translation, free translation, English equivalent
Syntactic	compositionality, syntactic function
Pragmatic	frequency, emotion, event (context), grade

Table 1. Property categories of CIKB.

There are three fields of translation as we can see in Table 1. In spite of the fact that a literal translation of an idiom will not reflect its metaphorical meaning generally, it will still be of value to those who expect to get familiar with the constituent characters and may want to connect its literal meaning with its metaphorical meaning, especially for those learners of Chinese as a foreign language.

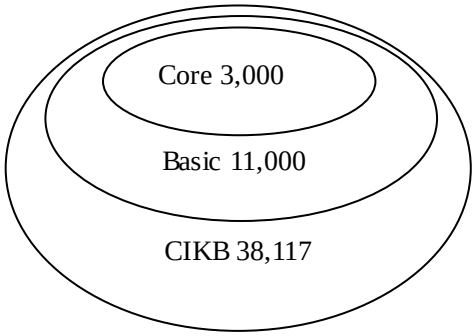

Figure 1. The hierarchical structure of CIKB.

The idioms are classified into three grades in terms of appearance in texts and complexity of annotation. The most commonly used 3,000 idioms serve as the core idioms based on the statistics obtained from the corpus of People's Daily (the year of 1998), a newspaper that has the largest circulation in China. Another 11,000 idioms are selected into a category named as basic idioms (fully annotated in every field) and the total 38,117 forms the whole knowledge base. Its hierarchical structure can be seen in Figure 1.

The syntactic category aims at NLP tasks like automatic identification or machine translation. Compared with English idioms, the identification of Chinese idioms is not so difficult for its fossilized structure, i.e. continuity in a text. To build a lexicon like CIKB will complete the task successfully. As for machine translation, however, it is completely another story because the compositional complexity of Chinese idioms enables them to function as different syntactic constituents with variable part-of-speech (POS). We classify them into nine categories according to its compositional relations of the morphemes and into seven categories according to its syntactic functions that they may serve in a sentence, as is shown in Table 2.

[2] Pinyin (拼音, literally "phonetics", or more literally, "spelling sound" or "spelled sound"), or more formally Hanyu Pinyin (汉语拼音, Chinese Pinyin), is currently the most commonly used Romanization system for standard Mandarin. The system is now used in mainland China, Hong Kong, Macau, parts of Taiwan, Malaysia and Singapore to teach Mandarin Chinese and internationally to teach Mandarin as a second language. It is also often used to spell Chinese names in foreign publications and can be used to enter Chinese characters on computers and cell phones.

[3] full Pinyin, a form of Pinyin that replaces the tone marks with numbers 1 to 5 to indicate the five tones of Chinese characters for the convenience of computer processing.

[4] bianti, a variant form of the idiom that was caused by random misuse, literary malapropism, etc.

No.	Compositionality	Tag	No.	Syntactic function	Tag
1	modifier-head construction	pz	1	as a noun	IN
2	subject-predicate phrase	zw	2	as a verb	IV
3	Coordination	bl	3	as an adjective	IA
4	predicate-object phrase	db	4	as a complement	IC
5	predicate-complement	dbu	5	as an adverbial	ID
6	predicate-object-complement	dbb	6	as a classifier	IB
7	serial verb	ld	7	as a modifier	IM
8	pivotal verb	jy			
9	Repetition	fz			

Table 2. Compositionality and syntactic functions of idioms.

Upon the completion of CIKB, a few research projects have been conducted to investigate possible applications. Li (2006) investigates the frequency and formation of idiom usage in People's Daily and Wang (2010) selects 1,000 popular idioms from CIKB to compile a book for Chinese learners. On the basis of CIKB, we also made a couple of attempts on the automatic classification of idioms to identify the token-level characteristics of an idiom. This paper will focus on the emotion classification of idioms with machine learning method and the work will be elaborated in section 4. Here we define the emotion types as "appreciative (A)", "derogatory (D)" and "neutral (N)".

3 Related Work on Idiom Knowledge Base and Its Applications

There has not been much work on the construction of an idiom corpus or an idiom knowledge base. With this respect, Birke and Sarkar (2006) and Fellbaum (2007) are exceptions. Birke and Sarkar (2006) constructed a corpus of English idiomatic expressions with automatic method. They selected 50 expressions and collected about 6,600 examples. They call the corpus TroFi Example Base, which is available on the Web.

As far as idiom identification is concerned, the work is classified into two kinds: one is for idiom types and the other is for idiom tokens. With the former, phrases that can be interpreted as idioms are found in text corpora, typically for lexicographers to compile idiom dictionaries. Previous studies have mostly focused on the idiom type identification (Lin, 1999; Baldwin et al., 2003; Shudo et al., 2004). However, there has been a growing interest in idiom token identification recently (Katz and Giesbrecht, 2006; Hashimoto et al., 2006; Cook et al., 2007). Our work elaborated in section 4 is also an attempt in this regard.

Despite the recent enthusiasm for multiword expressions, the idiom token identification is in an early stage of its development. Given that many language teaching and learning tasks like TCFL have been developed as a result of the availability of language resources, idiom token identification should also be developed when adequate idiom resources are provided. To this end, we have constructed the CIKB and hope to find applications of value, for example, emotion classification, event classification and text analysis based on idiom usage and its context.

According to the granularity of text, emotion analysis of texts can be divided into three levels: text (Pang et al., 2002; Cui et al., 2006), sentence (Pang et al., 2004), word (Hatzivassiloglou et al., 1997; Wiebe 2000). According to the sources of emotion prediction, classification methods can be divided into knowledge based methods and machine learning based methods. The former uses lexicons or knowledge bases to build a new lexicon that contains emotion words. WordNet is often used to compute the emotion prediction of words (Hatzivassiloglou et al., 1997; Andrea 2005). Meanwhile, incorporating knowledge into the machine learning architecture as features is a popular trend and untagged copra are often used to do emotion classification research (Turney et al., 2002; Akkaya et al., 2009).

4 An NLP Application of Emotion Classification on CIKB

In this paper, we focus on the emotion prediction of idioms conducted by machine learning method. To do this, we aim to investigate how the compositional constituents of an idiom affect its emotion orientation from the token level, especially for multi-word expressions with so obvious an exaggerative and descriptive nature like idioms. From CIKB, 20,000 idioms are selected as the training corpus and 3,000 idioms as the test corpus. The detailed distribution of idioms in each emotion group is shown in Table 3. We can see that neutral has the largest number of idioms, accounting for 41.08% and 36.67% in the training and test corpus respectively, but there is not a big difference between groups.

	Training corpus		**Test corpus**	
	number	**percentage**	**number**	**Percentage**
Appreciative(A)	6967	34.84%	1011	33.70%
Neutral(N)	8216	41.08%	1100	36.67%
Derogatory(D)	4817	24.08%	889	29.63%

Table 3. The distribution of idioms in each emotion group.

Support Vector Machine (SVM) (Cortes and Vapnik, 1995) is adopted as the classification method to predict emotions in idioms. LIBLINEAR (Fan et al., 2008), a library for large SVM linear classification, is used for implementation. The solver is set be L2-loss SVM dual. Parameter *C* is set to be 2^{-5}. Three classes of features and their various combinations are examined and used, including Chinese characters, words and part-of-speeches. Detailed features and related abbreviations are shown as in Table 4.

Because Chinese sentences are written in a consecutive string of characters, we need to segment a sentence into individual words to obtain the word feature. ICTCLAS (Zhang et al., 2003), a tool developed by the Institute of Computing Technology of Chinese Academy of Sciences (ICT), is used for word segmentation and part-of-speech tagging. We adopt precision, recall and F-score (ß=1) as the evaluation parameters. From Table 5 we can see that i_cb has a better performance than i_cu, which indicates that a bigram model usually performs better than a unigram model. But when we segment the idioms and use i_wu, we find that the performance gets bad. This may be because the compositionality of Chinese idioms is quite fossilized and the errors caused by segmentation introduce some noise.

Features and their abbreviations		**Idiom(i)**	**Explanation(e)**
Chinese characters	character unigram(i_cu, e_cu)	√[5]	√
	character bigram(i_cb, e_cb)	√	√
Words	word unigram(i_wu, e_wu)	√	√
	word bigram(i_wb, e_wu)	×	√
Word/part-of-speech	word/pos unigram(i_wpu, e_wpu)	√	√
	word/pos bigram(i_wpb, e_wpb)	×	×

Table 4. Features selected for emotion prediction.

[5] " √ " indicates the feature is selected while " × " indicates the feature is not selected.

We want to know whether we will have a better performance if we add more features from the other fields of CIKB. Obviously the most relevant feature will be the explanation of an idiom. Therefore we add the texts in the explanation field as features in the experiment. We find that by adding more features from the explanation field, the performance does improve. But when the POS feature is introduced, the performance gets bad. This may be because as Chinese idioms keep grammatical properties of ancient Chinese language and its POS is very different from the setting of the tool designed primarily for modern Chinese, more noise is introduced by using POS here. Finally we can see that the combination i_cu+i_cb+e_wu+e_wb achieves the best performance in both Chinese character features and word features.

Most importantly, we notice that although for idioms themselves segmentation does not affect the performance in a positive way, segmentation of the explanations does improve the performance. Thus we may conclude that the compositionality of an idiom is very different from its explanation which is written in modern Chinese while the idiom itself is still character-based and keeps its original morphemes that are inherited from ancient Chinese language.

Features or features combined	Result		
	Precision	Recall	F-score
i_cu	63.23%	75.16%	68.68%
i_cb	65.78%	78.24%	71.47%
i_wu	62.51%	73.42%	68.35%
i_wpu	60.03%	71.89%	65.43%
i_cu+e_wu	66.40%	80.05%	72.59%
i_cu+e_wpu	65.68%	77.95%	71.29%
i_cu+e_wb	65.08%	76.14%	70.18%
I_cu+i_cb	67.33%	80.82%	73.46%
i_cu+i_cb+e_wu	68.55%	81.37%	74.41%
i_cu+i_cb+e_wu+e_wb	70.18%	82.71%	75.93%

Table 5. The result of emotion classification with idioms and their explanations.

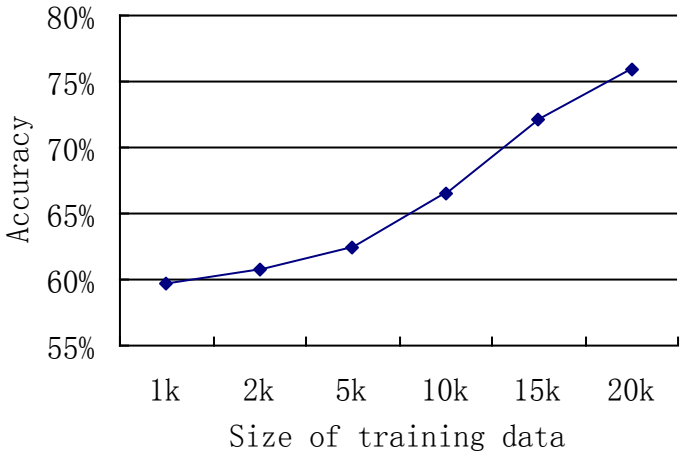

Figure 2. Learning curve of the feature combination i_cu+i_cb+e_wu+e_wb.

Figure 2 shows the learning curve of the best classifier with the feature combination i_cu+i_cb+e_wu+e_wb. We can see that the accuracy keeps improving with the increase of the size of training set, and peaks at 20,000 idioms. It shows the potential to improve the

performance of emotion classification by enlarging the training data set.

5 Conclusions and Future Work

This paper introduces the construction of CIKB by ICL at Peking University and its several applications so far. One application – the emotion classification of idioms – was elaborated to show our effort in exploring the token-level characteristics of Chinese idioms. Therefore we select a number of idioms from CIKB to classify them into three emotion groups. SVM is employed for automatic classification. Three classes of features are examined and experiments show that certain feature combinations achieve good performance. The learning curve indicates that performance may be further improved with the increase of training data size.

Now we also hope to classify the idioms into categories according to their usage in context, i.e., under what circumstances they are often used (event classification). Various linguistic features and real-world knowledge will be considered to incorporate into the machine learning classifier to improve classification result. The work is in progress and we hope the emotion classification and the event classification will be compared to determine their underlining relations and hope that more applications can be found in our future work based on CIKB.

Acknowledgements

The work in this paper is supported by a grant from the 973 National Basic Research Program of China (No. 2004CB318102). The authors are grateful to Dr. Li Yun and Professor Zhu Xuefeng for their work on CIKB and the anonymous reviewers for their helpful advice to improve the paper.

References

Akkaya, Cem, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity Word Sense Disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1:* pp.190-199.

Andrea, Esuli. 2005. Determining the Semantic Orientation of Terms through Gloss Classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*: pp.617-624.

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18:* pp.89-96.

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling Their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*: pp. 41-48.

Cortes, Corinna and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning*, 20(3): pp. 273-297.

Cui, Hang, Vibhu Mittal, and Mayur Datar. 2006. Comparative Experiments on Sentiment Classification for Online Product Reviews. In *Proceedings of the 21st National Conference on Artificial Intelligence-Volume 2*: pp.1265-1270.

Fan, Rong-En, Chang Kai-Wei, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9 (2008): pp.1871-1874.

Fellbaum, Christiane. 2007. *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies (Research in Corpus and Discourse)*. Continuum International Publishing Group Ltd., London, UK.

Hashimoto, Chikara, Satoshi Sato, and Takehito Utsuro. 2006. Japanese Idiom Recognition: Drawing a Line between Literal and Idiomatic Meanings. In *Proceedings of the COLING/ACL on*

Main Conference Poster Sessions: pp. 353-360.

Hatzivassiloglou, Vasileios, and Kathleen McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics:* pp.174-181.

Katz, Graham, and Eugenie Giesbrecht. 2006. Automatic Identification of Non-compositional Multi-word Expressions Using Latent Semantic Analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*: pp.12-19.

Li, Yun, Zhang Huarui, Wang Hongjun, and Yu Shiwen. 2006. Investigation on the Frequency and Formation of Idioms in People's Daily. In *Proceedings of the 7th Chinese Lexicon and Semantics Workshop:* pp.241-248.

Lin, Dekang. 1999. Automatic Identification of Noncompositional Phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*: pp.317-324.

Lo, Wing Huen. 1997. *Best Chinese Idioms (Vol. 3)*. Hai Feng Publishing Co., Hong Kong, China.

McArthur, Tom. 1992. *The Oxford Companion to the English Language*. Oxford University Press, Oxford, UK.

Pang, Bo and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*: pp.271-278.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumb up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing:* pp.79-86.

Shudo, Kosho, Toshifumi Tanabe, Masahito Takahashi, and Kenji Yoshimura. 2004. MWEs as Nonpropositional Content Indicators. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*: pp.32-39.

Turney, Peter D. 2002. Thumps Up or Thumps Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*: pp.417-424.

Wang, Lei. Forthcoming 2010. *1,000 Idioms for Chinese Learners*. Peking University Press, Beijing, China.

Wiebe, Janyce. 2000. Learning Subjective Adjectives from Corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence:* pp.735-740.

Zhang, Huaping, Yu Hongkui, Xiong Deyi, Liu Qun. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*: pp.184-187.