Modeling the Idiomaticity of Chinese Quadra-syllabic Idiomatic Expressions

Shu-Kai Hsieh **Yu-Hsiang Tseng** **Chiung-Yu Chiang**
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

This paper proposes a computational model of idiomaticity for Chinese Quadra-syllabic idiomatic expressions based on variations, compoundness and compositeness measure. Two classification experiments are conducted to test the model, together with linguistic analysis of the connection to wordnet. The result is promising and we believe that it will shed more light on our understanding of cognitive dynamics that underlies multiword expressions processing.

1 Introduction

Multiword expressions (MWEs) as the *habitual recurrent word combinations* in our daily language use have been regarded as the bottleneck in current NLP technology. In this paper, we will focus on a special type of idiomatic expressions of even length in Chinese called Quadra-syllabic Idiomatic Expressions (QIEs), which have pervasive presence in the Sinosphere (e.g., Japan, Korea, Vietnam, and other ethnic groups like the Naxi) due to the influence of emblematic logographic writing systems (Tsou, 2012).

Traditionally, idioms/idiomatic expressions are defined as MWEs for which the semantic interpretation is not a compositional function of their composing units. Over the past years, a rich amount of analytic works on them for mainly European languages has been proposed. Main efforts have been made to their linguistic and statistic characteristics, and the computational treatment as well. However, due to the lack of cross-language comparative work, QIEs as an idiosyncratic and indispensable part in Chinese and other languages haven't been well studied in the area of current MWE paradigm.

From the usage-based emergentist perspective, as one type of MWEs, Chinese QIEs are characterized by a holistic storage format that reveals high-level entrenchment and constructionist accounts of complex linguistic strings in the minds of language users. However, it is also noted that corpus evidence and acceptability ratings support that idioms are subject to variation too (Geeraert et al., 2017). This paper takes the challenge in modeling the QIE's idiomatic behaviour along three crucial dimensions, and explores their mapping to the synset of Chinese wordnet.

2 Chinese QIEs

The notion of *idiomaticity* has been proposed since (Chafe, 1968) and the issues debated in NLP have been well-recognized (Sag et al., 2002).

Quadra-syllabic Idiomatic Expressions (QIEs) in Chinese can be considered as a special type of idiomatic expressions of even length (i.e., four characters). In this paper, we further divide QIEs into two main types: idioms ('chengyu') and prefabs (Hsieh et al., 2017). Idiom-QIEs often involves Locus Classicus and awareness of cultural background with classical Chinese, they are formed through ages of constant use, well-compiled in dictionary and learned in school (e.g., 化險為夷 hua4 xian3 wei2 yi2,'turn danger to safety'). With their archaic origins, idiom-QIE in particular, are still prevalent in modern use and behaves more vividly than its synonyms represented by common lexemes.

tsou2012 observes some defining characteristics of QIEs which cannot find direct equivalents in En-

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 68-75, Hakodate, Japan, September 13-15, 2019

glish: they consist of four syllables or logographs, have relatively fixed structure and patterns, and carry figurative meaning and semantic opacity. Prefabs-QIEs, on the other hands, are more compositionally dependent, direct results of language use. They are mainly conventional combination of four morphemes taken up and reproduced by speakers they heard before. It can be understood as the *variations-tolerant* lexical bundles composing of four characters/morphemes (e.g., 好久不見 hao2 jiu3 bu2 jian4 'long time no see').

3 QIEs model of idiomaticity

This section introduces our proposed computational model of idiomaticity for Chinese QIEs. The model is based upon idiomaticity theories in linguistics and leveraged resources in Chinese Wordnet (CWN).

Idioms are complex linguistic and psychological configurations. Researchers proposed various theories and frameworks to describe aspects of linguistic construct and processing of idioms (Healy, 1994; Fernando, 1996), where different definitory dimensions are used to capture the nature of idiomaticity (Langlotz, 2006). Basing on previous literature, this paper described idiomaticity of Chinese QIEs along three dimensions:

1. **Variation** indicates the degree of conventionalization of QIEs. Idioms have gone through a socio-linguistic process through which the speakers became familiar and conventionalize the expression. The resulting construct became unitized (Healy, 1994) or frozen ("recalcitrance to undergo transformations") (Fraser, 1970). That is, the constituents of a QIE cannot be replaced or altered in actual usage.

2. **Compoundness** denotes the degree of idiosyncrasies in QIEs' compound structure. Past studies argued English idioms showed constructional idiosyncrasies, such as *trip the heavy fantastic*, which is otherwise ungrammatical (Langlotz, 2006). Similarly, Chinese idioms etymologically came from classical Chinese, their morphology and grammatical rules are different from contemporary Mandarin Chinese when compounding single-character words into QIEs.

3. **Compositeness** represents the extent of semantic un-compositionality, namely *opaqueness*, of QIEs. The uncompositional nature is the defining feature of idiom, that is the meaning of the idioms is not the compositional results of their constituent parts. Therefore two levels of meanings are to be distinguished: the literal meaning (the sum of constituent meanings) and the idiomatic meaning (the lexicalized meaning of the idiom). The more distinct these two levels of meanings of an idiom has, the more *opaque* an idiom is.

The computational model of idiomaticity formalized variation, compoundness, and compositeness with three indices respectively. These indices not only shed light on the nature of any given QIEs, but facilitate QIE candidates selection when incorporating QIEs into CWN.

3.1 Variation

Variation measures the extent of lexicogrammatically restriction of a QIE. The restriction is operationalized as usage variation frequency of a QIE in a corpus. These variations were further defined as two types: (1) substitution, where the second or the third character of a QIE was replaced by another character; and (2) insertion, where characters were placed between the spaces of the four characters in a QIE. The frequency of these variation patterns were identified and summed together, along with the frequency of the QIE itself, the index of variation can be computed as:

$$\text{variation} = \log \frac{\text{variations frequency} + 1}{\text{QIE frequency}} \qquad (1)$$

Higher variation values indicate more substitution or insertion patterns could be found for a given QIE, therefore the QIE is less likely to be *frozen*. For example, 狂風驟雨 kuáng fēng zòu yǔ "raining cats and dogs" is less conventionalized (variation $= 1.58$), since it is frequently found with the third character replaced with 暴 bào "fiercely" without changing the meaning. On the contrary, a low variation value implied a QIE is more likely to be conventionalized, thus fewer variation can be observed in corpus. For instance, 刮目相看 guā mù xiāng kàn "revere with

respect" has no variation form observed in corpus (variation $= -5.73$).

3.2 Compoundness

Compoundness indicates how probable the compound structure of a QIE follows morphological or grammatical rules in contemporary Mandarin Chinese. For instance, the idiom 虎頭蛇尾 hǔ tóu shé wěi "working industriously at first but carelessly in the end", followed common morphological rules in Mandarin Chinese. The first two characters 虎頭 literally mean "the head of a tiger", and the last two characters 蛇尾 is "the tail of a snake". The two parts both follow the same common, or probable, compound structure in Chinese word morphology. By contrast, the idiom 來龍去脈 lái lóng qù mài "the preceding and succeeding contexts of a subject matter" does not follow a common Chinese word compound rule. The first character 來 is often used as an adverb, but it seldom precedes a noun such as 龍 lóng "dragon". Similarly, the third character 去, which has comparable grammatical role as 來, is not commonly followed by a noun 脈 mài "context". Therefore this idiom has less probable compound structure.

To capture the common morphological rules in Chinese words, we constructed a morphological graph between Chinese words and characters. The graph incorporated the productive word morphology in Chinese, along with lexical and semantic relations encoded in CWN. From the morphological graph, we computed the embeddings of each character nodes (Grover and Leskovec, 2016), basing on which we devised a probability index to signify the compoundness of a QIE.

3.2.1 Morphological graph

The purpose of the morphological graph was to represent (1) the morphological relations between Chinese words and characters, and (2) the lexical and semantic relations between these words. The graph included only single- or two-character words, in the consideration that (1) Chinese words are predominantly bi-syllabic, words with one or two characters already account for 93.39% of word frequencies in a corpus; and (2) words longer than two characters potentially contaminated the graph with QIE compound information when modeling QIE compound structure.

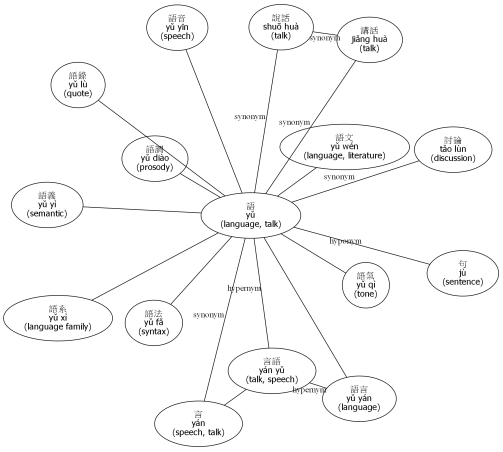

Figure 1: A sample morphological graph including 語 (yǔ , 'language, talk') and its immediate neighbors

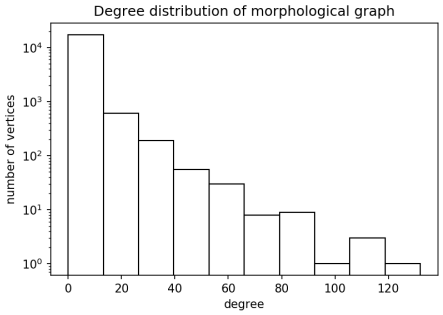

Figure 2: Degree distribution of the morphological graph.

The morphological graph was constructed from CWN. It had 18,251 vertices (including Chinese characters and two-character words) and 27,932 undirected edges (including morphological relations and CWN relations). Among the edges were 13,480 morphological relations, where characters were linked with their composed words. A sample graph was shown in 1. The degree distribution of the graph was shown in 2. There were 87% of vertices having 5 or fewer neighbors, while the most connected 20 vertices accounted for 70% connections in the graph.

Word morphology and semantic relations encoded in the graph allowed us to investigate the relations between characters, even ones not explicitly encoded in the graph. To efficiently explore the relations between characters in the graph, we computed a latent, low-dimensional node embeddings representation in

the graph. The index of compoundness was then defined basing on the embedding vectors.

3.2.2 Morphological vectors of Chinese characters

We used node2vec (Grover and Leskovec, 2016) to compute vector representations for each of the nodes in morphological graph. node2vec found a mapping $f : V \rightarrow \mathbb{R}^d$ from each vertices to a vector representation, and the mapping was optimized to maximize the log-probability of observing its neighbors in the graph given the vector. The mapping f was defined as:

$$\max_f \sum_{c \in \mathbb{C}} \log p(N(c)|f(c)) \tag{2}$$

where c is each of characters, $\mathbb{C}$, in the graph, and $N(c)$ denoted the neighbors of the character c in the graph. node2vec provided parameters to fine tune the random walk strategies when learning latent representations. In order to stress the homophily among characters, we chose $p = 2$ and $q = 0.5$ as random walk parameters. Since the probability of compounding would be evaluated on the character level, only embedding vectors of single characters were considered in following steps. We defined these vectors of characters as morphological vectors, $\mu_i = f(c_i)$, where subscript i denoted each character in the morphological graph.

Basing on morphological vectors μ_i, we first defined the compoundness of two characters as a conditional probability observing the second character given the first character. The conditional probability is based on the cosine similarity among morphological vectors, normalized to a categorical distribution with the softmax function, which could be formulated as follows:

$$p(c_2 \mid c_1) = \frac{\exp\left(\phi(\mu_1, \mu_2)\right)}{\sum_{i \in \mathbb{C}} \exp\left(\phi(\mu_1, \mu_i)\right)} \tag{3}$$

where $\phi(x, y)$ was cosine similarities between two vectors, and $\mathbb{C}$ denoted all characters in the morphological graph.

The compoundness of a QIE was defined through conditional probabilities. We assumed a linear dependency structure within QIE, that is, each character only dependent on its immediate predecessor. The compoundness of QIE then factored into a series of conditional probabilities between neighboring characters:

$$\begin{aligned} \text{compoundness} &= \log p(c_1, c_2, c_3, c_4) \\ &= \log p(\mu_1) p(\mu_2 \mid \mu_1) \\ &\quad\; p(\mu_3 \mid \mu_2)\, p(\mu_4 \mu_3) \end{aligned} \tag{4}$$

where μ_1, μ_2, μ_3, μ_4 denoted four morphological vectors of characters in the QIE. Higher probability signified stronger compoundness, i.e., the QIE followed a more common compound rules, such as the idiom 虎頭蛇尾 (compoundness $= -22.59$). Lower probabilities signified low compoundness, i.e. the QIE followed less common compound patterns, such as the idiom 來龍去脈 (compoundness $= -24.06$).

3.3 Compositeness

Semantic non-compositionality, or opaqueness, is the defining feature of idiomaticity. For example, 滄海桑田, cāng hǎi sāng tián, "drastic change of circumstances over time" is an opaque idiom. Each of its constituent characters: 滄, cāng, "blue", 海, hǎi, "ocean", 桑, sāng, "mulberry", 田, tián, "farm" bears no indication of the idiomatic meaning. As opposed to a more *transparent* idiom, 盡善盡美, jìn shàn jìn měi, "as perfect as possible" is more related to its constituents' meanings: 盡, jìn, "try to" 善, shàn, "good", 美, měi, "beauty".

In order to model compositeness, this paper took advantage of recent development of contextualized embeddings models, and the example sentences in CWN as a sense disambiguated lexical resources. We first constructed *sense vectors* from contextualized embeddings, and upon which we formalized idiomatic meaning and literal meaning of QIEs.

3.3.1 Idiomatic meaning of QIEs

Vector semantics received wide attentions in recent years, especially word embedding models such as word2vec (Mikolov et al., 2013). However, models of word semantics represented word meaning on lemma levels, which conflated different senses of a single word form (Camacho-Collados and Pilehvar, 2018). Recent advancement of contextualized embeddings, such as BERT model (Devlin et al., 2018) used a cloze task in training, allowing model to encode sentential contexts of the target word. Previ-

ous studies demonstrated that these contextualized embeddings, when combined with a set of disambiguated sense example sentences from CWN, resulted in sense vectors which can serve as a representation of CWN senses. These sense vectors, guided by linguistics constraints, help lexicographers find potential semantic relations in CWN (Tseng and Hsieh,).

Following the proposal of sense vectors, and the fact QIEs are predominately monosemic, we defined idiomatic meaning of a QIE as its contextualized embedding in the sentences. That is, a sense vector of QIE, σ_q, can be estimated by sampling sentences it occurred in, which can be formulated as an expectation over a set of sentences:

$$\sigma_q = \mathop{\mathbb{E}}_{\mathbf{w}\in\mathbb{W}} \left[\mathrm{CE}(\mathbf{w}) \cdot \mathrm{I_q}(\mathbf{w})\right] \tag{5}$$

where **w** was the list of words in a sentence, which was sampled from all the sentences the target QIE q occurred in, $\mathbb{W}$. $\mathrm{CE}(\mathbf{w})$ denoted the contextualized embeddings of the sentences, and $\mathrm{I_{target}}(\mathbf{w})$ was the indicator function to select out the embeddings of the target QIE.

In contrast of idiomatic meaning, the formalization of literal meanings was complicated by the fact most of the Chinese characters are polysemous (or homonymic). That is, to construct the sense vectors of a literal meaning, the character senses from which the literal meaning were composed should be first independently determined.

3.3.2 Literal meanings of QIEs

The task of determining character senses participated in QIE literal meanings, can be framed as finding the most probable sequence of sense composition. This view drew support from the semantic description view of Chinese word morphology, which argued meaning of the whole word came from the meaning of its constituent parts (Packard, 2000). That is, the compositionality of different senses should manifest itself on how surface word form compound to each other. In other words, the morphological vector space constructed in 3.2 could be regarded as an approximate estimate of sense composition space. The sense composition could be estimated by first projecting the sense vector into morphological vector space with projection matrix P:

$$\mathbf{P} = (\mathbf{S}^\intercal\mathbf{S})^{-1}\,\mathbf{S}^\intercal\mathbf{M} \tag{6}$$

where **M** was the morphological matrix with its rows being morphological vectors of each character, and **S** was the matrix with its rows being first sense vectors of each character in CWN (basing on the heuristic that the first sense of each character was the most frequently used sense). After obtaining the projection matrix P, we defined a function g mapping from sense vectors σ_{x_i} to an estimated morphological vector $\hat{\mu}_{x_i}$:

$$\hat{\mu}_{\mathrm{x_i}} = g(\sigma_{x_i}) = \mathbf{P}\cdot\sigma_{\mathrm{x_i}} \tag{7}$$

The joint probability of a given sense assignment can be computed based on the projected vector $\hat{\mu}_{\mathrm{x_i}}$, as defined in compoundness:

$$\begin{aligned} p(x_1,x_2,x_3,x_4) =& p(\hat{\mu}_{\mathrm{x_1}})p(\hat{\mu}_{\mathrm{x_2}}|\hat{\mu}_{\mathrm{x_1}}) \\ & p(\hat{\mu}_{\mathrm{x_3}}|\hat{\mu}_{\mathrm{x_2}})p(\hat{\mu}_{\mathrm{x_4}}|\hat{\mu}_{\mathrm{x_3}}) \end{aligned} \tag{8}$$

The most probable sense sequence in a given QIE q is then the sense assignment, $\mathbf{x}_q = (x_1, x_2, x_3, x_4)$ that maximize the joint probability:

$$\mathbf{x}_q = \mathop{\mathrm{argmax}}_{\mathrm{x}\in\mathrm{S}(q)} p(x_1,x_2,x_3,x_4) \tag{9}$$

where $\mathrm{S}(q)$ denotes all possible sense assignments in the given QIE q.

Basing on the probability, the most probable sense sequence $\mathbf{x}_q$ can then be decoded with beam search.

Equipped with the most probable sense sequence decoded in QIE, we defined index of compositeness as sum of (square root) distances between each of character sense vectors (literal meaning) and the QIE sense vector (idiomatic meaning). The index was calculated by:

$$\text{compositioness} = \sum_{x_i\in\mathbf{x}(q)} \frac{\sqrt{\|\sigma_{x_i} - \sigma_q\|^2}}{d} \tag{10}$$

where d was the dimension of sense vectors. Higher compositeness indicated literal meanings further away from idiomatic meaning, i.e., QIE was more opaque, such as the idiom 滄海桑田

(compositeness = 0.0997). Lower compositeness indicated literal meanings closer to idiomatic meaning, i.e. the QIE was more transparent, such as the idiom 盡善盡美 (compositeness = 0.0892).

4 Experiment

We presented two experiments, where three dimensions in model of idiomaticity were used as features to classify idioms and proper nouns from general QIEs.[1]

4.1 Idiom classifications

The purpose of this experiment was to illustrate the nature of QIEs, including prefabs and idioms. While Chinese idioms themselves were not a homogeneous class of linguist construct, prefabs, as a dynamic phenomena of language usage, should exhibit more variant behaviors with respect of variation, compoundness, and compositeness.

The experiment analyzed QIEs in a corpus of 1.2 billion characters, which included texts from news and online forum. In the corpus, we first extracted 319,201 quadgrams that occurred more than 32 times. Among these quadgrams, we selected 2,478 different prefabs that (1) were frequently occurred in the corpus, (2) has high PMI score (i.e. the four characters did not collocate by chance), and (3) did not frequently occurred in a fixed five-grams. Along with the prefabs, we referenced the idioms dictionary from Ministry of Education, Taiwan (MOE) to select a list of idioms as analyzing materials. Among 5,106 idioms included in the dictionary, we only included 1,518 idioms occurred more than 50 times.

Each of these 3,996 prefabs and idioms were computed for three features in model of idiomaticity. These three features were then used as classifying features in a gaussian-kernel SVM. The results classification was evaluated with a 5-fold cross validation, with mean accuracy of 70.80%, *SD* = 0.0146. Due to unequal number of prefabs and idioms, the random baseline was 62.01%.

Features distribution were shown in 3. In index of variation, the mean of idioms(*M* = -2.69, *SD* = 1.49) was lower than prefabs (*M* = -1.48, *SD* = 1.51), which was consistent to the observation that idioms were more conventionalized, therefore more resistant to usage variation. The compoundness distribution of prefabs (*M* = -23.25, *SD* = 0.37) had higher value than ones in idioms (*M* = -23.39, *SD* = 0.30), and exhibited fatter tail. It was consistent to the idea that idioms, comes from classic Chinese, followed a less common compound rule. However, compositeness distribution of prefabs and idioms showed greater overlap, and values of prefabs (*M* = 0.095, *SD* = 1.98e−3) were slightly higher than idioms (*M* = 0.094, *SD* = 2.02e−3).

One possible reason for higher compositeness (hence more opaque) of prefabs was some of which were proper nouns, such as names of locations or person. Since these proper nouns were often translated names, the characters meaning has no relations to the names they referring to, i.e., they are more opaque. Therefore, Proper nouns would serve as a clear materials to test the model of idiomaticity. Specifically, proper nouns should be opaque (high in composite index), and they would not follow morphological rules (hence low in compoundness index), and cannot allowed variations (low on variation index).

To test the hypothesis above, we conducted another experiment with proper nouns and other general prefabs.

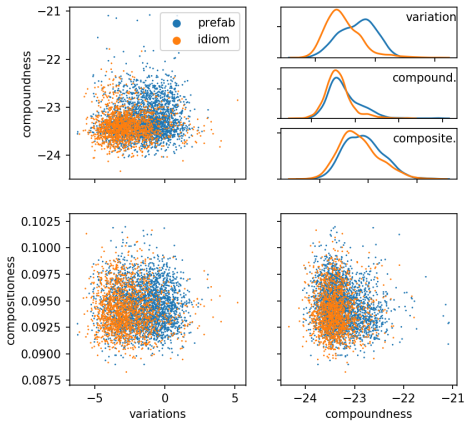

Figure 3: Distribution and scatter plots of three features in idioms and prefabs

[1] We intend to make code publicly available via github after the reviewing process.

4.2 Proper noun classification

This experiment was aimed to investigate the proper nouns with model of idiomaticity. The proper nouns were manually identified from the prefabs used in previous experiment. There were 108 proper nouns selected for this experiment, which were largely translated person names (e.g., 哈利波特, hā lì bō tè, "Harry Potter"), or locations (e.g., 巴基斯坦, bā jī sī tǎn, "Pakistan"). We randomly selected another 108 items (none of them were proper nouns) as general prefabs.

A proper noun classification task was performed and the results classification was also evaluated with a 5-fold cross validation. The mean accuracy was 71.29%, *SD* = 6.19%. The chance (baseline) level was 50.0%.

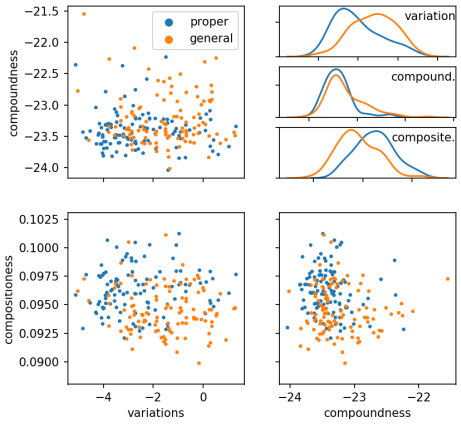

Figure 4: Distribution and scatter plots of three features in proper nouns and general prefabs.

4 showed the feature distribution of proper nouns and general prefabs. The overall patterns were consistent with the prior hypothesis. In index of variation, proper nouns (*M* = -2.54, *SD* = 0.14) were less likely to have variation forms, compared to general prefabs (*M* = -1.52, *SD* = 0.13). Proper nouns also had lower compoundness (*M* = -23.40, *SD* = 0.029) than general ones (*M* = -23.24, *SD* = 0.039). Compositeness showed a clear difference between proper nouns (*M* = 0.096, *SD* = 1.89e−4) and general prefabs (*M* = 0.095, *SD* = 1.95e−4).

The results of these two experiments demonstrated model of idiomaticity can be useful to shed light on properties, namely the variation, compoundness, and compositeness of Chinese QIEs.

4.3 Encoding QIEs in CWN

2478 QIEs-prefabs and 1518 idiom-QIEs are explored in this study. In considering the inclusion of wordnet , Idiom-QIEs are excluded, as they are well-studied in Chinese lexicography. What interests us more is the prefabs-QIEs and how we encode them into the organization of Chinese Wordnet.

We select top 200 prefabs-QIEs for manually clustering and determining their mapping to the current synsets with possible relations. Among these 200 QIEs, 109 QIEs could be justified as established concepts to incorporate into CWN (e.g., 移送法辦 'bring to justice'), 48 QIEs are more likely quasi-compounds with high frequency (e.g., 競選總部 'campaign headquarter'), and 43 QIEs are hard to be mapped into CWN as they carry mainly the pragmatic/discourse meaning (換句話說 'in other words').

5 Conclusion

In this paper, we demonstrate a proposed approach in modeling the idiomaticity of a special yet recurrent type of idiomatic expressions called QIE in Chinese. In contrast with English idioms, Chinese QIEs are different in that they are phonologically composed of four syllables, syntactically fixed structure, and semantically intransparent. Three dimensions are considered in modeling QIE's behaviour, and two classification experiments are conducted to test the model. In addition, the consequences of encoding QIEs in Chinese Wordnet is discussed.

References

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788, December.

Wallace Chafe. 1968. Idiomaticity as an anomaly in the chomskyan paradigm. *Foundations of Language*, 4:109–127.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

C. Fernando. 1996. *Idioms and Idiomaticity*. Describing English language. Oxford University Press.

Bruce Fraser. 1970. Idioms within a transformational grammar. *Foundations of Language*, 6(1):22–42.

Kristina Geeraert, John Newman, and R Harald Baayen. 2017. Idiom variation: Experimental data and a blueprint of a computational model. *Topics in cognitive science*, 9(3):653–669.

Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA. ACM.

Alice F. Healy. 1994. Letter detection: A window to unitization and other cognitive processes in reading text. *Psychonomic Bulletin & Review*, 1(3):333–344, Sep.

Shu-Kai Hsieh, Chiung-Yu Chiang, Yu-Hsiang Tseng, Bo-Ya Wang, Tai-Li Chou, and Chia-Lin Lee. 2017. Entrenchment and creativity in chinese quadrasyllabic idiomatic expressions. In *Workshop on Chinese Lexical Semantics*, pages 576–585. Springer.

A. Langlotz. 2006. *Idiomatic Creativity: A Cognitive-linguistic Model of Idiom-representation and Idiom-variation in English*. Human cognitive processing. J. Benjamins.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

J.L. Packard. 2000. *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press.

Ivan A. Sag, Baldwin Timothy, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15.

Y. H. Tseng and S. K. Hsieh. Augmenting chinese wordnet semantic relations with contextualized embeddings. submitted.

Benjamin K Tsou. 2012. Idiomaticity and classical traditions in some east asian languages. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 39–55.