Construction of a Chinese Idiom Knowledge Base and Its Applications

Lei Wang & Shiwen Yu

Key Laboratory of Computational Linguistics of Ministry of Education, Peking University

A few about idioms

- 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).

A few about idioms

- 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.

A few about idioms

- An idiom is a metaphor — a term requiring some background knowledge, contextual information, or cultural references.
- Idioms are not considered part of the language, but part of a nation's history, society or culture.

A few about idioms

- Some idioms can overcome cultural barriers and easily be translated across languages, and their metaphoric meanings can still be deduced.
- 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.

A few about idioms

- 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.

Chinese Idioms and Chinese Idiom Knowledge Base

- 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.

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

The Analects of Confucius

Chinese Idioms and Chinese Idiom Knowledge Base

- The majority of Chinese idioms consist of four characters, but some have fewer or more.

  鸿门宴　　天有不测风云

- 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.

Idioms and emotion

- 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.

Idioms and emotion

- 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.
  - Teaching Chinese as a Foreign Language (TCFL)
  - Political articles or editorials

Chinese Idioms and Chinese Idiom Knowledge Base

- Chinese idioms are often closely related with the fable, story or historical account from which they were originally born.
- Usually a Chinese idiom reflects the moral behind the story that it is derived. (Lo, 1997)

“破釜沉舟”(pò fǔ chén zhōu)

- “smash the cauldrons and sink the boats.”
- 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.
- (pic from www.sogou.com)

“瓜田李下” (guā tián lǐ xià)

- “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)

“瓜田李下”(guā tián lǐ xià)

- 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.
- (pic from www.sogou.com)

Most idioms: phonetic, semantic or formal expressiveness

- “欢天喜地”(huān tiān xǐ dì, metaphorically meaning “be highly delighted”) literally means “happy heaven and joyful earth”.
- “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.

Chinese Idioms and Chinese Idiom Knowledge Base

- An idiom bank with about 6,790 entries were included in the Grammatical Knowledge base of Contemporary Chinese (GKB) completed by the Institute of Computational Linguistics at Peking University (ICL).
- 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.

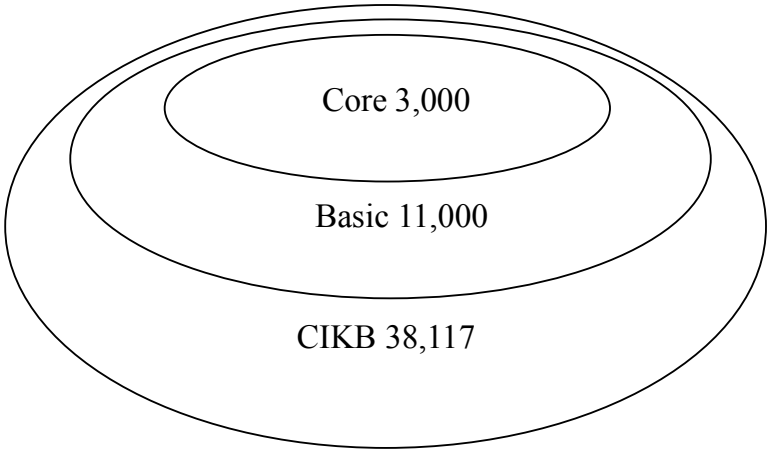
Chinese Idioms and Chinese Idiom Knowledge Base

- 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.

Property categories of CIKB

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

The hierarchical structure of CIKB



Research and 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.

About CIKB

- 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.

Compositionality and syntactic functions of idioms

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			

An NLP Application on CIKB– Emotion Prediction of idioms

- The emotion prediction of idioms conducted by machine learning method :
- Our 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.

	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%

Features selected for emotion prediction.

Features and their abbreviations		Idiom(i)	Explanation(e)
Chinese characters	character unigram(i_cu, e_cu)	√	√
	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)	×	×

The result of emotion classification with idioms and their explanations

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%

The experiments show that:

- Observation: Although for idioms themselves segmentation does not affect the performance in a positive way, segmentation of the explanations does improve the performance.
- 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.

Conclusions and Future Work

- Our work is on 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.

Conclusions and Future Work

- 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 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

- Our work 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.

- Thank you for your attention.
- Questions?