Implanting Rational Knowledge into Distributed Representation at Morpheme Level

Zi Lin,[1,3] **Yang Liu**[2,3]
[1]Department of Chinese Language and Literature, Peking University
[2]Institute of Computational Linguistics, Peking University
[3]Key Laboratory of Computational Linguistics (Ministry of Education), Peking University
{zi.lin, liuyang}@pku.edu.cn

Abstract

Previously, researchers paid no attention to the creation of unambiguous morpheme embeddings independent from the corpus, while such information plays an important role in expressing the exact meanings of words for parataxis languages like Chinese. In this paper, after constructing the Chinese lexical and semantic ontology based on word-formation, we propose a novel approach to implanting the structured rational knowledge into distributed representation at morpheme level, naturally avoiding heavy disambiguation in the corpus. We design a template to create the instances as pseudo-sentences merely from the pieces of knowledge of morphemes built in the lexicon. To exploit hierarchical information and tackle the data sparseness problem, the instance proliferation technique is applied based on similarity to expand the collection of pseudo-sentences. The distributed representation for morphemes can then be trained on these pseudo-sentences using word2vec. For evaluation, we validate the paradigmatic and syntagmatic relations of morpheme embeddings, and apply the obtained embeddings to word similarity measurement, achieving significant improvements over the classical models by more than 5 Spearman scores or 8 percentage points, which shows very promising prospects for adoption of the new source of knowledge.

Introduction

Nowadays, learning representations for the meanings of words has been a key problem in natural language processing (NLP). The basic unit in NLP is usually and mainly word. However, for parataxis languages like Chinese, which is made up of hieroglyphic characters, *word* is not a natural unit, and character can provide yet rich semantic information (Fu 1981; Xu 2004).

Theoretically, the meanings of Chinese words can be deduced from the meanings of characters. However, the same Chinese characters within words may hold different meanings, so the meanings of an identical character should be further differentiated. For example, in the words "huā qián" (to spend money) and "táo huā" (peach-blossom), the character meanings of "huā" are not the same. Chinese linguists use the term *yusu* (morpheme) to distinguish identical characters with different meanings, which is defined as *the smallest combination of meaning and phonetic sound in Chinese* (Zhu 1982). Previously, linguists define *morpheme* as *the smallest meaning-bearing unit of language as well as the smallest unit of syntax* (Matthews 1972). The Chinese morpheme is close to this definition in terms of semantics but requires a smallest phonetic sound, i.e., it should correspond to a character in form. Therefore, the character "huā" in "huā qián" and "táo huā" refers to different morphemes "huā_1" and "huā_2". We sometimes use the term sememe to refer to the meanings of morphemes, so the sememe of "huā_1" is *to spend*, while the sememe of "huā_2" is *flowers*.

The meanings of words and the meanings of morphemes are highly related to some extent, and the sub-units of words will form patterns during word-building. For example, in Chinese words "táo huā" (peach-blossom) and "hé huā" (lotus-blossom), one will find "huā", which means *flowers*, as a common component. The components before "huā", as different modifiers, also hold their meanings. These words are of Modifier-Head structure, and the morphology is somewhat similar to that of syntax in Chinese.

Therefore, morphemes and their combination patterns are very important for the generation of the meanings of words. Researchers have addressed the importance of morphological compositionality and word-formation analysis. (Lazaridou et al. 2013) explored the application of compositional distributed semantic models, originally designed to learn phrase meanings. (Luong, Socher, and Manning 2013) combined recursive neural networks with neural language models to consider contextual information in learning morphologically-aware word representations.

As for Chinese, (Chen et al. 2015) proposed a character-enhanced word embeddings model (CWE) by regarding the word as the simple combination of characters and obtained character embeddings on the corpus, without further knowledge of morphemes or sememes input. (Niu et al. 2017) employed HowNet to learn word representation on corpus for the SE-WRL model. By using attention schema for rough word sense disambiguation, HowNet sememe embeddings were thus obtained directly from the large-scale corpus.

To avoid sense disambiguation on the corpus, there was also work trying to obtain word sense embeddings by leveraging the dictionary definitions or lexical resources. `Dict2vec` used the English version of Cambridge, Oxford, Collins and dictionary.com to build new word pairs so that semantically-related words are moved closer, and negative

sampling filters out pairs whose words are unrelated in dictionaries (Tissier, Gravier, and Habrard 2017). As there is no structured knowledge can be exploited in the dictionaries, such selection of word pairs tended to be cumbersome and complicated. `WordNet2vec` (Bartusiak et al. 2017) created vectors for each word from WordNet by first simplifying the WordNet structure into a graph and then utilizing the shortest paths to encode the words, which cannot distinguish and retain the specific semantic relations.

To the best of our knowledge, all these works aimed to get better representations for words, while little emphasis has been laid on creation and analysis of relatively independent morpheme embeddings. Also, all the works relied highly on corpora or simple dictionary knowledge, and have not been able to form distributed representation directly from the structured rational knowledge.

In this paper, after constructing the Chinese lexical and semantic ontology based on word-formation, we propose a novel approach to implanting the structured rational knowledge into distributed representation at morpheme level without using any text corpus. We first introduce the construction of rational knowledge of Chinese morphemes at Peking University. Then we extract this knowledge to design a template to create instances and proliferate instances based on similarity, as a source to train data for morpheme embeddings. For evaluation, we validate the paradigmatic and syntagmatic relations for morpheme embeddings, and apply the obtained embeddings to word similarity measurement, achieving significant improvements over the classical models [1].

Constructing Rational Knowledge of Morphemes

In recent years, Chinese lexical and semantic ontologies such as Hantology (Chou 2005) and HowNet (Dong, Dong, and Hao 2007) have exhibited valuable work related to morphemes.

One is Hantology. It's known that a Chinese character sometimes can be roughly deduced to a radical as a part of it, which may help to predict the category of the character. Some radicals themselves even stand alone, known as radical characters. Hantology exploits some 540 radical characters of *ShuoWen*[2] as basic semantic symbols of Chinese. However, it is also limited to the use of radical characters, regardless of thousands of common Chinese characters involving about 20,000 morphemes. Hantology is obviously lacking in fine-grained representation of morphemes.

Another is HowNet. The contributors hypothesize that all the lexicalized concepts can be reduced to the relevant sememes. They, by personal introspection and examination, allegedly arrive at a set of around 2,800 language-independent sememes. Having no morphological analysis as the basis, the definitions of HowNet sememes are just abstract and prototypic, failing to be linked with any real and existing Chinese radical characters or characters or morphemes. Such assumptions may lead to doubt about its objectivity and coverage of Chinese semantics.

According to the above analysis, for construction of language resources, we want to ensure method objectivity as well as data coverage, and fully consider the characteristics of Chinese, i.e., the close relationship between a word and its morphemes.

For years of development, the *Chinese Object-Oriented Lexicon* (*COOL*) of Peking University has been in process (Liu, Lin, and Kang 2018). We adopt the *Synonymous Morpheme Set* (SMS) to denote the *Morphemic Concept* (MC) and build the hierarchy of MCs. On this basis, *COOL* further describes word-formation pattern and forms strict bindings between morphemes as sub-units of words and the MCs. Such rational knowledge may be applied to the fields of humanities as well as to industries. We plan to release our lexicon in the near future for research purpose.

Constructing MCs and the Hierarchy

Morpheme Extraction and Encoding To make sure of full coverage and fine granularity, the collection of our morphemes and words is from the 5th edition of *Xiandai Hanyu Cidian* (Contemporary Chinese Dictionary, CCD) by the Commercial Press, the most influential dictionary in China. CCD contains Chinese morphemes as well as their sense definitions, which have been carefully made by lexicographers for tens of years.

As different morphemes may originate from the identical character, we then set a unique encoding for each morpheme in CCD in the format `H_X1_X2_X3`, where `H` represents the character as host, and `X1` means that the current morpheme is the $\texttt{X1}^{th}$ entry of this host in the dictionary, `X2` means that there are `X2` sememes in all for this entry, and `X3` means that the current is the $\texttt{X3}^{th}$ sememe. For example, the character "shù" carries 4 sense definitions in the dictionary, corresponding to 4 different morphemes and sememes, and acquires encodings as shown in Table 1.

Encoding	Sense Definition (Sememe)
树1_04_01	木本植物的通称 (general term of woody plant)
树1_04_02	移植，栽培 (to plant)
树1_04_03	树立，建立 (to set up)
树1_04_04	姓氏 (surname)

Table 1: Examples of morphemes

We excavated data from CCD and collected a total of 8,514 Chinese characters and their 20,855 morphemes. On account of the fact that morphemes can have parts of speech (POSs), even when a morpheme is not necessarily a word, the POSs of morphemes could somehow be drawn from the POSs of words (Yin 1984). We further classified all these morphemes into 13 POS types, specifically, nominal, verbal, adjectival, adverbial, numeral, classifier, pronominal, prepositional, auxiliary, conjunctional, onomatopoetic, interjec-

[1] The data of morpheme embeddings and word similarity measurement is available at https://github.com/zi-lin/MC for research purpose.

[2] *Shuowen* is historically the first dictionary to analyze the structures and meanings of Chinese characters and give the rationale behind them.

tion and affix morphemes. The free morphemes have postags in the dictionary, and for the bound morphemes, we manually annotated the postags of them as complement. We found that nominal, verbal and adjectival morphemes hold a total of 88.74% of Chinese morphemes as the main body, while the rest amounts to 11.26%.

Forming MCs Some morphemes in Chinese actually correspond to the same or similar sememes. For example, morphemes "shù1_04_01" and "mù1_07_01" all refer to the meaning of *general term of woody plant*. Clustering such morphemes together will help to get basic semantic units with high information density. Therefore, we are inspired to merge and represent morphemes in form of SMSs, and try to exploit such pieces of knowledge for further use.

To get reliable SMSs and considering that lexicographers used same or similar brief wordings for some sense definitions of morphemes, we measure the sememe similarity by using the co-occurrence model. The automatic clustering is just a reference for the substantial manual annotation, which ensures both the efficiency and the quality of the construction. For a particular sense definition, according to its semantic similarity score with others in descending order, and by hand-correcting, we form a corresponding SMS. Repeat this process until all the meanings of morphemes with the same POS are covered, and then turn to other POSs until all sense definitions are covered.

These SMSs just refer to the MCs of Chinese. By now, we have achieved 4,198 MCs for morphemes of the main body, including 2,018 nominal MCs, 1,630 verbal MCs and 550 adjectival MCs respectively. These MCs then form a collection of all the smallest semantic units of Chinese, showing a clear advantage of data coverage and method objectivity.

For example, here we list samples of verbal MCs in Table 2, with the different cardinality of SMS (the number of morphemes in the SMS, **#Mor** for short), i.e., the size of the MC. **#MC** then refers to the total number of MCs with regards to a particular #Mor. Here we just list one MC example for each #Mor, along with its MC definition. For layout problem, we just remove the morpheme encodings and show characters as morphemes in the figure (so multiple occurrences of an identical character are allowed).

Building the Hierarchy of MCs Up to this point, the MCs are still discrete concepts. However, some MCs are highly related to others semantically. For example, as paradigmatic relation, the nominal MCs which mean *herbage* and *xylophyta* all denote the meaning of *plant*; as syntagmatic relation, the verbal MCs which mean *to sprout* and *to grow*, and the adjectival MCs which mean *luxuriant* and *sere*, are relevant to the MCs in connection with *plant*. Therefore, to express those relations, a hierarchical structure for all the MCs is needed in organizing the knowledge base, with the purpose of facilitating reasoning and computing afterwards.

Inspired by WordNet (Miller and Fellbaum 1998), the nominal MCs are structuralized based on hypernymy relation. As for MCs of other POSs, we are enlightened by the Generative Lexicon Theory (Pustejovsky 1995) to build a hierarchy where the nominal MCs are of the core structure. The verbal MCs refer to actions of the nominal, while adjectival MCs refer to attributes of the nominal. In this way, the hierarchy of the main body of Chinese morphemes, comprising the nominal, verbal and adjectival, is internally isomorphic in general.

By making use of this solution, we have obtained paradigmatic relations within the same POS and syntagmatic relations across different POSs for Chinese morphemes.

#Mor	#MC	%	MC Example	MC Definition
77	1	0.06	溻炆焗卧焗俏熯炸馓熯汆汆媪煲熯扒熬热蒸烤…	to cook
34	2	0.12	谴斥咎责呵非谪病诮吐谇赖怨说怪骂貶驳批埋…	to blame
32	2	0.12	储存存厝寘寇庋囤囊痦蓄贮蹲辐蓄窝隐藏养居…	to store
29	1	0.06	称诵谓陈语云道曰话叙言具咧拉齿谙讲聊扯吭…	to communicate
28	1	0.06	刨剐剖劈异溃掐溃披敞析撑彀睁解捭展抡巴启…	to open
26	2	0.12	亡化卒危夭尽徂故殁殂殇殪殪殪没终绝薨逝弱…	to die
25	4	0.25	了会博喻审悉悟惺憭懂明晓曝照省知解谂通醒…	to know, to understand
24	1	0.06	诈唊缃赚拐哄冤诳欺诱逗饵骗钓罔谩蒙虞诈诖…	to lie, to deceive
23	2	0.12	仗挠揪攀扒攉挝搏捉攫掺揭搦攫执捞手持秉端…	to hold
22	3	0.18	上傅刮刷图垩墅墩打抆抹抹抹抿拭搵擦敷油涂…	to paint, to scrub
21	4	0.25	向近凑靠压附守拢临挨傍贴毗薄濒就迫促即接…	to approach
20	2	0.12	与予命给付授效畀丐赋发缴奉致施付献投供把	to give, to send
19	10	0.61	会恬参摊撞晤寔碰见靓靓谒赶迕逢遘遭邂	to meet sb.
18	6	0.37	上下之到即如幸往往徂徂抵至莅赴赶趟适造	to get somewhere
17	5	0.31	串化为变变变成拜愈可嬗失革移构迁济	to change
16	4	0.25	少少欠下拉亏该缺离短差脑匮乏憋阙	to lack
15	8	0.49	上到及底彀勾够偏满齐平臻达致等	to achieve
14	11	0.67	号哮嗥哮呶吆呼喝唱喊嚎嚷吼呐	to yell out
13	17	1.04	烊溶熔化化焊融铄泮镅销冶炀	to melt, to dissolve
12	14	0.86	遴择择擢抡拔拣刷挑选铨调	to select, to pick up
11	19	1.17	倒砸败败败输负胜破衄北	to fail
10	28	1.72	作做写修书渤撰著编纂	to literarily create
9	20	1.23	夺剽劫掠攫抢越掳逼	to rob, to loot
8	39	2.39	僦租租赁赁贳贷假	to rent
7	45	2.76	奖称嘉褒赉夸与	to think highly of
6	59	3.62	施舍斋赏食捯	to donate
5	86	5.28	赌扳竞角斗	to compete
4	121	7.42	掠鞭箠抶	to whip
3	163	10.00	冤屈骫	to feel a sense of injustice
2	305	18.71	丈测	to measure
1	645	39.57	倍	to double
N/A	1630	100.00	N/A	N/A

Table 2: Samples of verbal MCs and their definitions

Word-Formation Tagging

As Chinese linguists argued, morphemes as sub-units of words have particular word-formation patterns in word-building (Chao 1968; Liu 1990). It is necessary to explore these patterns for understanding larger language units like words.

Word-formation Pattern	Example	Percentage
定中(Modifier-Head)	红旗(red-flag)	37.94%
联合(Parallel)	买卖(buy-sell)	21.90%
述宾(Verb-Object)	植树(plant-tree)	15.62%
状中(Adverb-Verb)	广播(widely-broadcast)	8.09%
后附加(Suffixation)	里头(inside-∅)	4.43%
单纯词(Noncompound)	克隆(clone)	3.99%
连谓(Verb-Verb)	剪贴(clip-paste)	3.28%
前附加(Prefixation)	老虎(∅-tiger)	1.34%
述补(Verb-Complement)	击毙(shoot-died)	1.21%
主谓(Subject-Predicate)	地震(earth-quake)	1.01%
重叠(Overlapping)	星星(star-star)	0.59%
介宾(Preposition-Object)	从小(from-youth)	0.30%
名量(Noun-Classifier)	纸张(paper-piece)	0.15%
数量(Quantifier)	一天(one-day)	0.11%
复量(Classifier-Classifier)	人次(person-(per)time)	0.04%

Table 3: Word-formation patterns and examples

For the selection of word-formation patterns, Chinese linguists generally hold two different views - one based on syntactics (Chao 1968) and one based on semantics (Liu 1990).

Word	POS	Word-Formation	1st-MC POS	2nd-MC POS	1st-MC	2nd-MC
植树(plant-tree)	动词(Verb)	述宾(Verb-Object)	动语素(Verbal)	名语素(Nominal)	养1_11_02	木1_07_01

Table 4: Demo of the piece of rational knowledge at morpheme level. For brevity, we use the first morpheme encoding in each MC to denote the MC itself.

<B>	<B>-<1st-MC POS>	<POS>	<1st-MC>	<2nd-MC>	<Word-Formation>	<E>
B	B-Verbal	Verb	{养1_11_02}	{木1_07_01}	Verb-Object	E

Table 5: The designed template (the first line) and the pseudo-sentence for the word "zhí shù" (*plant-tree*) (the second line). <B> and <E> refers to the begin-position and the end-position respectively.

Semantic labels have advantages of naturalness and intuition but are hard to unify and too complicated to be processed by computers. In contrast, syntactic labels are relatively simple and uniform and are somewhat consistent with syntactic structures (Fu 2003). Therefore, we eventually choose the labels of word-formation geared towards syntactics to facilitate the construction of the resources. It is noteworthy that after the strict binding between morphemes and the achieved MCs, we actually acquire semantic word-formation knowledge to some extent.

After data analysis, we adopted a collection of 15 labels for Chinese word-formation tagging. The example and percentage of each word-formation pattern are listed in Table 3. By now, 52,108 Chinese disyllabic words in CCD have all been labelled with their word-formation patterns.

Binding between Sub-units of Words and MCs

After the work of and , this procedure aims to provide character-to-morpheme bindings for Chinese words, i.e., we want to assign specific MCs to morphemes as sub-units of words.

For all these 52,108 Chinese disyllabic words, we list all the possible morphemes for the first and second character, among which we choose the appropriate ones. For example, for word "zhí shù" (*plant-tree*), the first sememe is *to plant*, while the second sememe is *general term of the woody plant*. Since each sememe corresponds to a unique morpheme encoding, the word is now formalized as <zhí1_04_01, shù1_04_01>. It is noted that each morpheme encoding belongs to a unique MC, and this actually fulfills bindings between sub-units of words and MCs. The morphemes within words would now be constrained by the hierarchy of MCs.

Taking advantage of such lexical and semantic knowledge representation, *COOL* may meet a variety of needs. In humanities, it shows potential in Chinese lexicography (e.g. concept-based browser), Chinese teaching (e.g., level evaluation standard), language study (e.g., Chinese ontology), etc. Such interdisciplinary applications can benefit from these pieces of rational knowledge (Liu, Lin, and Kang 2018). As for NLP, to cope with the difficulties of full semantic prediction of unknown words, it can give specific lexical and semantic generation according to tasks and requirements, showing a high level of flexibility and tailorability. We leveraged the rich information of *COOL* to predict word-formation patterns, morphemes and their postags within words. Our result of prediction is simple and easy for applications (Tian and Liu 2016).

Training Distributed Representation for Morphemes

Vector-space word representation has been successful in recent years across a variety of NLP tasks (Luong, Socher, and Manning 2013). Apart from rational methods in the abovementioned application, we are motivated to implant such valuable knowledge into distributed representation. However, how to generate the so-called *corpus* based on such knowledge is a central issue and makes a big challenge. And till now there has not been such practice and approach reported.

To address this issue, we design a template based on the structured rational knowledge to generate the instances, and conduct instance proliferation to exploit hierarchical information and tackle data sparseness problem. Such proliferated instances of the word by semantic word-formation, as pseudo-sentences, have thus formed a *corpus* relevant to rational knowledge built in the lexicon. Then word2vec is applied to such *corpus* to obtain distributed representation for morphemes.

Template Design

Instead of using context words to predict the target word, we try to make full use of the piece of rational knowledge to generate the instantiated pseudo-sentence of morphemes. To achieve this, we propose to design a template to create the instances merely from the pieces of rational knowledge built in the lexicon. Under this assumption, word by semantic word-formation actually represents a certain and real occurrence of the combination of morphemes as in their respective MCs. Each pseudo-sentence in the so-called *corpus* now refers to the instance of a word by semantic word-formation.

The piece of knowledge for use at morpheme level is shown in Table 4. As for such piece of information, we hence design the template as shown in the first line in Table 5. Accordingly, the pseudo-sentence for the word "zhí shù" (*plant-tree*) is generated as shown in the second line.

By now, we get an instantiated pseudo-sentence of morphemes through the application of the template. Hence a total of 52,108 instances as pseudo-sentences are generated by

	C_a	C_b	$\mathbb{C}_a$	$\mathbb{C}_a$
Seed Word	养1_11_02	木1_07_01	养1_11_02 (to plant) 浇1_04_03 (to water) ... ... 耕1_02_01 (to cultivate)	木1_07_01 (tree) 李1_03_01 (fruit) ... ... 禾1_03_02 (crop)
Pseudo-Sentences	<养1_11_02, 木1_07_01>, <养1_11_02, 李1_03_01>, ..., <养1_11_02, 禾1_03_02> <浇1_04_03, 木1_07_01>, <浇1_04_03, 李1_03_01>, ..., <浇1_04_03, 禾1_03_02> ... ... <耕1_02_01, 木1_07_01>, <耕1_02_01, 李1_03_01>, ..., <耕1_02_01, 禾1_03_02>			

Table 6: Demo of instance proliferation by similarity for the seed word "zhí shù" (*plant-tree*) (<yǎng1_11_02 mù1_07_01 >), where $\mathbb{C}_a$ and $\mathbb{C}_b$ refer to a set of similar MC C_a and C_b.

applying the template to all the disyllabic words in the lexicon.

Instance Proliferation

To exploit hierarchical information and tackle data sparseness problem, we go on to expand our lexicon based on similarity measurement of the achieved MCs. The similarity score between two MCs (C_1, C_2) in the hierarchy of MCs is defined as

$$sim(C_1, C_2) = \frac{2 \times |path(C_1) \cap path(C_2)|}{|path(C_1)| + |path(C_2)|} \quad (1)$$

where $path(C)$ is the set of all the tree nodes along the path from the root to the tree node C.

According to the threshold, for a certain MC C, a set of similar MCs can be achieved as $\mathbb{C}$. For every morpheme $a \in C_a$ and $b \in C_b$, if they ever happen to form a disyllabic word ab in the lexicon, we then generate more pseudo-sentences between $\mathbb{C}_a$ and $\mathbb{C}_b$. As for the missing knowledge of POS and word-formation, it is naturally assumed to be the same as the original one. In this way, the seed word can now be proliferated into n pseudo-sentences, where $n = |\mathbb{C}_a \times \mathbb{C}_b|$. For example, the proliferated instances we get for the seed word "zhí shù" (*plant-tree*) (<yǎng1_11_02 mù1_07_01 >) are listed in Table 6.

We set the threshold equal to 0.85 in the experiments and finally get a total of **54,880,628** pseudo-sentences from **52,108** real disyllabic words as the seeds input.

Data Training

Word2vec (Mikolov et al. 2013) is an algorithm to learn distributed word representation using a neural language model. It has two models, the continuous bag-of-words one (CBOW) and the skip-gram one.

In this paper, we train the distributed representation for morphemes based on CBOW, which aims at predicting the target word given context words in a sliding window. For morpheme embeddings on these 54,880,628 pseudo-sentences, we set the dimension to 20 and context window size to 3 to include all the rational knowledge when the MC is the target word.

Experimental Results and Evaluation

By the above approach we proposed, the rational knowledge of morphemes is now implanted into distributed representation. To evaluate such representation, intrinsic evaluation, such as paradigmatic and syntagmatic relations among morphemes, and extrinsic evaluation like word similarity measurement are taken into consideration.

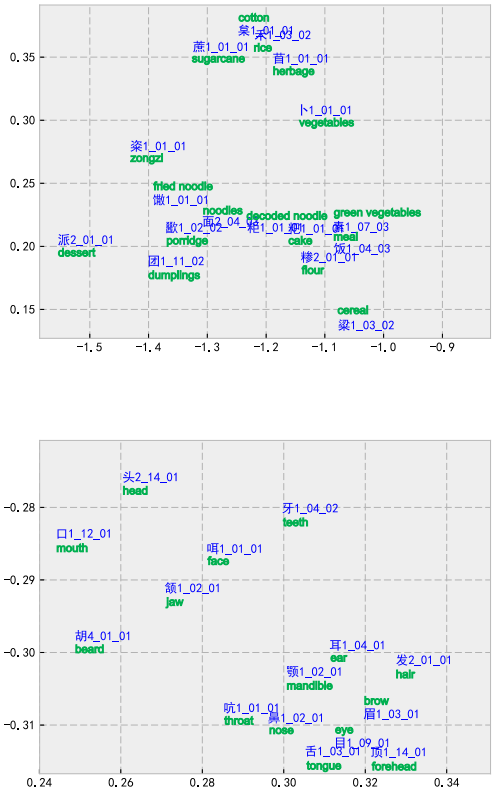

Figure 1: Illustration of paradigmatic MCs in 2D planes

Paradigmatic Relation Validation

To evaluate the effectiveness of the new approach, we use Principal Component Analysis (PCA) to conduct dimen-

Morpheme	MC	Nearest MCs
法1_07_01 (law)	{宪1_03_02, 制1_05_04, ..., 辟4_01_01}	{命2_03_01, 号3_04_01, ..., 禁1_04_03} (command) {禅1_02_02, 佛3_05_03, ..., 藏1_02_02} (Dharma) {墨1_10_08, 黥1_02_02, ..., 刑1_03_02} (punishment)
法1_07_02 (method)	{道1_12_03, 方3_02_01, ..., 筹1_03_03}	{典1_06_01, 度2_15_06, ..., 师1_07_02} (model) {揆1_04_02, 理1_07_02, ..., 谛1_02_02} (theory and principle) {准2_06_01, 标1_10_04, 杠1_07_07} (standard)
法1_07_04 (to emulate)	{则1_05_03, 宗1_08_05, ..., 象2_02_02}	{问1_06_01, 叩1_04_03, ..., 难1_02_02} (to address inquiries) {绎1_01_01, 验1_03_01, ..., 搜1_02_02} (to search) {造1_03_02, 编1_09_05, ..., 蔑2_01_01} (to make up)

Table 7: Different morphemes with identical character "fǎ" and their nearest MCs

MC	Top-Related MCs
{骏1_01_01, 马1_03_01, ..., 驹1_02_02} (horse)	{骏1_01_01, 马1_03_01, ..., 驹1_02_02} (horse) {镫1_01_01, 鞍1_01_01, ..., 鞒1_01_01} (saddle) {兵1_05_02, 军1_03_01, ..., 卒1_03_01} (soldier)
{鸡1_02_01, 鸭1_01_01,..., 鹅1_01_01} (fowl)	{仔3_01_01, 子1_13_08, ..., 雏1_02_02} (chick) {野1_07_04} (wild) {坤1_02_02, 母1_06_03, ..., 牝1_01_01} (female)
{牛1_04_01, 牦1_01_01, ..., 犊1_01_01} (cattle)	{乳1_05_03, 奶1_03_02} (milk) {牛1_04_01, 牦1_01_01, ..., 犊1_01_01} (cattle) {土1_07_01,垆1_01_01, ..., 壤1_03_01} (soil)

Table 8: Different MCs and their related MCs in word-building

sional reduction on morpheme embeddings to show paradigmatic relations as internal knowledge input. The results are illustrated in Figure 1.

We further take different morphemes with the identical character for example and list their corresponding MCs. Table 7 illustrates the nearest 3 MCs of each morpheme for observation. It can be observed that the MCs with similar meanings are naturally gathered together in general.

Syntagmatic Relation Validation

In addition to paradigmatic relations, we also explore the syntagmatic relations as internal knowledge input, i.e., which morphemes are more likely to form words.

For each MC, we predict its context words, which are the rational knowledge already input, and extract the most probable MCs in the context words.

We list some of the MCs and 3 of their top-related MCs in Table 8, from which it can be observed that given a specific MC, which MCs are prone to be involved in word-building. From the table, we notice that the cases of combination obtained from morpheme embeddings are quite consistent with human judgment. Take MC which means *horse* for example, many words can be formed between the MC and its top-related MCs, such as "jùn mǎ" (steed), "mǎ jū" (foal), "mǎ ān" (saddle) and "qí bīng" (cavalryman).

Word Similarity Measurement

The above are all intrinsic evaluation. For extrinsic evaluation, word similarity computation is an appropriate task. In order to compute the semantic distance between word pairs, many previous works take words as basic units and learn word embeddings according to external knowledge (contexts), ignoring the internal knowledge of words (morphemes and word-formations). However, as we argued, the internal knowledge also plays an important role in Chinese. The ideal way to measure semantic similarity may be to combine external and internal knowledge together.

Word-Formation Pattern	1st-MC	2nd-MC
后附加(Suffixation)	1	0
述补(Vreb-Compliment)	0.8	0.2
述宾(Verb-Object)	0.6	0.4
联合(Parallel)	0.5	0.5
单纯词(Noncompound)	0.5	0.5
定中(Modifier-Head)	0.45	0.55
状中(Adverb-Verb)	0.45	0.55
主谓(Subject-Predicate)	0.4	0.6
前附加(Prefixation)	0	1

Table 9: Weight assignments for different word-formation patterns

As the meaning of a word is contributed by the morphemes as its sub-units, which now refers to the MCs in the hierarchy, we assign different weights to the morphemes appearing in different word-formation patterns. For example, for *Modifier Head* structure, the head will contribute more to the meaning of the word, while for *Prefixation* structure, the prefix can hardly be related to the meaning of the word. Eventually, 9 types of word-formation pattern in the test sets (see description below) are assigned with different weights for the morphemes, as shown in Table 9.

Based on this, we try to obtain word embedding for each word by a weighted average of its n morpheme embeddings. It is calculated as $v_i = \sum_{k=1}^{n} w_{i_k} \times c_{i_k}$, where v_i stands for the word vector of the i^{th} word in the lexicon, w_{i_k} is the weight assigned by the above table and c_{i_k} is the morpheme embedding which the k^{th} character corresponds to. This is how our MC model will work on word similarity by purely exploiting internal knowledge.

As for CBOW and skip-gram, we use the corpus of *Baidu Encyclopedia*, which contains 203.69 million words in total. In the experiments, the dimension is set to 50, and the context window size is set to 5. Cosine similarity is applied to measure word similarity score for models of CBOW, skip-gram and MC individually. We also combine similarity scores obtained from the classical model (CBOW and skip-gram respectively) and MC model with the same weight assignment, namely the hybrid models of CBOW+MC and skip-gram+MC.

In the experiments, wordsim-296 (Jin and Wu 2012) and PKU-500 (Wu and Li 2016) are used as evaluation datasets. We extract the disyllabic words in the datasets and get a total of 141 word pairs in wordsim-296 and 232 word pairs in PKU-500 respectively. These serve as the test sets for word similarity measurement. Spearman's correlation ρ (Myers, Well, and Lorch 2010) is then adopted to evaluate all the outputs on the test sets. The experimental results of the 5 models are shown in Table 10.

Model	wordsim-296	PKU-500
CBOW	57.43	34.82
Skip-gram	62.17	40.19
MC	46.28	30.57
CBOW+MC	64.35	42.74
Skip-gram+MC	**67.58**	**45.91**

Table 10: Evaluation results on wordsim-296 and PKU-500 ($\rho \times 100$)

Note that our morpheme embeddings are trained with only 52,108 original pieces of semantic word-formation knowledge (approximately 2.79 MB of storage as in the experiments), without a corpus of data harnessed as before. The MC model, by purely exploiting such internal knowledge, alone achieves a fairly good performance, compared with the classical models. Furthermore, experiments on test sets show that the hybrid models of CBOW+MC and Skip-gram+MC, by exploiting external and internal knowledge, achieve significant improvements over the classical models by more than 5 Spearman scores or 8 percentage points. This indicates that both sources of knowledge are very valuable and highly complementary in expressing the meanings of words.

As for the combination for hybrid models, we set different weight assignments for similarity scores obtained from the classical model (CBOW and skip-gram respectively) and MC model to explore the cases. The correlation between internal knowledge adopted and performance is shown in Figure 2. Considering the global optimum for both test sets, the most ideal weight assignment for internal and external knowledge is $0.35 : 0.65$ in the experiments, which has been adopted and yielded the results in Table 10.

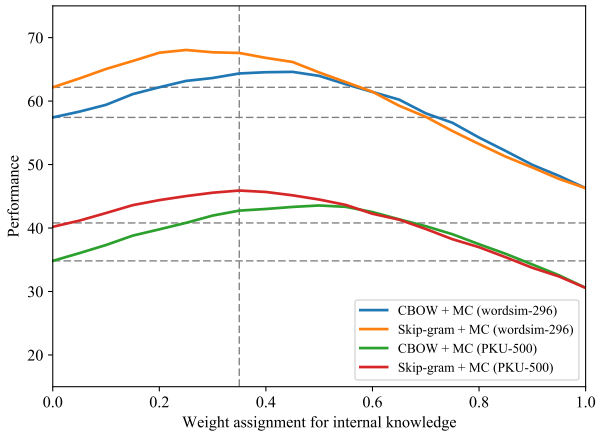

Figure 2: Correlation between weight assignment for internal knowledge and performance

Conclusion

In this paper, after constructing the Chinese lexical and semantic ontology based on word-formation, we try to implant the structured rational knowledge into distributed representation at morpheme level without using any text corpus. For evaluation, we validate the paradigmatic and syntagmatic relations of morpheme embeddings, and apply the obtained embeddings to word similarity measurement, achieving significant improvements over the classical models by more than 5 Spearman scores or 8 percentage points.

The key contributions of this work are as follows: (1) We, for the first time, put forward an approach to implanting the structured rational knowledge into distributed representation by merely using the lexicon. As the form of such piece of knowledge is almost common to most knowledge bases, we actually present an inspiring way of obtaining distributed representation for the desired language units described in the lexicons. (2) For parataxis languages like Chinese, morphemes as the basic units play an important role in expressing the exact meanings of words. It is a convenient way by obtaining unambiguous morpheme embeddings simply based on the descriptions in the lexicon, which naturally avoids heavy disambiguation in the corpus as before (Luo et al. 2018a; 2018b).

Currently, we focus on the original meanings of Chinese disyllabic words, which make up the majority of the vocabulary of CCD. However, some words may have metaphoric or transferred meanings, or comprise of more than two characters. Such work is in progress in our group according to the solution. Also, to gain better word embeddings for certain tasks, the topic of compositionality of word embeddings is reserved for further research. After completion of these works, we hope to release the *COOL* system.

Acknowledgments

This work is supported by National Social Science Fund of China (General Project, No. 16YY137). We thank the anonymous reviewers for their helpful comments. The corresponding author of this paper is Yang Liu.

References

Bartusiak, R.; Augustyniak, Ł.; Kajdanowicz, T.; Kazienko, P.; and Piasecki, M. 2017. Wordnet2vec: Corpora agnostic word vectorization method. *Neurocomputing*.

Chao, Y. R. 1968. *A grammar of spoken Chinese*. USA: CA: University of California Press.

Chen, X.; Xu, L.; Liu, Z.; Sun, M.; and Luan, H.-B. 2015. Joint Learning of Character and Word Embeddings. In *IJCAI*, 1236–1242.

Chou, Y.-M. 2005. *Hantology: The Knowledge Structure of Chinese Writing System and Its Applications*. PhD dissertation, National Taiwan University.

Dong, Z.; Dong, Q.; and Hao, C. 2007. Theoretical Findings of HowNet. *Journal of Chinese Information Processing* 21(4):3–9.

Fu, H. 1981. The Relationship between Lexical Meaning and Meanings of Morphemes Constituting Words. *Lexicographical Studies* (01):98–110.

Fu, A. 2003. Word Formation and the Recognition of Compounds in Chinese Language Understanding. *Applied Linguistics (China)* (04):25–33.

Jin, P., and Wu, Y. 2012. Semeval-2012 task 4: Evaluating Chinese word similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 374–377. Association for Computational Linguistics.

Lazaridou, A.; Marelli, M.; Zamparelli, R.; and Baroni, M. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1517–1526.

Liu, Y.; Lin, Z.; and Kang, S. 2018. Towards a Description of Chinese Morphemic Concepts and Semantic Word-Formation. *Journal of Chinese Information Processing* 32(2):11–20.

Liu, S. 1990. *Hanyu Miaoxie Cihuixue (Descriptive Lexicology of Chinese)*. China: Beijing: The Commercial Press.

Luo, F.; Liu, T.; He, Z.; Xia, Q.; Sui, Z.; and Chang, B. 2018a. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 1402–1411.

Luo, F.; Liu, T.; Xia, Q.; Chang, B.; and Sui, Z. 2018b. Incorporating glosses into neural word sense disambiguation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2473–2482. Association for Computational Linguistics.

Luong, T.; Socher, R.; and Manning, C. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 104–113.

Matthews, P. H. 1972. *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*, volume 6. CUP Archive.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G., and Fellbaum, C. 1998. Wordnet: An electronic lexical database.

Myers, J. L.; Well, A.; and Lorch, R. F. 2010. *Research design and statistical analysis*. Routledge.

Niu, Y.; Xie, R.; Liu, Z.; and Sun, M. 2017. Improved Word Representation Learning with Sememes. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 2049–2058.

Pustejovsky, J. 1995. *The Generative Lexicon*. USA: Mass: MIT Press.

Tian, Y., and Liu, Y. 2016. Lexical Knowledge Representation and Sense Prediction of Chinese Unknown Words. *Journal of Chinese Information Processing* 30(6):26–34.

Tissier, J.; Gravier, C.; and Habrard, A. 2017. Dict2vec: Learning word embeddings using lexical dictionaries. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, 254–263.

Wu, Y., and Li, W. 2016. Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word similarity measurement. In *Natural Language Understanding and Intelligent Applications*. Springer. 828–839.

Xu, T. 2004. The 'Character' in Chinese Semantics and Syntax. *Preliminary Studies in Chinese Research Methodology* 276–301.

Yin, B.-y. 1984. Hanyu Yusu De Dingliang Yanjiu (A Quantitative Study of Chinese Morphemes). *Zhongguo Yuwen* 05:340.

Zhu, D. 1982. *Yufa Jiangyi (Lectures on Grammar)*. China: Beijing: The Commercial Press.