IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 26, 2019, accepted September 27, 2019, date of publication September 30, 2019, date of current version October 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944572

Idiomaticity Prediction of Chinese Noun Compounds and Its Applications

CHENGYU WANG[1], YAN FAN[2], XIAOFENG HE[3], (Member, IEEE), HONGYUAN ZHA[4], AND AOYING ZHOU[5], (Member, IEEE)
[1] School of Software Engineering, East China Normal University, Shanghai 200062, China
[2] Alibaba Group, Hangzhou 311121, China
[3] School of Computer Science and Technology, East China Normal University, Shanghai 200062, China
[4] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[5] School of Data Science and Engineering, East China Normal University, Shanghai 200062, China

Corresponding author: Xiaofeng He (xfhe@sei.ecnu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1000904.

ABSTRACT Idiomaticity refers to the situation where the meaning of a lexical unit cannot be derived from the usual meanings of its constituents. As a ubiquitous phenomenon in languages, the existence of idioms often causes significant challenges for semantic NLP tasks. While previous research mostly focuses on the idiomatic usage detection of English verb-noun combinations and the semantic analysis of Noun Compounds (NCs), the idiomaticity issues of Chinese NCs have been rarely studied. In this work, we aim at classifying Chinese NCs into four idiomaticity degrees. Each idiomaticity degree refers to a specific paradigm of how the NCs should be interpreted. To address this task, a Relational and Compositional Representation Learning model (RCRL) is proposed, which considers the relational textual patterns and the compositionality levels of Chinese NCs. RCRL learns relational representations of NCs to capture the semantic relations between two nouns within an NC, expressed by textual patterns and their statistical signals in the corpus. It further employs compositional representations to model the compositionality levels of NCs via network embeddings. Both loss functions of idiomaticity degree classification and representation learning are jointly optimized in an integrated neural network. Experiments over two datasets illustrate the effectiveness of RCRL, outperforming state-of-the-art approaches. Three applicational studies are further conducted to show the usefulness of RCRL and the roles of idiomaticity prediction of Chinese NCs in the fields of NLP.

INDEX TERMS Representation learning, idiomaticity prediction, noun compound, relational pattern, compositionality analysis.

I. INTRODUCTION

Idiomaticity is ubiquitous in natural languages. It refers to the phenomenon where the meaning of a lexical unit is unpredictable from the usual meanings of its individual constituents [1]. Typical examples of idioms include ‘‘cloud nine’’, ‘‘white coal’’, ‘‘kick the bucket’’, etc. As a class of Multiword Expressions (MWEs), the presence of idioms often changes the default meanings of natural languages. Hence, such phenomenon has been regarded as ‘‘a pain in the neck’’ in Natural Language Processing (NLP) for decades [2], [3]. For example, the performance of machine translation drops significantly in an idiom-rich corpus [4].

Due to its prevalence, the detection and processing of idiomatic languages are key research areas in NLP and computational linguistics. For several NLP tasks (e.g., analysis of verb semantics [5]), idioms often need to be processed separately. In previous research, idiom token classification is an NLP task of distinguishing idiomatic MWEs and expressions with literal meanings. A majority of related research focuses on English Verb-Noun Combinations (VNCs) (e.g., ‘‘push one’s luck’’, ‘‘blow the whistle’’) [6]–[9]. The goal of this task is to classify VNC usages in sentences as idiomatic or literal. Another closely related task is the compositionality prediction of Noun Compounds (NCs). This is because the compositionality levels of NCs have strong correlations with their idiomaticity degrees [10]–[13]. For example, the noun phrase ‘‘apple tree’’ is decomposable, as its meaning can be inferred

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

by combining the meanings of the two nouns (''trees where apples grow''). In contrast, the idiomatic NC ''cloud nine'' is indecomposable, since the meaning of this compound is unrelated to the meanings of ''cloud'' and ''nine'', separately.

Despite the success, existing research works mostly focus on the English language and are not sufficient to process idiomatic Chinese NCs. The reasons are briefly stated below:

- Chinese is a highly idiomatic language, containing a large portion of metaphorical expressions [14]. The ''hidden relations'' within Chinese NCs are mostly commonsense to native speakers and are often omitted in texts. Hence, pattern-based methods for English (e.g., [9]) are difficult to apply to analyze and interpret the meanings of Chinese NCs (especially idiomatic NCs).
- The compositionality level prediction of English NCs usually relies on the morphological and structural analysis of English words (e.g., [10], [15]). In contrast, Chinese words have little morphology, using unbound morphemes and character orders to convey meanings [16]. Hence, it is difficult to apply these methods to the Chinese language without major modification.
- In existing datasets related to the compositionality of NCs, an NC is associated with a numeral score, indicating a scale of literality [13], [17]. However, this labeling scheme only implies how idiomatic the meanings of NCs are, instead of providing an explicit, classification-based framework for machines to understand and interpret the meanings of NCs directly.

In this paper, we propose a novel, neural computational framework to predict the idiomaticity degrees of Chinese NCs based on representation learning.[1] Different from existing numerical score prediction tasks for compositionality analysis of NCs [10]–[13], our work aims at classifying a Chinese NC into one of the four idiomaticity degrees introduced by Wang and Wang [18]. Each idiomaticity degree corresponds to one specific paradigm to interpret the NC.[2] Hence, the relations between idiomaticity analysis and natural language understanding of Chinese NCs can be established. The four idiomaticity degrees w.r.t. a noun-noun compound N_1N_2 are briefly summarized as follows, with an increasing level of idiomaticity:

1) **Transparent** (N_1 describes a property of N_2 explicitly.)
2) **Partly opaque** (N_1 is not a property of N_2. Instead, there exists a hidden, implicit verbal relation between N_1 and N_2.)
3) **Partly idiomatic** (No direct relation exists between N_1 and N_2. N_1 modifies N_2 metaphorically, not literal.)
4) **Completely idiomatic** (N_1 and N_2 are completely indecomposable and should be treated as a whole unit whose semantic meaning is not explicitly related to N_1 or N_2.)

To address this task, a Relational and Compositional Representation Learning framework (RCRL) is proposed. RCRL learns relational representations of NCs. Such representations encode the hidden semantic relations between two nouns within an NC via several specially designed pattern-based features for the Chinese language. As idiomaticity degrees of NCs are closely related to compositionality [11], [12], the RCRL model also learns compositionality representations of NCs efficiently via network embeddings [20], so that NCs with similar compositionality levels have similar representations. Both the representation learning and the idiomaticity prediction tasks are jointly optimized in an integrated neural network via multi-task iterative learning.

In the experiments, we evaluate the RCRL model over two Chinese NC datasets and compare it with strong baselines. The experimental results show that RCRL consistently outperforms these baseline methods. Based on the application of the RCRL model, we further conduct three studies, with the findings summarized as follows. i) We detect idiomatic NCs in a large-scale Chinese Web corpus. The results reveal that 50.8% of the Chinese NCs have idiomatic meanings to some degree. ii) We find that the idiomaticity degrees of NCs affect the machine translation quality significantly. A higher degree of idiomaticity is associated with poorer machine translation quality. iii) We show that the RCRL model is not entirely language-dependent. With simple modification, the RCRL model can also achieve state-of-the-art performance for the English language over the compositionality analysis task.

The rest of this paper is organized as follows. Section II introduces the background of this research. The RCRL model is described in Section III in detail. The experimental results are shown in Section IV. Finally, we present the three studies in Section V and draw the conclusion in Section VI.

II. BACKGROUND

In this section, we introduce our research background and discuss the idiomaticity issues of Chinese NCs.

A. A SUMMARY OF RELATED WORK

1) NC SEMANTICS

NC is a class of MWEs [3], which is a fixed expression consisting of multiple nouns. The semantics of NCs (particularly noun-noun compounds) are typically expressed by abstract verbal relations from a fixed inventory [21], [22]. For example, the semantics of ''olive oil'' can be expressed by the verbal relation ''made-of'' (i.e., ''oil'' that is made of ''olive''). However, a finite set of relations (or verbs) are insufficient to represent complicated NC relations. For finer-grained representations, Cruys *et al.* [23] use multiple paraphrases involving verbs and/or prepositions to express the semantics of NCs. In SemEval-2013 Task 4 [24], participants are allowed to use

[1] We select Chinese as the target language in this work due to its high level of idiomaticity [14]. However, our method is not completely language-dependent. Please refer to the study on how to apply our method to the English language in Section V for details.

[2] In this work, we specifically focus on noun-noun compounds (denoted as N_1N_2), consisting of one modifier (N_1) and one head word (N_2) [19], because noun-noun compounds are most common among NCs. Longer NCs can also be processed similarly by parsing and bracketing the NCs beforehand.

free paraphrases to represent relations within NCs. The above works deal with English NCs only. For Chinese, the study of NC semantics is insufficient. In [18], the authors present a taxonomy on semantics of Chinese noun-noun compounds. We employ the four idiomaticity degrees in [18] to represent the semantics of Chinese NCs. Each idiomaticity degree can be viewed as a specific paradigm to interpret Chinese NCs.

2) IDIOM TOKEN CLASSIFICATION

The detection of idiomatic tokens is vital for processing MWEs, as it recognizes idiomatic language usages in free texts. The research of idiom token classification dates back to [25], which uses an SVM classifier to distinguish whether a Japanese phrase is idiomatic or literal. Peng *et al.* [7] assume that contexts of idioms are usually different from words in local topics and detect such idioms based on topic models. Recently, the usage of word embeddings has been extensively applied to the task of idiom token classification. For example, Salton *et al.* [8] use sentences containing a target phrase as inputs, and classify the phrase as idiomatic or literal based on distributional representations of the corresponding words. Gharbieh *et al.* [26] show that by using word embeddings as features, both unsupervised and supervised models for idiomatic token classification outperform traditional methods. King and Cook [9] further improve the performance of word embedding-based methods by considering lexico-syntactic linguistic knowledge. A similar linguistically motivated work is proposed by Liu and Hwa [27], which propose a literal usage metric to measure the probability of a certain idiom is intended literally in the text corpus. A potential drawback is that they deal with idiomatic VNCs, leaving idiomatic NCs and other similarly structured MWEs unexplored.

3) COMPOSITIONALITY OF NCS

The compositionality analysis of NCs is also closely related to our research. This is because the meanings of idiomatic NCs are usually different from its components. Hence, these NCs tend to be indecomposable. Early attempts devise a number of measures to describe the levels of compositionality based on vector space representations of terms. Notable approaches in this field include [13], [28] and many others. With the wide application of neural language models, word embeddings have been extensively exploited for compositionality prediction. For example, Salehi *et al.* [29] combine word embedding techniques and previous compositionality measures proposed by Reddy *et al.* [13] to detect indecomposable NCs. Yazdani *et al.* [12] and Cordeiro *et al.* [11] present a range of distributional semantic models to learn the non-compositionality of MWEs. By comparing the difference between the compound embeddings of NCs and the individual embeddings of the two component nouns, the compositionality degrees of NCs can be measured, in the form of real-value scores. For Chinese, the research work is highly insufficient. Qi *et al.* [30] incorporate Chinese sememe knowledge into compositionality prediction models, and learn the semantic representations of Chinese multiword expressions based on sememes. However, this work is potential restricted by the coverage of sememes. In this work, we solve the problem of semantic compositionality from another aspect. We classify Chinese NCs into a fine-grained taxonomy of idiomaticity degrees, which does not rely on existing sememe knowledge bases and is more suitable for the natural language understanding of the Chinese language.

B. IDIOMATICITY OF CHINESE NCS

As discussed, the semantics of Chinese NCs are not sufficiently studied. In linguistics, "semantic transparency" is a property of NCs. It describes to which extent an NC retains its literal meaning in its actual meaning [31]. Hence, higher transparency in NCs also means lower idiomaticity.

In this work, we consider a taxonomy of semantic transparency of Chinese NCs [18] as the standard to characterize the idiomaticity degrees of Chinese NCs. It has four levels of semantic transparency, with increasing idiomaticity degrees. In the following, we summarize the four classes briefly, with examples presented in Table 1[3]:

1) **Transparent**: N_1 modifies N_2 explicitly, describing a property/attribute of N_2. It also means N_1N_2 is a type of N_2. For example, "固体(solid)" is a physical property of "燃料(fuel)" in the NC "固体燃料(solid fuel)".
2) **Partly Opaque**: N_1 does not directly modify N_2. Instead, there exists an implicit verbal relation between N_1 and N_2. Again, N_1N_2 is a type of N_2. Consider the NC "办公用品(office supplies)". "办公(office)" is not a property of "用品(supplies)", but refers a kind of "用品(supplies)" that are used in "办公(office)". In this case, "used-in" is the verbal relation between them.
3) **Partly Idiomatic**: N_1 and N_2 are decomposable but the usage of N_1 is idiomatic, not literal. The relation between N_1 and N_2 is metaphorical. Hence, there is no direct relation between N_1 and N_2. However, we still can infer N_1N_2 is a type of N_2. In "计划经济(planned economy)", the economic system is not about plans themselves. Instead, "计划(plan)" refers to the most important characteristics in the system where the allocation of goods and other resources is managed by plans made by governments.
4) **Completely Idiomatic**: N_1N_2 is completely indecomposable, referring to a concept that is not a type of N_1 or N_2. For example, "夫妻肺片(Mr and Mrs Smith)" is the name of a specific Chinese beef dish. The NC is neither a type of "夫妻(couple)" nor "肺片(lung piece)".

The reason for choosing the four-value scale to model the idiomaticity degrees of Chinese NCs rather than numerical

[3] In linguistics, the term "semantic transparency" is specifically used to describe such property of NCs. The term "idiomaticity" is more frequently used in the NLP community, referring to a broader spectrum of linguistic phenomena. In this paper, because we focus on the semantics of NCs only, we do not consider the strict differences between these terms.

TABLE 1. Four levels of idiomaticity degrees of Chinese NCs. Each Chinese NC is accompanied by English literal translation and correct translation. Modifiers of Chinese NCs are underlined with linguistic heads printed in bold.

Idiomaticity Degree	Example	English Translation
Type I: Transparent	固体燃料 (Solid **Fuel**)	Solid fuel
	Interpretation: **Fuels** that *are* solid.	
	沿海地区 (Close to sea **Area**)	Coastal area
	Interpretation: **Areas** that *are* close to seas.	
Type II: Partly Opaque	办公用品 (Office **Appliance**)	Office supplies
	Interpretation: **Supplies** that *are used* in the office.	
	国家联盟 (Country **Alliance**)	Coalition of nations
	Interpretation: A **coalition** that *is formed by* nations.	
Type III: Partly Idiomatic	计划经济 (Plan **Economy**)	Planned economy
	Interpretation: An **economic system** where the allocation of resources, production, investment and pricing is performed through plans.	
	纳米技术 (Nanometer **Technology**)	Nanotechnology
	Interpretation: A type of science and **technology** that manipulates matter on an atomic or molecular scale.	
Type IV: Completely Idiomatic	夫妻肺片 (Husband and wife **Lung piece**)	Mr and Mrs Smith (Sliced beef and ox tongue in Chilli sauce)
	Interpretation: Paraphrased as sliced beef and ox tongue in Chilli sauce, a Sichuan dish invented by a couple in the 1930s. The dish was said to contain lung pieces by mischievous children.	
	意识形态 (Consciousness **Character**)	Ideology
	Interpretation: A set of normative beliefs, conscious and unconscious ideas (especially political beliefs) held by people.	

scores is as follows. Numeral compositionality scores for NCs (e.g., [11], [17]) may lack explicit relational modeling between such scores and other NLP tasks. For instance, if the score of “计划经济(planned economy)” is predicted to be 0.7/1.0, we only know that this NC is idiomatic to some extent. It is still unclear how the meanings of the NC should be interpreted. In contrast, by adopting the framework [18], connections between NC semantics and other NLP tasks can be established, including noun phrase interpretation [32], hypernym generation [33], etc. For example, if “固体燃料(solid fuel)” is transparent, the two semantic relations can be directed generated:

(固体燃料, 具有属性, 固体) (solid fuel, has-property, solid)
(固体燃料, 属于, 燃料) (solid fuel, is-a, fuel)

For “办公用品(office supplies)”, we can also extract the two relations:

(办公用品, 用于, 办公) (office supplies, used-in, office)
(办公用品, 属于, 用品) (office supplies, is-a, supplies)

where the relation predicate “used-in” can be inferred via noun phrase interpretation techniques [32]. As for “夫妻肺片(Mr and Mrs Smith, sliced beef and ox tongue in Chilli sauce)”, we can infer that the previously mentioned “has-property” and “is-a” relations are not correct in this case. In summary, the extracted relations or induced rules are particularly useful for taxonomy induction, knowledge base completion, commonsense reasoning, etc. In this paper, we restrict the scope of this work to the idiomaticity degree prediction of Chinese NCs and a few related studies. Further applications of RCRL in NLP are left as future work.

III. THE RCRL MODEL

In this section, we present the RCRL model in detail for the prediction of idiomaticity degrees for Chinese NCs.

A. AN OVERVIEW OF RCRL

A basic approach for predicting the idiomaticity degree of the NC N_1N_2 is to leverage word embeddings of N_1 and N_2. This is because the semantic relations between N_1 and N_2 inferred from word embeddings have close connections to idiomaticity degrees [11]. However, this method considers distributional representations of words only, and may suffer from the “lexical memorization” problem [34]. In the experiments, we also find that it has low accuracy for Chinese due to the ignorance of Chinese language characteristics. In this work, we learn relational and compositional representations of Chinese NCs based on the following two observations:

- **Observation 1**: Some relational textual patterns w.r.t. N_1 and N_2 in the corpus are important signals to predict the idiomaticity degree of N_1N_2.

For example, given a sentence containing a Chinese NC N_1N_2, if the head word N_2 also occurs solely in the same sentence, it is likely that the author uses other expressions containing N_2 to describe N_1N_2. Hence, the meaning of N_1N_2 is likely to be decomposable. This gives little probability that N_1N_2 is completely idiomatic. Consider the sentence: “流行音乐是一种以盈利为主要目的而创作的音乐(Pop music is a type of music created with profit as its main purpose)”. The head word of “流行音乐 (pop music)” is “音乐 (music)”, which also appears in the same sentence, besides in the NC “流行音乐 (pop music)” . Here, we may infer “流行音乐 (pop music)” is not completely idiomatic, since its meaning relates to “音乐 (music)” to some degree.

- **Observation 2**: NCs with similar compositionality levels share similar idiomaticity degrees.

For example, two NCs “solid fuel” and “liquid oil” are both decomposable. In these NCs, “solid” and “liquid” describe a property (physical form) of the objects (“fuel” and "oil"). This also implies that the two NCs are semantically transparent.

Let $y^{(i)}$ be the true idiomaticity degree of an NC $x^{(i)}$. $\tilde{y}(x^{(i)})$ is the predicted idiomaticity degree by any machine learning models. Denote L and U as the training and unlabeled sets of Chinese NCs. Based on two observations, RCRL learns relational representation $\mathbf{x}_r^{(i)}$ and compositional representation $\mathbf{x}_c^{(i)}$ for each $x^{(i)} \in D \cup U$. It minimizes the loss function defined in Eq. (1)[4]:

$$\mathcal{J} = \sum_{x^{(i)} \in L} sl(y^{(i)}, \tilde{y}(x^{(i)})) + \lambda \sum_{x^{(i)}, x^{(j)} \in L \cup U} \alpha_{i,j} ul(\tilde{y}(x^{(i)}), \tilde{y}(x^{(j)})) \qquad (1)$$

Here, $sl(y^{(i)}, \tilde{y}(x^{(i)}))$ is the supervised loss of idiomaticity degree prediction errors over the training set. The features used for idiomaticity degree classification are designed based on Observation 1. $\alpha_{i,j} = \text{sim}(\tilde{y}(x^{(i)}), \tilde{y}(x^{(j)}))$ is the compositional similarity between two NCs $x^{(i)}$ and $x^{(j)}$. $ul(\tilde{y}(x^{(i)}), \tilde{y}(x^{(j)}))$ is the unsupervised loss, forcing compositionally similar NCs to have similar idiomaticity predictions, which addresses Observation 2. λ is the balancing factor w.r.t. the two types of losses. Note that we utilize both training and testing (i.e., unlabeled) data to compute the unsupervised loss in order to exploit Observation 2 in both datasets.

B. LEARNING RELATIONAL REPRESENTATIONS

Unlike general word embeddings, the relational representation characterizes how two nouns relate to each other within an NC. To encode the knowledge of relational textual patterns (i.e., Observation 1), we define a collection of raw features $\mathcal{F}_r^{(i)}$ of an NC $x^{(i)}$ over a text corpus. The relational representations of $x^{(i)}$ can be calculated via $\mathbf{x}_r^{(i)} = \mathbf{M}_r \mathcal{F}_r^{(i)}$ where $\mathbf{M}_r$ is a linear projection matrix. Based on characteristics of different idiomaticity degrees, we design following raw features $\mathcal{F}_r^{(i)}$.

1) AUXILIARY FEATURE

In Chinese, if the pattern "N_1的N_2" exists, N_1 explicitly modifies N_2 as a property, where "的(de)" is a common Chinese auxiliary word. Hence, it is probable that N_1N_2 is transparent. Because the pattern may be expressed multiple times and contains noise, similar to [35], we define r_a as a pattern redundancy factor, typically set to a small, positive integer. To speed up text retrieval, we construct a sentence-level inverted index over the text corpus. We denote S_q^k as the collection of top-k sentences that returns for query q. Assume that "N_1的N_2" strongly indicates that N_1N_2 is *transparent* if the pattern appears at least r_a times. Using such notations, the auxiliary feature is defined as follows[5]:

$$f_{aux}(N_1, N_2) = \min\{1, \frac{1}{r_a} \sum_{s \in S_{q_{aux}}^k} I(q_{aux} \in s)\}$$

where query q_{aux} ="N_1的N_2". $I(\cdot)$ is the indicator function that returns 1 if $(\cdot)$ is true and 0 otherwise.

2) VERB FEATURE

This feature models to which degree there may exist verbs describing the relations between the two nouns. In Chinese, the detection of verbal relations suffers from low accuracy due to the flexible language expressions and the existence of light verb constructions [36]. To address the two issues, we propose a statistical approach to increase error tolerance, as shown in Algorithm 1. Let query q_{verb} be "(N_1 AND N_2) NOT N_1N_2". For each sentence $s \in S_{q_{verb}}^k$, we extract contextual verbs from s which may indicate relations between N_1 and N_2. Inspired by [37], we treat a verb v as a contextual verb of N_1 and N_2 if it is in the dependency chain or syntax path between N_1 and N_2.

Denote $V(N_1, N_2)$ as the multi-set of the contextual verb collection w.r.t. N_1 and N_2 where $c(v)$ is the count of v. Similarly, let r_v as a verb redundancy factor (i.e., a small, positive integer). $V_{r_v}(N_1, N_2)$ is the subset of $V(N_1, N_2)$ with top-r_v frequency counts. Assume there is strong presence of relational verbs between the two nouns if there are at least r_v verbs where each verb has at least r_v frequency counts. The verb feature is computed as follows:

$$f_{verb}(N_1, N_2) = \min\{1, \frac{1}{r_v^2} \sum_{v \in V_{r_v}(N_1, N_2)} c(v)\}$$

Algorithm 1 Verb Feature Extraction Algorithm

```
Initialize multiset $V(N_1, N_2) = \emptyset$;
for each sentence $s \in S_{q_{verb}}^k$ do
    if $N_1 \in s$ and $N_2 \in s$ then
        Add contextual verbs w.r.t. $N_1N_2$ to $V(N_1, N_2)$;
    end if
end for
for each verb $v \in V(N_1, N_2)$ do
    Compute frequency count $c(v)$;
end for
Extract the verb collection $V_{r_v}(N_1, N_2)$ with top-$r_v$ counts
    from $V(N_1, N_2)$;
return Feature value $f_{verb}(N_1, N_2)$;
```

3) HEAD CO-OCCURRENCE FEATURE

As in Observation 1, the frequent co-occurrence of N_1N_2 and N_2 in the same sentences indicates that N_1N_2 is not completely idiomatic. Let r_c be the head word co-occurrence factor. q_{head} ="N_2 AND N_1N_2". $I_c(s, N_1N_2)$ is an indicator function that returns 1 iff $N_1N_2 \in s$ and $N_2 \in s \setminus \{N_1N_2\}$. The head co-occurrence feature is defined as follows[6]:

$$f_{head}(N_1, N_2) = \min\{1, \frac{1}{r_c} \sum_{s \in S_{q_{head}}^k} I_c(s, N_1N_2)\}$$

[4] For simplicity, we omit the regularization terms of the model parameters.

[5] We use $x \in y$ to represent x is the substring of y.

[6] We use $\min\{1, \cdot\}$ in all three features because: i) all the feature values are self-normalized, and ii) there is enough evidence for idiomaticity prediction if the corresponding patterns appear over certain times in the corpus.

TABLE 2. Detailed raw feature template of RCRL.

Feature Name	Mathematical Definition
Auxiliary feature	$f_{aux}(N_1, N_2) = \min\{1, \frac{1}{r_a} \sum_{s \in S^k_{q_{aux}}} I(q_{aux} \in s)\}$
Verb feature	$f_{verb}(N_1, N_2) = \min\{1, \frac{1}{r_v^2} \sum_{v \in V_{r_v}(N_1, N_2)} c(v)\}$
Head co-occurrence feature	$f_{head}(N_1, N_2) = \min\{1, \frac{1}{r_c} \sum_{s \in S^k_{q_{head}}} I_c(s, N_1 N_2)\}$
Modifier-expanded auxiliary feature	$f^m_{aux}(N_1, N_2) = \frac{1}{\tau} \sum_{n_1 \in C_p(N_1)} \text{top}(f_{aux}(n_1, N_2), \tau)$
Head-expanded auxiliary feature	$f^h_{aux}(N_1, N_2) = \frac{1}{\tau} \sum_{n_2 \in C_{p/2}(N_2)} \text{top}(f_{aux}(N_1, n_2), \tau)$
Modifier-expanded verb feature	$f^m_{verb}(N_1, N_2) = \frac{1}{\tau} \sum_{n_1 \in C_p(N_1)} \text{top}(f_{verb}(n_1, N_2), \tau)$
Head-expanded verb feature	$f^h_{verb}(N_1, N_2) = \frac{1}{\tau} \sum_{n_2 \in C_{p/2}(N_2)} \text{top}(f_{verb}(N_1, n_2), \tau)$
Modifier-expanded head co-occurrence feature	$f^m_{head}(N_1, N_2) = \frac{1}{\tau} \sum_{n_1 \in C_p(N_1)} \text{top}(f_{head}(n_1, N_2), \tau)$
Head-expanded head co-occurrence feature	$f^h_{head}(N_1, N_2) = \frac{1}{\tau} \sum_{n_2 \in C_{p/2}(N_2)} \text{top}(f_{head}(N_1, n_2), \tau)$

4) EXPANDED FEATURES

Although previous features capture linguistic characteristics of idiomaticity degrees, they are insufficient. This is because in some cases, such explicit expressions are commonsense to humans and do not frequently appear in the corpus. Recently, word embedding based query expansion techniques [38] have been applied to improve the recall of sentence retrieval. Here, we derive expanded features using query expansion. For example, the pattern "液体的燃料(liquid fuel)" gives us additional knowledge to predict "固体燃料(solid fuel)" is transparent, because "liquid" and "solid" are semantically similar. In this case, we do not even need to see "固体(solid)" and "燃料(fuel)" to co-occur in the same sentence to make a confident prediction.

Let $C_p(w)$ be the p-nearest neighbors of word w, where the semantic similarity of words is quantified by the cosine similarity of word embeddings. For the auxiliary feature, we replace N_1 with each word $w \in C_p(N_1)$ and compute feature values. The modifier-expanded auxiliary feature is:

$$f^m_{aux}(N_1, N_2) = \frac{1}{\tau} \sum_{n_1 \in C_p(N_1)} \text{top}(f_{aux}(n_1, N_2), \tau)$$

where $\text{top}(f_{aux}(n_1, N_2), \tau) = f_{aux}(n_1, N_2)$ if $f_{aux}(n_1, N_2)$ is the top-τ largest among all values $f_{aux}(\tilde{n}_1, N_2)$ ($\tilde{n}_1 \in C_p(N_1)$) and equals 0 otherwise. Hence, $f^m_{aux}(N_1, N_2)$ is the top-τ averaged feature values in the "neighborhood" of N_1. Similarly, the head-expanded auxiliary feature is:

$$f^h_{aux}(N_1, N_2) = \frac{1}{\tau} \sum_{n_2 \in C_{p/2}(N_2)} \text{top}(f_{aux}(N_1, n_2), \tau)$$

We heuristically use $\frac{p}{2}$-nearest neighbors rather than p because the change of heads affects more on meanings of NCs than modifiers. We also introduce the modifier-expanded and head-expanded features for verbs and head co-occurrences. Refer to Table 2 for the detailed feature template of RCRL. In summary, $\mathcal{F}^{(i)}_r$ is the concatenation of all the above features. The matrix $\mathbf{M}_r$ is used for generating relational representations via $\mathbf{x}^{(i)}_r = \mathbf{M}_r \mathcal{F}^{(i)}_r$ (which will be learned via the neural network introduced afterwards).

C. LEARNING COMPOSITIONAL REPRESENTATIONS

While it is relatively straightforward to derive relational representations based on textual patterns, learning compositional representations to encode Observation 2 is challenging, because word embeddings of nouns in NCs are not sufficient to characterize the meanings of NCs with idiomatic meanings [11]. We improve the work [11] for multi-way classification of idiomaticity degrees. Let two NCs $x^{(i)}$ and $x^{(j)}$ be N_1N_2 and $N_1^{'}N_2^{'}$. We define the compositional similarity score $\alpha_{i,j}$ (in Eq. (1)) between $x^{(i)}$ and $x^{(j)}$ as:

$$\alpha_{i,j} = \frac{1}{2} | \cos(\vec{v}(N_1N_2), \vec{v}(N_1 + N_2)) - \cos(\vec{v}(N_1^{'}N_2^{'}), \vec{v}(N_1^{'} + N_2^{'}))|$$

where $\vec{v}(N_1N_2)$ is the compound embedding of the NC N_1N_2, and $\vec{v}(N_1 + N_2)$ is the sum of the normalized embeddings of two nouns N_1 and N_2 separately:

$$\vec{v}(N_1 + N_2) = \frac{\vec{v}(N_1)}{\|\vec{v}(N_1)\|} + \frac{\vec{v}(N_2)}{\|\vec{v}(N_2)\|}$$

From this setting, we can see that i) $\alpha_{i,j} \in [0, 1]$ and ii) NCs with similar compositionality degrees have similar $\alpha_{i,j}$ scores.

Recall that $\mathbf{x}^{(i)}_c$ is the compositional representation of the NC $x^{(i)}$. In order to minimize the unsupervised loss $\sum_{x^{(i)}, x^{(j)} \in L \cup U} \alpha_{i,j} ul(\tilde{y}(x^{(i)}), \tilde{y}(x^{(j)}))$ in Eq. (1), we adopt the idea of graph embeddings to learn such compositional representations. Let $G(\Phi, \Psi, W)$ be a graph with edge weights, where Φ and Ψ denote the node and edge sets, respectively. Φ corresponds to all training and testing data instances (i.e., all $x^{(i)} \in L \cup U$). W is an edge weight vector that assigns a weight $w_{i,j}$ to each $(x^{(i)}, x^{(j)}) \in \Psi$. In the graph, each NC $x^{(i)} \in D \cup U$ is associated with a compositional representation $\mathbf{x}^{(i)}_c$. To ensure that compositionally similar NCs have similar compositional representations $\mathbf{x}^{(i)}_c$, we propose a variant of the DeepWalk [20] and node2vec [39] algorithms as follows.

Let $N(x^{(i)})$ be the collection of "neighbors" of NC $x^{(i)}$ in G where "neighbors" of $x^{(i)}$ are compositionally similar to $x^{(i)}$. Based on [20], [39] We re-write the unsupervised loss as the negative log likelihood function:

$$-\sum_{x^{(i)} \in L \cup U} \sum_{x^{(j)} \in N(x^{(i)})} \log \Pr(x^{(j)} | \mathbf{x}^{(i)}_c)$$

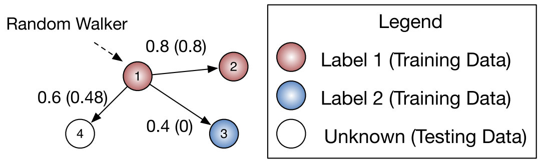


FIGURE 1. A simple example of how a random walker goes to three adjacent nodes with label information. Assume $\alpha_{1,2} = 0.8$, $\alpha_{1,3} = 0.4$, $\alpha_{1,4} = 0.6$ and $\gamma = 0.8$. Based on Eq. (2), we have $w_{1,2} = 0.8$, $w_{1,3} = 0$ and $w_{1,4} = 0.48$. The random walk probabilities can be computed as $\Pr(1 \rightarrow 2) = \frac{0.8}{0.8+0.48}$, $\Pr(1 \rightarrow 3) = 0$ and $\Pr(1 \rightarrow 4) = \frac{0.48}{0.8+0.48}$.

where $\Pr(x^{(j)}|\mathbf{x}_c^{(i)})$ is the probability of predicting $x^{(j)}$ as the "neighbor" of $x^{(i)}$ given its compositional representation $\mathbf{x}_c^{(i)}$.

Computing $\Pr(x^{(j)}|\mathbf{x}_c^{(i)})$ is infeasible due to the expensive computation cost of the partition function. Besides, the label information of NCs in the training set is not considered. We propose a Label-sensitive Weighted Random Walk (LWRW) process to sample sequences of compositionally similar NCs. Let G be a complete graph. $w_{i,j}$ is computed as follows:

$$w_{i,j} = \begin{cases} \alpha_{i,j} & x^{(i)} \in L, x^{(j)} \in L, y^{(i)} = y^{(j)} \\ 0 & x^{(i)} \in L, x^{(j)} \in L, y^{(i)} \neq y^{(j)} \\ \alpha_{i,j} \cdot \gamma & \text{Otherwise} \end{cases} \tag{2}$$

where $\gamma \in (0, 1)$ is a decay factor that gives a relatively low confidence to unlabeled data. The LWRW process assumes a random walker travels from $x^{(i)}$ to $x^{(j)}$ with probability $\propto \alpha_{i,j}$ if they have the same label, and with zero probability if they have different labels. When at least one of the labels of the NCs ($x^{(i)}$ or $x^{(j)}$) is unknown, we set the probability $\propto \alpha_{i,j}\gamma$. Refer to a simple example in Figure 1.

Denote an LWRW sequence as $\mathcal{S} = \{x^{(1)}, \cdots, x^{(|\mathcal{S}|)}\}$, and l as a window size parameter. The optimization objective is re-formulated as:

$$-\sum_{\mathcal{S}} \sum_{x^{(i)} \in \mathcal{S}} \sum_{j=i-l(j \neq i)}^{i+l} \log \Pr(x^{(j)}|\mathbf{x}_c^{(i)}) \tag{3}$$

The general learning algorithm is presented in Algorithm 2. Compositional representations $\mathbf{x}_c^{(i)}$ in both U and L are initialized randomly, with $G(\Phi, \Psi, W)$ constructed. Next, it starts a number of iterations. In each iteration, it samples an LWRW sequence $\mathcal{S}$ from the graph starting from a randomly selected node. It uses $\mathcal{S}$ to update the compositional representations with optimization details elaborated later.

D. JOINT OPTIMIZATION

Figure 2 shows the general neural network architecture to optimize Eq. (1) via multi-task learning. For each NC $x^{(i)}$, RCRL extracts raw features $\mathcal{F}_r^{(i)}$ and computes relational representation $\mathbf{x}_r^{(i)}$ by multiplying $\mathbf{M}_r$. For compositional representation, $x^{(i)}$ is mapped to $\mathbf{x}_c^{(i)}$, which is used to predict its neighbors $N(x^{(i)})$. $\mathbf{x}_r^{(i)}$ and $\mathbf{x}_c^{(i)}$ are jointly fed into a neural network. The model predicts the label $\tilde{y}(x^{(i)})$ based on the two sets of features $\mathbf{x}_r^{(i)}$ and $\mathbf{x}_c^{(i)}$. We re-formulate Eq. (1) as:

Algorithm 2 Compositional Learning Algorithm of RCRL

```
for each $x^{(i)} \in L \cup U$ do
    Randomly initialize the compositional representation $\mathbf{x}_c^{(i)}$;
end for
Construct the graph $G(\Phi, \Psi, W)$;
for $i = 1$ to max iteration do
    Sample a node $x^{(*)} \in \Phi$ from $G$ uniformly;
    Sample $\mathcal{S} = \{x^{(1)}, \cdots, x^{(|\mathcal{S}|)}\}$ where $x^{(1)} = x^{(*)}$;
    Minimize $-\sum_{x^{(i)} \in \mathcal{S}} \sum_{j=i-l(j \neq i)}^{i+l} \log \Pr(x^{(j)}|\mathbf{x}_c^{(i)})$;
end for
```

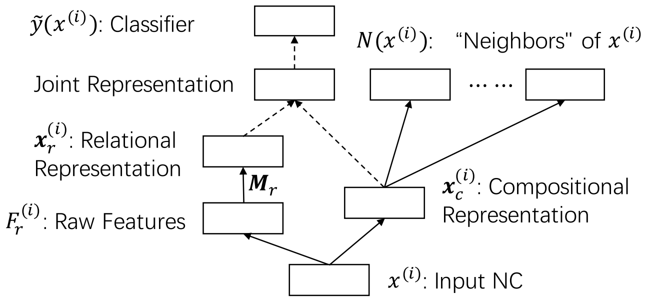

$\tilde{y}(x^{(i)})$: Classifier
$N(x^{(i)})$: "Neighbors" of $x^{(i)}$
Joint Representation
……
$\mathbf{x}_r^{(i)}$: Relational Representation
$\mathbf{M}_r$
$\mathbf{x}_c^{(i)}$: Compositional Representation
$F_r^{(i)}$: Raw Features
$x^{(i)}$: Input NC

FIGURE 2. General neural network architecture for joint optimization. Solid arrows denote direct connections. Dotted arrows refer to hidden layers. In this work, we only use one hidden layer.

$$\mathcal{J} = -\sum_{x^{(i)} \in L} \sum_{t \in T} I(t = y^{(i)}) \log \Pr(\tilde{y}(x^{(i)}) = t|\mathcal{F}_r^{(i)}, \mathbf{x}_c^{(i)})$$
$$-\lambda \sum_{\mathcal{S}} \sum_{x^{(i)} \in \mathcal{S}} \sum_{j=i-l(j \neq i)}^{i+l} \log \Pr(x^{(j)}|\mathbf{x}_c^{(i)})$$

where T is the label collection (i.e., four idiomaticity degrees).

In practice, $\Pr(x^{(j)}|\mathbf{x}_c^{(i)})$ is difficult to optimize due to the existence of the normalization factor over all possible values of $\mathbf{x}_c^{(i)}$. To speed up the training process, we employ the negative sampling technique for the Skip-gram model [40] to approximate $-\lambda \sum_{\mathcal{S}} \sum_{x^{(i)} \in \mathcal{S}} \sum_{j=i-l(j \neq i)}^{i+l} \log \Pr(x^{(j)}|\mathbf{x}_c^{(i)})$. In this technique, a binary logistic regression classifier is trained to predict whether an arbitrary NC $x^{(j)}$ is a "neighbor" of the center NC $x^{(i)}$ via noise-contrastive estimation. The positive samples are generated from $\mathcal{S}$. The negative samples are randomly paired from $L \cup U$. By minimizing the classification error, the values of compositional representations $\mathbf{x}_c^{(i)}$ can be updated. Integrating the compositional learning process into idiomaticity degree classification, the general training algorithm is shown in Algorithm 3. The algorithm converges when the joint loss $\mathcal{J}$ (with additional regularization terms omitted) does not decrease significantly.

IV. EXPERIMENTAL EVALUATION

In this section, we conduct extensive experiments to evaluate the effectiveness of the RCRL model over two Chinese

Algorithm 3 General Learning Algorithm of RCRL

```
for each $x^{(i)} \in L \cup U$ do
    Randomly initialize the compositional representation $\mathbf{x}_c^{(i)}$;
    if $x^{(i)} \in L$ then
        Compute raw features $\mathcal{F}_r^{(i)}$;
    end if
end for
Construct the graph $G(\Phi, \Psi, W)$;
while not converge do
    for $i = 1$ to max iteration do
        Sample a node $x^{(*)} \in \Phi$ from $G$ uniformly;
        Sample $\mathcal{S} = \{x^{(1)}, \cdots, x^{(|\mathcal{S}|)}\}$ where $x^{(1)} = x^{(*)}$;
        Update compositional representations $\mathbf{x}_c^{(i)}$ by training the negative sampling based classifier;
    end for
    Train the idiomaticity prediction classifier over $L$ based on $\mathbf{x}_c^{(i)}$ and $\mathcal{F}_r^{(i)}$;
end while
```

NC datasets. We also compare it with several recent approaches to make the convincing conclusion.

A. DATA SOURCE AND EXPERIMENTAL SETTINGS

The Chinese text corpus is obtained by crawling 1.3 million entity pages of *Baidu Baike*,[7] consisting of 1.1 billion words (after Chinese word segmentation). A Skip-gram model [40] is trained over the corpus and the word embeddings are set to 50 dimensions. We filter out incomplete sentences and build a sentence-level inverted index using Apache Lucene. The FudanNLP toolkit [41] is employed for NLP analysis such as Chinese word segmentation, syntactic parsing, etc.

To the best of our knowledge, SemTransCNC [42] is the only Chinese NC dataset of semantic transparency. However, it focuses on how Chinese characters form Chinese words, i.e., word formation. It does not have a clear labeling of idiomaticity degrees as well. For example, this dataset considers how the Chinese word "马虎 (carelessness)" is formed by the two Chinese characters "马 (horse)" and "虎 (tiger)". Additionally, the linguistic characteristics of Chinese word formations in this dataset are significantly different from the problem w.r.t. Chinese NCs that we consider in this work. Hence, although relevant, this dataset is not suitable for evaluating our task. Another dataset used in existing research is created by Qi *et al.* [30], which contains Chinese short expressions (not necessarily noun phrases) derived from a sememe knowledge base *HowNet*.[8] This dataset is used sememe prediction and semantic similarity computation, which is also not suitable evaluating our task.

We construct two new datasets to evaluate the RCRL model. The first dataset is *CNCBaike*, a subset of entity-category pairs taken from *Baidu Baike*. 2,500 pairs are randomly selected and sent to a group of native Chinese speakers with sufficient linguistic knowledge to label the idiomaticity degrees. We discard pairs with inconsistent labels across different annotators, and generate the *CNCBaike* dataset, consisting of 1,330 NCs and their labeled idiomaticity degrees. The second dataset is *CNCWeb*, consisting of 815 labeled NC pairs. The NCs are extracted from the same corpus, detected by POS rules and methods in [43]. The annotation process of this dataset is the same as of *CNCBaike*.

In the experiments, we set the default values of hyper-parameters as $k = 500$, $r_a = r_v = 3$, $\tau = 2$, $r_c = 20$ and $p = 16$ for raw feature generation. We randomly partition the *CNCBaike* dataset into training, development and testing sets, with the ratio as 70%:10%:20%. Because the size of *CNCWeb* is relatively small, all the pairs in *CNCWeb* are taken as the testing set with all pairs in *CNCBaike* as the training set. For LWRW based sampling, we run the algorithm in 5000 iterations with $|\mathcal{S}| = 100$. We train the compositional representation learning model with $l = 5$, $\lambda = 0.1$, $\gamma = 0.8$ and dimensions of two representations $d = 50$. We also report how changes of hyper-parameters affect the model performance over the development set in subsequent sections.

B. GENERAL PERFORMANCE COMPARISON

To our knowledge, there is no prior work that directly deals with the prediction of idiomaticity degrees of Chinese NCs. However, our task is closely related to several NLP tasks, such as lexical relation classification, idiom token classification and compositionality analysis of NCs. In this work, we consider the following models as strong baselines:

- **Lexical relation classification**: Three classical distributional models are employed to classify idiomaticity degrees using word embeddings of N_1 and N_2 as features, including the *Concat* model $\vec{v}(N_1) \oplus \vec{v}(N_2)$, the *Sum* model $\vec{v}(N_1) + \vec{v}(N_2)$ and *Diff* model $\vec{v}(N_1) - \vec{v}(N_2)$. They are frequently used as baselines for lexical relation classification [44], [45]. An SVM classifier is employed to predict idiomaticity degrees over these features.
- **Idiom token classification**: It is a recent neural network model for the task of idiom token classification based on word embeddings and lexico-syntactic patterns [9]. We train the model for four-way classification of idiomaticity degrees, instead of the two-way classification in the original paper (i.e., idiomatic vs. literal).
- **Compositionality prediction**: We consider two word embedding based models [11], [29] to compute a compositionality score for each NC. Because our task is a classification task, instead of real-value score prediction, we learn threshold-based cuts over the measures using the development set. Using three cuts, we map the learned compositional scores in (0, 1) to four idiomaticity degrees as model prediction results.

[7] *Baidu Baike* (https://baike.baidu.com) is one of the largest online encyclopedia websites in China.

[8] http://www.keenage.com/html/c_index.html

TABLE 3. Performance summarization of different approaches over two datasets: *CNCBaike* and *CNCWeb*.

Dataset	*CNCBaike*			*CNCWeb*		
Method	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
Sum Model	0.622	0.631	0.626	0.512	0.508	0.510
Concat Model	0.663	0.657	0.660	0.508	0.472	0.489
Diff Model	0.567	0.606	0.586	0.597	0.478	0.531
Model [9]	0.664	0.691	0.682	0.563	0.582	0.572
Model [29]	0.675	0.663	0.669	0.705	0.648	0.675
Model [11]	0.704	0.693	0.698	0.723	0.652	0.686
Pattern	0.770	0.766	0.768	0.745	0.687	0.715
RRL	0.785	0.776	0.780	0.762	0.703	0.731
RCRL	**0.801**	**0.783**	**0.792**	**0.784**	**0.733**	**0.758**

- **Raw pattern-based method**: It is an SVM-based classification model based on the proposed pattern-based raw features. We denote it as the *Pattern* model.
- **Variant of RCRL**: It is a variant of our approach without compositional representations, denoted as *RRL*.

Experimental results are in Table 3. As seen, distributional lexical relation classification models are not effective for our task, with F1 score generally around 40% to 60%. The most possible cause is that they simply learn the lexical meanings of two nouns within NCs, rather than how the two nouns are related to each other within NCs. A similar phenomenon (called lexical memorization) is also reported in [34] for other supervised relation prediction tasks. In contrast, our RCRL model leverages linguistics-motivated relational representations to learn how nouns with Chinese noun compounds are related to each other. The method [9] has similar performance, compared to the three distributional models.

The compositionality prediction models [11], [29] are most relevant to our task. The performance is still not satisfactory because they do not model how these compositionality scores can be mapped to the four idiomaticity degrees. Additionally, the Chinese language characteristics (i.e., specific language patterns) are not considered in these baselines. By comparing the performance of RCRL and its variants *Pattern* and *RRL*, we can see that the compositional representations contribute to idiomaticity prediction, increasing the F1 score from 2% to 5%. Compared to all the baselines, the proposed RCRL model improves the performance by a large margin, which clearly proves the effectiveness of RCRL.

C. PARAMETER ANALYSIS

During the raw feature extraction process, we employ the following steps to determine feature values: k, p, r_a, r_v, r_c and τ. The choices of k and p are mostly related to the size and the quality of the text corpus. We carry out a preliminary experiment to set the default values of k and p. As for k, we vary the value of k from {50, 100, 200, 500, 1000, 2000} and randomly select 50 textual patterns as queries to retrieve their corresponding top-k sentences. We find that when $k \geq 500$, almost no more sentences that match the search queries can be retrieved. Therefore, k is set to 500 to guarantee the high recall of sentence retrieval. A similar tuning process is conducted to determine the choice of p. When p is overly small, the expanded patterns are semantically similar to the original patterns, but the effect of query expansion is limited. In contrast, a large p may lead to the "semantic drift" phenomenon. We suggest that a suitable choice over our corpus is $p = 16$.

TABLE 4. Cases of prediction errors by RCRL. Each Chinese NC is accompanied by English literal translation and correct translation. Modifiers of Chinese NCs are underlined with linguistic heads printed in bold.

Example	Translation	Predicted	Truth
知识青年 (Knowledge **Youth**)	Sent-down youth	Type I	Type III
青铜时代 (Bronze **Age**)	The Bronze Age	Type II	Type III
邮政编码 (Postal service **Coding**)	Postal code	Type IV	Type II
宗教仪式 (Religion **Ceremony**)	Religious ceremony	Type III	Type II

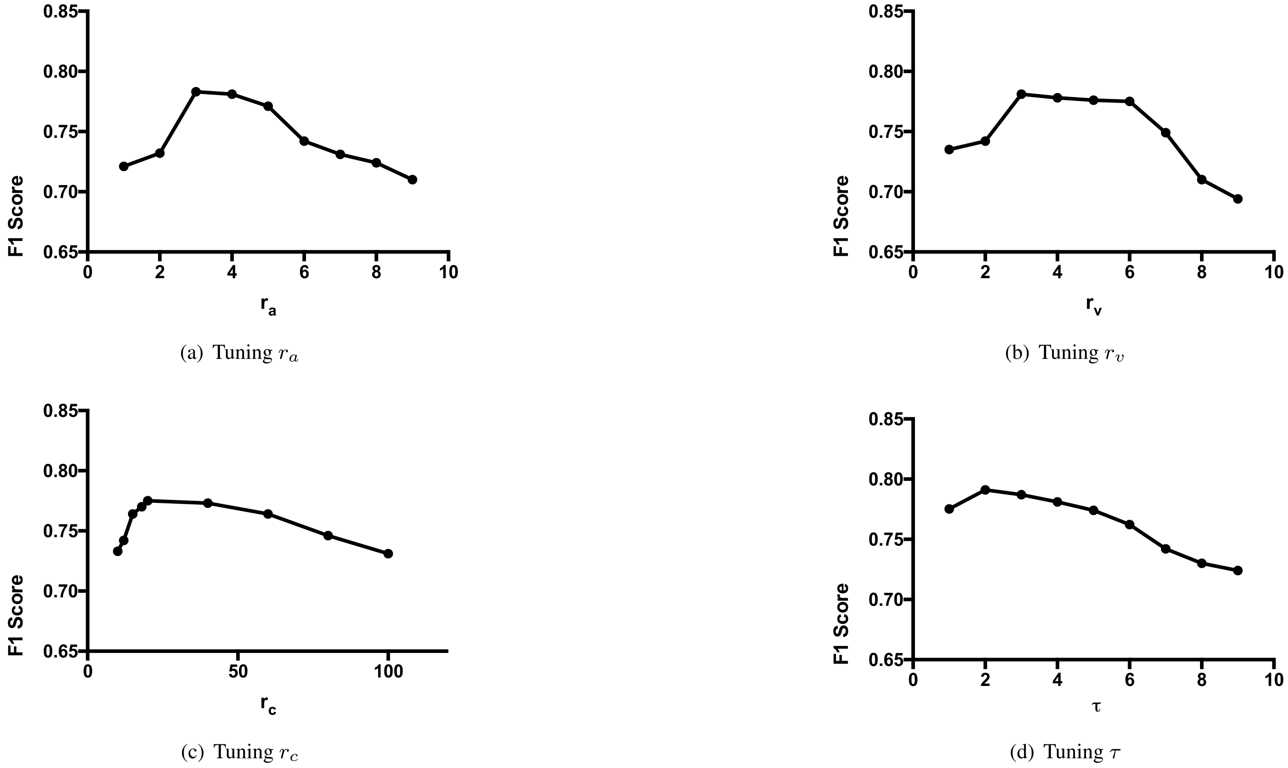
Next, we tune the values of r_a, r_v, r_c and τ. To determine which values are the most suitable in an efficient way, we employ a classifier-based trick. Each time after the raw features $\mathcal{F}_r^{(i)}$ based on a specific parameter configuration are extracted, we directly train a logistic regression classifier for idiomaticity degree prediction using $\mathcal{F}_r^{(i)}$ over the development set. The macro-averaged F1 score is utilized to measure the "goodness" of parameter settings for feature extraction. The optimal settings are as: $r_a = r_v = 3$, $\tau = 2$ and $r_c = 20$. We also report how the changes of parameters affect the performance in Figure 3. We vary one parameter once, while the rest of parameters are fixed to default values.

After raw features are extracted, we tune two parameters of RCRL (i.e., γ and d). Figure 4(a) and Figure 4(b) illustrate the prediction performance in term of F1 over the development set. We set the default values as $\gamma = 0.8$ and $d = 60$ and change one parameter each time. From the experiments, we can draw the two conclusions. i) The use of γ in the LWRW process enhances the representation learning of compositionality, mostly because this process considers both training and testing data in a transductive learning setting. iii) When the dimension of representations d is set to the number close to the dimensions of word embeddings, the model is more accurate.

Additionally, we report how the model performance changes during iterations in Algorithm 3. Each time, we run the compositional representation learning algorithm in 500 iterations, train the idiomaticity degree prediction model and report the F1 score. The result is shown in Figure 4(c). As seen, the F1 score increases steadily when the algorithm iterates. After 4,500 iterations (i.e., the algorithm runs for nine outer loops), the performance becomes stable.

D. ERROR ANALYSIS AND CASE STUDIES

We analyze prediction errors made by our model, with several cases in Table 4. In total, 300 cases are presented to human annotators to determine the underlying causes of such errors. Overall, there exist two types of errors: MDE (Metaphor Detection Error) and LPE (Lack-of-Pattern Error).



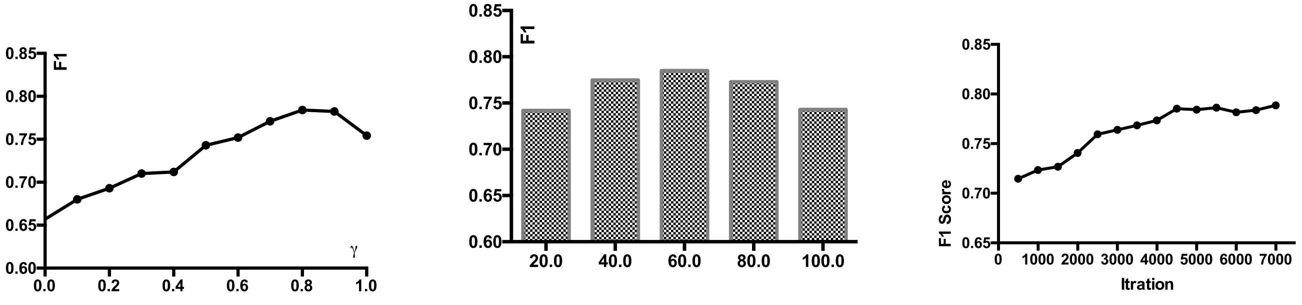
(a) Tuning r_a

(b) Tuning r_v

(c) Tuning r_c

(d) Tuning τ

FIGURE 3. Tuning of parameters r_a, r_v, r_c and τ for raw feature extraction.



(a) Tuning γ

(b) Tuning d

(c) Varying the number of iterations

FIGURE 4. Feature analysis for RCRL.

MDE accounts for 43.8% of all the prediction errors. This type of errors occurs when the model underestimates idiomaticity degrees of NCs. This is because it is challenging to detect signs of idiomaticity within NCs and use them as features. For example, although "知识青年(sent-down youth)"[9] is idiomatic, patterns such as "有知识的青年(youth with knowledge)", "没有知识的青年(youth without knowledge)" can be matched, misleading the classifier due to the existence of the auxiliary word "的(de)". The rest of errors is LPE. RCRL requires the presence of textual patterns to make certain predictions. The lack of such textual patterns causes the model to make more "idiomatic" predictions for transparent NCs. In the future, our model can be refined by incorporating additional fine-grained linguistic knowledge.

V. STUDIES AND APPLICATIONS

In this section, we conduct three data-driven studies related to the idiomaticity degree prediction of Chinese NCs. We discuss the roles of idiomaticity prediction of Chinese NCs in NLP and also show how our work benefits the understanding of natural languages and several NLP applications.

[9] "Sent-down youth" refers to young, educated people who left urban areas to live and work in rural areas during 1950s until the end of the so-called "Cultural Revolution" in China. It is literally translated as "knowledge youth".

TABLE 5. Idiomaticity degree distributions of Chinese NCs detected from the Web corpus (%).

Method	Type I	Type II	Type III	Type IV
Pattern	47.2	33.6	15.0	4.2
RRL	53.1	31.2	14.2	1.5
RCRL	51.1	34.6	12.2	2.1
Human Est.	**(49.2±1.4)**	**(38.1±1.5)**	**(10.8±0.4)**	**(1.9±1.4)**

A. OVERALL LANGUAGE IDIOMATICITY ANALYSIS OF THE CHINESE LANGUAGE

Idiomaticity of NCs affects the overall idiomaticity degrees of natural languages. In this work, we study a data-driven study on how the idiomaticity degrees of Chinese NCs are distributed over the Web corpus. We randomly sample 5,000 NCs from the previously acquired Chinese text corpus and employ three approaches with high performance (i.e., Pattern, RRL and RCRL) to predict the idiomaticity degrees of these NCs. In Table 5, we present the idiomaticity degree distributions based on the predictions of the three models. We also sample 400 NCs from the collection three times and ask human annotators to provide human re-annotation results. The standard *t-test* is used to estimated the confidence intervals, with the significance level to be $\alpha = 0.05$. We can see that the distribution generated by RCRL is closest to the human estimated results, compared to Pattern and RRL. We do not list the results of other models due to their unsatisfying performance. Another finding is that over half of the Chinese NCs (50.8%) are not transparent and have idiomatic meanings to some extent.

The statistics estimated here partially reveal the difficulty of machine understanding of Chinese. We suggest that, by treating Chinese idiomatic expressions separately in downstream NLP applications, the performance of these tasks can be further improved. For example, when computing the semantic similarity of Chinese MWEs, we should first model how the meanings of MWEs are formed by its component nouns, rather than averaging the representations of all the words together. Some of the previous studies (such as [30]) also draw the similar conclusions.

B. HOW IDIOMATICITY AFFECTS MACHINE TRANSLATION

As shown in the literature [4], the presence of idiomatic expressions may cause significant challenges for machine translation. In this work, we specifically study the relations between language idiomaticity and machine translation accuracies in terms of Chinese NCs. We consider two popular machine translation engines that are widely used in the industry (i.e., Google Translation and Bing Microsoft Translator) to translate Chinese NCs in *CNCBaike* into English. Because classical metrics such as BLEU aim at evaluating sentence translation qualities, hence they are not suitable for evaluating translation qualities of NCs. Here, we ask human annotators to label the correctness of results and report the accuracies directly, with results reported in Table 6. It is shown that the translation accuracies strongly correlate with idiomaticity degrees. The machine-generated translations of transparent NCs are generally correct, i.e, 98.2% and 97.4% accurate. NCs with higher degrees of idiomaticity have poorer translation results.

TABLE 6. Machine translation accuracy of Chinese NCs divided into four idiomaticity degrees (%).

Accuracy	Type I	Type II	Type III	Type IV
Google Translation	98.2	92.6	75.0	64.2
Microsoft Translator	97.4	90.2	78.2	58.2

In Table 7, we present three cases of translation results generated by the two translation engines, together with their literal and true translation results. We can see that errors occur when machines translate Chinese NCs (especially NCs with cultural-specific meanings) word-by-word, ignoring their idiomatic meanings and indecomposable nature (e.g., "couple lung pieces" for "Mr and Mrs Smith"). In a few cases, the translation engine even gives unexplainable, random outputs (e.g., "becoming" for "citizen-managed teacher").

This phenomenon indicates that current machine translation models still have difficulty in dealing with idiomatic expressions. The most possible cause is that machine translation systems (either statistical machine translation or neural machine translation that employ attention mechanisms such as [46]) encode word alignments across languages in order to generate translated results. The learning of word alignments ignores the processing of idiomatic expressions, which heavily involve semantic composition of words. We can see that our task has the potential to be combined with other NLP tasks to increase the models' ability for semantic understanding.

C. EXPERIMENTS OVER THE ENGLISH LANGUAGE FOR COMPOSITIONALITY PREDICTION

Although our work mostly addresses the idiomaticity issue of Chinese NCs, we investigate whether RCRL can be applied to the English language. A closely related task in English is to predict the compositionality of NCs. In this paper, we implement a variant of RCRL to address to the NC compositionality prediction task for English, and evaluate it over a widely used dataset Reddy *et al.* [13]. In the implementation, because we exploit the auxiliary pattern in Chinese for raw feature extraction, we manually translate such pattern into English. The patterns that we use include: "[…] of […]", "[…]'s […]", "[…] that is […]", "[…] which is […]", etc. The compositional representations are computed using the same approach for Chinese, as it is language-independent. For compositional score prediction, we replace the cross entropy loss of idiomaticity degree classification with the MSE regression loss. The regression loss is defined as follows:

$$\mathcal{J} = \sum_{x^{(i)}\in L} (\tilde{y}(x^{(i)}; \mathcal{F}_r^{(i)}, \mathbf{x}_c^{(i)}) - y(x^{(i)}))^2 - \lambda \sum_{\mathcal{S}} \sum_{x^{(i)}\in\mathcal{S}} \sum_{j=i-l(j\neq i)}^{i+l} \log \Pr(x^{(j)}|\mathbf{x}_c^{(i)})$$

TABLE 7. Machine translation results from Google Translation and Microsoft Translator.

Chinese NC	Result of Google Translation	Result of Microsoft Translator
夫妻肺片	Couple lungs	Couple lung slices
Correct translation: *Mr and Mrs Smith (Sliced beef and ox organs in chili sauce)*		
竹书纪年	Bamboo book year	The Annals of Bamboo Books
Correct translation: *Bamboo Annals (A chronicle of ancient China)*		
民办教师	Private teacher	Becoming
Correct translation: *Citizen-managed teacher (teacher who is employed by private schools in China)*		

TABLE 8. The performance of NC compositionality prediction in terms of Spearman Correlation Coefficient ρ.

Method	Spearman Correlation Coefficient ρ
Model [13]	0.71
Model [29]	0.80
Model [11]	0.82
RCRL	0.81

where $\tilde{y}(x^{(i)}; \mathcal{F}_r^{(i)}, \mathbf{x}_c^{(i)})$ is the predicted compositionality score of the English NC $x^{(i)}$, given the features $\mathcal{F}_r^{(i)}$ and $\mathbf{x}_c^{(i)}$. $y(x^{(i)})$ is the ground-truth compositionality score.

During the testing stage, the prediction scores are evaluated against the ground truth ratings using the Spearman Correlation Coefficient ρ. We follow the exact experimental settings and use the same English text corpus as in [11], with the experimental results presented in Table 8. As seen, our approach has the performance $\rho = 0.81$, which outperforms previous methods for NC compositionality prediction (e.g., [13]) and is comparable to two recent state-of-the-art approaches [11], [29]. Therefore, RCRL is not entirely language-specific and can be applied to English as well.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a Relational and Compositional Representation Learning model (RCRL) to predict the idiomaticity degrees of Chinese NCs. The model predicts idiomaticity of Chinese NCs based on relational textual patterns and compositionality analysis via an integrated neural network. We conduct extensive experiments over two datasets to evaluate RCRL. The experimental results show that RCRL outperforms all the baselines. Additionally, the usefulness of RCRL and the roles of idiomaticity prediction of NCs in NLP are illustrated by three studies. Future works include: i) applying our work to the interpretation and machine understanding of Chinese NCs; and ii) extending our method to other languages and lexical units.

REFERENCES

[1] C. Sporleder, L. Li, P. Gorinski, and X. Koch, "Idioms in context: The IDIX corpus," in *Proc. Int. Conf. Lang. Resour. Eval.*, May 2010, pp. 1–8.

[2] V. Shwartz and I. Dagan, "Still a pain in the neck: Evaluating text representations on lexical composition," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 403–419, Feb. 2019.

[3] M. Constant, G. Eryigit, J. Monti, L. Van Der Plas, C. Ramisch, M. Rosner, and A. Todirascu, "Multiword expression processing: A survey," *Comput. Linguistics*, vol. 43, no. 4, pp. 837–892, Dec. 2017.

[4] Y. Shao, R. Sennrich, B. L. Webber, and F. Fancellu, "Evaluating machine translation performance on Chinese idioms with a blacklist method," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, May 2018, pp. 31–38.

[5] W. Cui, X. Zhou, H. Lin, Y. Xiao, H. Wang, S.-W. Hwang, and W. Wang, "Verb pattern: A probabilistic semantic representation on verbs," in *Proc. 30th AAAI Conf. Artif. Intell.*, May 2016, pp. 2587–2593.

[6] A. Fazly, P. Cook, and S. Stevenson, "Unsupervised type and token identification of idiomatic expressions," *Comput. Linguistics*, vol. 35, no. 1, pp. 61–103, Mar. 2009.

[7] J. Peng, A. Feldman, and E. Vylomova, "Classifying idiomatic and literal expressions using topic models and intensity of emotions," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Feb. 2014, pp. 2019–2027.

[8] G. Salton, R. J. Ross, and J. D. Kelleher, "Idiom token classification using sentential distributed semantics," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1–12.

[9] M. King and P. Cook, "Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of english verb-noun combinations," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2018, pp. 345–350.

[10] S. S. Walde, A. Hätty, and S. Bott, "The role of modifier and head properties in predicting the compositionality of english and german noun-noun compounds: A vector-space perspective," in *Proc. 5th Joint Conf. Lexical Comput. Semantics*, Aug. 2016, pp. 148–158.

[11] S. Cordeiro, C. Ramisch, M. Idiart, and A. Villavicencio, "Predicting the compositionality of nominal compounds: Giving word embeddings a hard time," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Aug. 2016, pp. 1986–1997.

[12] M. Yazdani, M. Farahmand, and J. Henderson, "Learning semantic composition to detect non-compositionality of multiword expressions," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2015, pp. 1733–1742.

[13] S. Reddy, D. McCarthy, and S. Manandhar, "An empirical study on compositionality in compound nouns," in *Proc. 5th Int. Joint Conf. Natural Lang. Process.*, 2011, pp. 210–218.

[14] X. Lu and B. P. Wang, "Towards a metaphor-annotated corpus of mandarin chinese," *Lang. Resour. Eval.*, vol. 51, no. 3, pp. 663–694, Sep. 2017.

[15] Y. Sawai, H. Shindo, and Y. Matsumoto, "Semantic structure analysis of noun phrases using abstract meaning representation," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. Asian Fed. Natural Lang. Process.*, Jul. 2015, pp. 851–856.

[16] C. N. Li and S. A. Thompson, *Mandarin Chinese: A Functional Reference Grammar*, vol. 42. Hamlin, TX, USA: JAS, 1989, no. 3.

[17] C. Ramisch, S. Cordeiro, L. Zilio, M. Idiart, and A. Villavicencio, "How naked is the naked truth? A multilingual lexicon of nominal compound compositionality," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Aug. 2016, pp. 156–161.

[18] L. Wang and M. Wang, "A study on the taxonomy of chinese noun compounds," in *Proc. 16th Workshop Chin. Lexical Semantics*, 2015, pp. 262–269.

[19] M. Pasca, "Interpreting compound noun phrases using Web search queries," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2015, pp. 335–344.

[20] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 701–710.

[21] C. Dima and E. W. Hinrichs, "Automatic noun compound interpretation using deep neural networks and word embeddings," in *Proc. 11th Int. Conf. Comput. Semantics*, Apr. 2015, pp. 173–183.

[22] V. Shwartz and C. Waterson, "Olive oil is made of olives, baby oil is made for babies: Interpreting noun compounds using paraphrases in a neural model," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 218–224.

[23] T. V. de Cruys, S. D. Afantenos, and P. Müller, "MELODI: A supervised distributional approach for free paraphrasing of noun compounds," in *Proc. 7th Int. Workshop Semantic Eval., (SemEval NAACL-HLT)*, Jun. 2013, pp. 144–147.

[24] I. Hendrickx, Z. Kozareva, P. Nakov, Ó. Séaghdha, S. Szpakowicz, and T. Veale, "Semeval-2013 task 4: Free paraphrases of noun compounds," in *Proc. 7th Int. Workshop Semantic Eval., (SemEval NAACL-HLT)*, Jun. 2013, pp. 138–143.

[25] C. Hashimoto and D. Kawahara, "Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct. 2008, pp. 992–1001.

[26] W. Gharbieh, V. Bhavsar, and P. Cook, "A word embedding approach to identifying verb-noun idiomatic combinations," in *Proc. 12th Workshop Multiword Expressions*, Aug. 2016, pp. 112–118.

[27] C. Liu and R. Hwa, "Heuristically informed unsupervised idiom usage recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Jan. 2018, pp. 1723–1731.

[28] D. Kiela and S. Clark, "Detecting compositionality of multi-word expressions using nearest neighbours in vector space models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct. 2013, pp. 1427–1432.

[29] B. Salehi, P. Cook, and T. Baldwin, "A word embedding approach to predicting the compositionality of multiword expressions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, May/Jun. 2015, pp. 977–983.

[30] F. Qi, J. Huang, C. Yang, Z. Liu, X. Chen, Q. Liu, and M. Sun, "Modeling semantic compositionality with sememe knowledge," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, Jul. 2019, pp. 5706–5715.

[31] P. Nakov, "On the interpretation of noun compounds: Syntax, semantics, and entailment," *Natural Lang. Eng.*, vol. 19, no. 3, pp. 291–330, Jul. 2013.

[32] I. Dagan and V. Shwartz, "Paraphrase to explicate: Revealing implicit noun-compound relations," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1200–1211.

[33] A. Gupta, R. Lebret, H. Harkous, and K. Aberer, "Taxonomy induction using hypernym subsequences," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 1329–1338.

[34] O. Levy, S. Remus, C. Biemann, and I. Dagan, "Do supervised distributional methods really learn lexical inference relations?" in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2015, pp. 970–976.

[35] J. Betteridge, A. Ritter, and T. M. Mitchell, "Assuming facts are expressed more than once," in *Proc. 27th Int. Florida Artif. Intell. Res. Soc. Conf.*, May 2014, pp. 431–436.

[36] L. Qiu and Y. Zhang, "ZORE: A syntax-based system for chinese open relation extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct. 2014, pp. 1870–1880.

[37] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam, "Open information extraction: The second generation," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, Jun. 2011, pp. 3–10. [Online]. Available: http://www.cse.iitd.ac.in/~mausam/index.html

[38] S. Kuzi, A. Shtok, and O. Kurland, "Query expansion using word embeddings," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2016, pp. 1929–1932. s

[39] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 855–864.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent. (ICLR)*, May 2013. [Online]. Available: https://arxiv.org/pdf/1301.3781.pdf

[41] X. Qiu, Q. Zhang, and X. Huang, "Fudannlp: A toolkit for chinese natural language processing," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics (Syst. Demonstrations)*, Aug. 2013, pp. 49–54.

[42] S. Wang, C. Huang, Y. Yao, and A. Chan, "Building a semantic transparency dataset of chinese nominal compounds: A practice of crowdsourcing methodology," in *Proc. Workshop Lexical Grammatical Resour. Lang. Process. (LG-LP COLING)*, Aug. 2014, pp. 147–156.

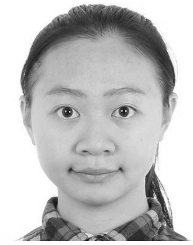
[43] V. I. N. T. Vincze and G. Berend, "Detecting noun compounds and light verb constructions: A contrastive study," in *Proc. Workshop Multiword Expressions: From Parsing Gener. Real World, (MWE ACL)*, Jun. 2011, pp. 116–121.

[44] S. Roller, K. Erk, and G. Boleda, "Inclusive yet selective: Supervised distributional hypernymy detection," in *Proc. 25th Int. Conf. Comput. Linguistics*, Aug. 2014, pp. 1025–1036.

[45] P. D. Turney and S. M. Mohammad, "Experiments with three approaches to recognizing lexical entailment," *Natural Lang. Eng.*, vol. 21, no. 3, pp. 437–476, May 2015.

[46] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015. [Online]. Available: https://arxiv.org/pdf/1409.0473.pdf

CHENGYU WANG received the B.E. degree in software engineering from East China Normal University (ECNU), in 2015, where he is currently pursuing the Ph.D. degree. He is also involved in the construction and application of large-scale knowledge graphs. His research interests include web data mining, information extraction, and natural language processing.



YAN FAN received the master's degree in software engineering from East China Normal University (ECNU), in 2019. She is currently an Algorithm Engineer with Alibaba Group. Her research interests include question answering, dialogue systems, and the construction and application of knowledge graphs.

XIAOFENG HE received the Ph.D. degree from Pennsylvania State University. He worked with Microsoft, Yahoo Laboratories, and the Lawrence Berkeley National Laboratory. He is currently a Professor of computer science with the School of Computer Science and Technology, East China Normal University, China. His research interests include machine learning, data mining, and information retrieval.

HONGYUAN ZHA received the Ph.D. degree in scientific computing from Stanford University, in 1993. He has been working on information retrieval, machine learning applications, and numerical methods. He is currently a Professor with East China Normal University and with the School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology. He was a recipient of the Second Prize of Leslie Fox Prize, in 1991, of the Institute of Mathematics and its Applications, the Outstanding Paper Awards of the 26th International Conference on Advances in Neural Information Processing Systems (NIPS 2013), and the Best Student Paper Award (advisor) of the 34th ACM SIGIR International Conference on Information Retrieval (SIGIR 2011). He served as an Associate Editor for the IEEE Transactions on Knowledge and Data Engineering.

AOYING ZHOU is a currently a Professor with East China Normal University. He is also acting as the Vice-Director of ACM SIGMOD China and the Database Technology Committee of the China Computer Federation. His research interests include data management for data-intensive computing and memory cluster computing. He is also serving as a member of the editorial boards of *The VLDB Journal*, the *WWW Journal*, and so on.