

Received August 13, 2019, accepted September 22, 2019, date of publication October 2, 2019, date of current version October 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2945061

Semantic Relata for the Evaluation of Distributional Models in Mandarin Chinese

HONGCHAO LIU¹, EMMANUELE CHERSONI², NATALIA KLYUEVA², ENRICO SANTUS³, AND CHU-REN HUANG²

¹School of Literature, Shandong University, Jinan 250100, China

²Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong

³Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA

Corresponding author: Hongchao Liu (jiye12yuran@126.com)

This work was supported by the Shandong University under Grant 11070079614059 and Grant 10000086393101.

ABSTRACT Distributional Semantic Models (DSMs) established themselves as a standard for the representation of word and sentence meaning. However, DSMs provide quantitative measurement of how strongly two linguistic expressions are related, without being able to automatically classify different semantic relations. Hence the notion of semantic similarity is underspecified in DSMs. We introduce Evaluation-MAN in this paper as an effort to address this underspecification problem. Following the EVALution 1.0 dataset for English, we present a dataset for evaluating DSMs on the task of the identification of semantic relations in Mandarin Chinese. Moreover, we test different types of word vectors on the automatic learning of these semantic relations, and we evaluate them both in a unsupervised and in a supervised setting, finding that distributional models tend, in general, to assign higher similarity scores to synonyms and that deep learning classifiers are the best performing ones in the identification of semantic relations.

INDEX TERMS Computational semantics, ontologies, relation classification, semantic relations, vector space models, lexical resources.

I. INTRODUCTION

Discovery and processing of semantic relations is a crucial issue shared by cognitive language processing, language technology and knowledge engineering [1]. The distributional nature of linguistic relations is an important premise of Distributional Semantics, one of the most successful research paradigms in the recent history of Natural Language Processing (NLP). The models based on the so-called *Distributional Hypothesis* [2], generally known either as *Distributional Semantic Models* (DSMs) or as *Vector Space Models* (VSMs) [5], [6], assume that linguistic items can be represented as *vectors*, whose dimension values depend on their distributional behaviour in text. The most recent newcomers in the family of the distributional models, the *word embeddings* [7], [8], have been shown to be a powerful and easy-to-use solution to the need of automatically deriving semantic representations from language corpora, leading to improved performances both in classical NLP [9] and in psycholinguistic tasks [10].¹

The associate editor coordinating the review of this manuscript and approving it for publication was Wajahat Ali Khan².

¹*Distributional models* and *Word Embeddings* have been used interchangeably in this paper and, unless specified otherwise, they refer to the same type of Vector Space Model.

An important role in this framework is played by *semantic similarity* measurement [6], [12]. Since words occurring in the same contexts are assumed to have related meanings, and distributional models can indeed retrieve similar words simply by measuring distances between vectors. It is easy to understand why this feature is potentially very appealing for several information processing tasks, as well as for the construction and enrichment of lexical and ontological resources (thesauri, electronic dictionaries etc.) [13].

However, semantic vectors have also an important limitation from this point of view, that is, the concept of distributional similarity conflates together different ways in which word meanings can be related. For example, the meaning of *house* is surely related to the meaning of *building* and to the meaning of *roof*, but nobody would say that the relation holding between them is the same. From these considerations, an emerging line of research developed around the issue of using vector representations for the identification of semantic relations (for a more extensive overview, see [13], [14]). Yet such research crucially relies on the public availability and shareability of datasets and resources for both system development and evaluation, especially for languages other than English.

In the present contribution, we introduce a new version of Evaluation-MAN [15], a dataset of semantically related word pairs (*relata*) for evaluating vectors in Mandarin Chinese. After presenting the main features of the dataset, we also describe a preliminary evaluation that we conducted on it, and we compare the performance of the models with the analogue dataset already available for the English language [16]. To the best of our knowledge, this is the first time that such an evaluation is carried out for Mandarin Chinese. Moreover, the proposed dataset will be made publicly available.²

II. SEMANTIC RELATIONS

In this section, we briefly define and describe the semantic relations described in Evaluation-MAN 2.0. These relations have been chosen as they have been widely investigated in linguistics and cognitive science [1], on the one hand, while on the other hand they are the most commonly represented in similar resources for the evaluation of NLP systems [16]–[18].

A. SYNONYMY

Words expressing similar meanings are commonly referred to as synonyms, such as *rich-wealthy* or *ask-enquire*. We adopt Cruse's [19] notion of 'cognitive' synonymy, which states that words that are interchangeable in some but not all contexts and designate the same concept are synonyms. Several computational approaches to synonymy detection rely on distributional vectors, and use vector proximity as a proxy for meaning similarity: since synonyms are the related words with the highest degree of distributional similarity, and they tend to share the most of the contexts.

Other approaches to synonymy detection are based on lexicographic resources that propose instead more fine-grained distinctions, with the Wordnet lexical network being the most famous example [20], [21]. However, for simplicity, our resource will make reference to the more generic synonymy.

B. ANTONYMY

Antonymy can be defined as a semantic relation expressing opposition and contrast, as in word pairs like *big* and *small*, or *up* and *down*. Although it might seem counter-intuitive, antonyms are generally very similar words: according to Cruse [19], they are actually similar in all dimensions of meaning except one (e.g. *giant* and *dwarf* differ for the dimension of size). Such a feature of antonyms can be observed also in their distributional patterns, as they tend to occur in similar contexts, and thus their discrimination from synonyms is a non-trivial problem for DSMs [22].

C. HYPERNYMY-HYPONYMY

Hypernymy (also known as IS-A) defines the relation between a broader semantic category and semantically more specific categories, e.g., *bird* is a hypernym of *swal-*

low, *seagull*, *sparrow*, *eagle*, its hyponyms. Generally, each hyponym inherits all the features and properties of its hypernym, and adds at least one feature that distinguishes it from the hypernym itself and from the other hyponyms (*coordinates* or *cohyponyms*) of the same super-ordinate (*inheritance property of hyponymy*) [18]. Since two cohyponyms inherit the same set of properties from their common hypernym, they are sometimes said to be attributionally similar.

Similarly to synonymy and antonymy, hypernymy and hyponymy have been widely studied in Natural Language Processing, as the recognition of such relations between lexical items is fundamental for many applications related to textual entailment, automatic summarization and ontology extraction [23], [24].

D. COHYPONYMY

A semantic relation strictly connected to hypernymy is cohyponymy, which concerns words that are semantically similar as they share a common hypernym, as in the case of *dog*, *cat* that share the hypernym *animal*. Previous evaluation showed that, in terms of distributional patterns, distinguishing between synonyms, cohyponyms and hypernyms can be a non-trivial task [18]. Cohyponyms are also referred to as *coordinates* in the literature on DSMs [17].

E. MERONYMY AND HOLONYMY

Meronymy relates parts to a whole, that is, it relates meronyms, such as *page*, *cover*, *binding*, to their holonyms, such as *book*. Among the semantic relations, it is probably the one connecting the least similar terms, and it is often expressed by means of lexico-syntactic patterns such as *x is part of y* [25]–[27].

III. RELATED WORK

The research around DSMs and semantic relations went through different stages. In a first stage, researchers focused on the type of metric, trying to find alternatives to cosine similarity that can be used to set apart a specific semantic relation from the other ones, e.g. hypernymy [23], [28], synonymy [29] or antonymy [22]. In parallel, the first datasets for evaluating the identification of semantic relations were being released, including relations such as hypernymy, cohyponymy and antonymy [16], [17], [28].

In a second phase, following the increasing popularity of publicly-available frameworks for training word embeddings such as Word2Vec [7] and Glove [8], the focus quickly shifted on the usage of these vectors as features for supervised classifiers. Some of these methods train classifiers directly on pairs of vectors [18], [30], [31], while others compute DSMs-based distance metrics first and then use them as features [32].

Some of the most recent contributions proposed even more sophisticated classification methods. [33], [34] aim at integrating word embeddings with information coming from lexical patterns, which proved to be extremely accurate for detecting relations. Other researchers introduced modifications to the structure of the vector spaces with

²https://github.com/LHongchao/EVALution_MAN/blob/Version2.0/pairs.txt

the goal of identifying a specific type of semantic relation, for example by modifying the objective function of the Word2Vec training to inject external knowledge from a lexical resource (e.g. WordNet) [35], [36], or by adding an extra postprocessing step that projects the word vectors into a new space, expressly specialized for modeling the target relation [37].

However, one thing to be said about these contributions is that they mostly tried to address one relation at a time, with rare attempts of tackling the problem of relation classification in a multiclass setting. The shared task organized in coincidence with the COLING workshop on Cognitive Aspects of the Lexicon [13] was one of the few exceptions, and the low results achieved by most systems (the top F-score being 0.44) showed the inherent difficulty of distinguishing between multiple relations at once. For such a task, the data had been extracted from Evaluation, a dataset for English introduced by Santus *et al.* [16] and containing more than 7500 word pairs labeled with the semantic relations holding between them. The construction process of the proposed resource closely follows the one adopted for the original English dataset.

IV. EVALUTION-MAN 2.0

A. DATA COLLECTION

There are three main steps involved in the construction of EVALution-MAN 2.0, starting with the collection of candidates for the related word pairs. These collected pairs, which instantiate different semantic relations, are then checked for their reliability by several human raters. Words are also labeled with part-of-speech information in order to enrich the features that can be used for the evaluation and application of the dataset using DSMs. We apply two different methods, namely extraction and elicitation tasks, to collect candidates for the related word pairs. These two ways are complementary in their efficiency and reliability [38].

On the one hand, automatically extracting the pairs from existing resources, such as the Academia Sinica Bilingual Ontological WordNet [11] and the Chinese WordNet [39], can be quick and easy. Data extracted from Sinica BOW has the advantage of being able to be compared with English data from Princeton WordNet, while the data from the Chinese WordNet have the advantage of manual checking of the relations. On the other hand, not all the relation pairs in these resources are fully verified and inconsistencies can be expected due to the nature of manual tagging. The inconsistencies are caused by several reasons. Mostly, they are caused by the uncertainty about the relations between these pairs, because of insufficient contextual information or under-specification of meanings [38].

For example, some entries appear in abbreviated forms in the Chinese WordNet [39], and hence the relations involving these terms can hardly be recognized unless additional background information or context is available.

(1) 亚热带 (*ya4re4dai4*, subtropics) is a kind of 文化 (*wen2hua4*, culture).

The word 亚热带 (*ya4re4dai4*, subtropics) is an abbreviation of 亚热带文化 (*ya4re4dai4 wen2 hua4*, subtropical civilization) in this entry in the Chinese WordNet [39]. If no context is provided, it will be difficult for human raters to grasp its meaning and as a result, its relation with 文化 (*wen2hua4*, civilization) cannot be easily confirmed. Another factor contributing to uncertainty is that some words in the Chinese WordNet are rarely used in modern Mandarin Chinese as they are archaic words, used only in ancient Chinese. It is natural for raters to classify the relations between these words as unacceptable since they have never used them in their daily lives. The pair 俄 (*e2qin3*, instantly)-转眼间 (*zhuan3yan3jian1*, in a blink of an eye), which has been excluded from the current dataset because of rating inconsistency, is presented here as an example

(2) 俄 (*e2*, instantly) is a kind of 转眼间 (*zhuan3yan3jian1*, in a blink of an eye).

The word 俄 (*e2qin3*, instantly) is a word from classical Chinese and it is rarely, if ever, used in modern Chinese. Hence, for some of the raters, it is hard to confirm its meaning and thus the relation pair above was rated as unacceptable in the judgement task. Because of the reasons mentioned above, we introduced an elicitation task to collect more consistent candidates for our related word pairs. However, words created by participants in elicitation tasks could be missing in corpus, resulting in the absence of word embedding vectors and inability to predict them in the evaluation stage. Thus, these two methods must be combined to reach a balance between the quality and coverage of the word pairs in the dataset. The extraction method starts from processing the data from the Chinese WordNet. There are more than 20,000 word relation pairs in this knowledge source, but most of them are duplicated ones or contain null values. The following stage is to filter out these pairs, and this step left us only 10,000 pairs for the next stage. After that, we decided to rule out those pairs including *relata* appearing in less than two kinds of relations. In other words, we only kept the *relatum* with two or more relations with other *relata*, in order to increase the variability of *relata* in the dataset, following the criteria of the original EVALution [16]. More than 5,400 pair candidates were then selected in this step, before validation.

The second method, namely the elicitation task, involves two main steps, starting from picking words as the primes. 100 common nouns representing basic, concrete concepts were selected firstly (a similar initial selection was carried out by Baroni and Lenci for the BLESS dataset [17]). These words were then provided to our raters to produce semantically related words for different relations. These words are referred to as *relata* [17]. The three annotators are volunteers from China and they are all native Mandarin speakers with a Ph.D degree in linguistics. In this way, the quality of the produced words can be guaranteed. These 100 target words were then embedded into carrier sentences with descriptions about the desired relation type and blanks to fill in their missing *relata*. These sentences were built to elicit six kinds of relations, including holonymy, meronymy, hypernymy,

TABLE 1. Example elicitation form.

Target word	Relation	Answer
血(<i>xue3</i> , blood)	Synonym	血液(<i>xue3ye4</i> , blood)
血(<i>xue3</i> , blood)	Antonym	水(<i>shui3</i> , water)
血(<i>xue3</i> , blood)	Hyponym	鲜血(<i>xian1xue3</i> , fresh blood)
血(<i>xue3</i> , blood)	Hypernym	组织液(<i>zu3zhi1ye4</i> , tissue fluid)
血(<i>xue3</i> , blood)	Meronym	红细胞(<i>hong2xi4bao1</i> , red blood cell)
血(<i>xue3</i> , blood)	Holonym	身体(<i>shen1ti3</i> , body)

TABLE 2. The collected chinese data.

Word X	Relation	Word Y	Vote
足球(<i>zu2qiu2</i> , football)	Hypernym	球(<i>qiu2</i> , ball)	2
熊(<i>xiong2</i> , bear)	Hyponym	狗熊(<i>gou3xiong2</i> , black bear)	1
冰(<i>bing1</i> , ice)	Antonym	火(<i>huo3</i> , fire)	1

hyponymy, antonymy and synonymy, and thus each target word can form six elicitation sentences. Some examples are shown in Table 1. Note that senses are not pre-defined or prompted in this study. That is, the relata are created based on words and native speakers' intuition-based assignment of meaning. A good example is the assignment of antonym for the *blood/water* pair. In this context, the antonym relation assigned by the subjects depends on the metaphorical meanings of 'with shared kinship heritage' (blood) and 'without shared kinship heritage' (water). This is an important feature of the design of relata vs. WordNet or semantic taxonomy. A relatum contains all salient semantic relations between word pairs and allows us to build all possible links around a perceived word (and not the abstract sense of linguistic word with a single sense).

The three raters were asked to fill in the blanks within the 600 sentences (100 words, 6 relations). We did not limit the number of related words that can be inserted for each sentence, and thus the rater can supply more than one answer for each target relation. After the elicitation task, more than 1,000 word pairs were constructed. It should be noticed that duplicated ones were counted as only one item, although production frequency was attached (see Table 2).

B. RELIABILITY CHECK

Once we finished with the collection of candidates, we merged the pairs into one raw dataset obtaining a total of 6,600 related word pairs. These pairs, together with their relation type and the relation description, were finally validated in a judgement task by different annotators who had not participated in the previous elicitation tasks. The raters of this last group are all adult native speakers of Mandarin aged from 18 to 20 in university and their majors vary from arts to science. Each pair was checked for reliability by five annotators. When supplying pairs for checking, five possible ratings were presented to them ("totally agree", "agree", "don't know", "disagree" and "totally disagree") and they were asked to choose one of these options. Two extra options were provided ("don't know X" and "don't know Y"), for the case in which

TABLE 3. Sentence judgement task result (part).

Sentence to be checked	Results		
狗(<i>gou3</i> , dog) has the same hypernym with 龙(<i>long2</i> , dragon)	3	2	0
存(<i>cun2</i> , survive) has an opposite meaning with 亡(<i>wang2</i> , die)	0	3	2
弟兄(<i>di4xiong</i> , brother) is a kind of 关系(<i>guan1xi4</i> , relationship)	3	2	0

Note: For the "Results", the three columns represent "totally agree", "agree", "don't know". We also have other four options including "disagree", "totally disagree", "don't know X" and "don't know Y" respectively. Since they are all zero for these three relation pairs, we intend to ignore these columns here in order to display a readable table with smaller width.

TABLE 4. Part-of-speech candidates.

Word X	Candidates for X	Word Y	Candidates for Y
少(<i>shao4</i> , young)	VJ_411VH_7571Neqa_26	老(<i>lao3</i> , old)	Na_58ID_801VH_1051
歌星(<i>ge1xing1</i> , singer)	Na_41	人(<i>ren2</i> , human)	Na_23990
显示(<i>xian3shi4</i> , show)	VK_785INv_53	显示出(<i>xian3shi4chu1</i> , show)	VK_73

the annotators did not know the meaning of one of the words in the pair.

Finally, the rating results for each pair were collected and combined as shown in Table 3.

In order to keep consistency with the original EVALution dataset, we kept only the pairs that received either "totally agree" or "agree" ratings, and this procedure left us with 4091 related word pairs. At the following stage of the construction process, we introduced two new expert annotators (university assistant professors doing research on lexical semantic relations) to double-check the dataset and more than 100 pairs rated as unreliable by both the expert annotators at the same time were ruled out.

C. GENERATION OF RANDOM RELATA

The competition of the CogALex shared task was organized as a two-step classification task: first, the systems were asked to discriminate between related word pairs and randomly paired words, and then to distinguish between the different semantic relations [13]. In order to allow a similar evaluation strategy, for each related pair in the dataset, we generated a new pair by randomly pairing the first word with another unrelated word in the dataset. We labeled such word pairs as RANDOM relata.

D. PART-OF-SPEECH TAGGING

Except the relation type, we also added part-of-speech (hence, PoS) information for each relatum to supply more features for the training and evaluating based on this dataset.

In this phase, automatic mapping and manual checking were both applied. Firstly, the PoS distribution of each relata is extracted from Sinica corpus [3]. The reason why we choose this resource as a supplier for PoS candidates is that all the words in it are annotated with PoS tags manually and hence a high rating quality can be guaranteed. These candidates' PoS information, together with the frequency of each tag, are attached to the relatum as following in Table 4.

After that, the same expert annotators who had been involved in the last step of reliability check were asked to choose among these relata's PoS candidates. In this stage, no pairs would be abandoned even though there is incon-

TABLE 5. Distribution of classes in the dataset.

Relation	Pairs	Relata (without PoS)	Relata (with PoS)
Antonym	639	365	382
Synonym	658	414	420
Hypernym	1858	1500	1538
Cohyponym	362	181	186
Meronym	411	334	337
Random	3918	2219	2235
Total	7846	3124	3263

sistency between the annotation of the linguistics experts. Instead, these inconsistent pairs were further discussed by them based on sentences extracted from the Sinica corpus [3], which would provide more accurate contextual information and they were required to consistently choose the most appropriate one from these options. The final dataset includes 7846 word pairs (the related pairs, plus the random pairs). The number of pairs per relations can be seen in Table 5.

V. DATASET EVALUATION

For our experiments, we created vector representations for all the words in the dataset, using two very popular word embeddings libraries: the Word2Vec library by Mikolov and colleagues [7] and the FastText library by Bojanowski *et al.* [40]. For Word2Vec, we tested both the Skip-Gram and the Continuous Bag-of-Words model, generally obtaining better performances with the former.

The models for Chinese have been trained on the tagged version of the Chinese Gigaword [41], with the *window size* set to 5, and *min count* set to 1. All the other hyperparameters are the default ones of the Word2Vec and the FastText packages [7], [40]. Notice that the Chinese Gigaword Corpus contains data both from China and Taiwan. Our training used data from both varieties after normalizing the orthography.

We trained vectors with different dimensionality, ranging from 300 to 500, but in the end we report results only for the 300-dimensional ones, since we did not observe any significant performance difference.

In the following two subsections, we present the results of our preliminary experiments on the dataset: an unsupervised evaluation, based on cosine similarity distances and on Mean Average Precision [42]; a supervised evaluation, in which we used our word embeddings as features for some standard supervised classifiers, and we assessed performance in terms of precision, recall and F1-score. For the sake of comparison, we also illustrate the results for a supervised evaluation task on the English dataset by Santus *et al.* [16], on the basis of pretrained Word2Vec vectors made available on the Google Code archive.³

A. UNSUPERVISED EVALUATION

The goal of the unsupervised evaluation is to see which type of semantic relata tend to be closer in the distributional space. With this purpose in mind, we adopt the Average Precision

TABLE 6. Mean average precision scores per relation for each model on the chinese dataset.

Relation	Skip-Gram	CBOW	FastText
Synonym	0.12	0.07	0.15
Antonym	0.07	0.08	0.03
Meronym	0.03	0.04	0.01
Hypernym	0.02	0.03	0.06
Cohyponym	0.02	0.05	0.02
GLOBAL	0.05	0.05	0.06

metric, adapted from the field of Information Retrieval (IR). This metric is generally used in IR systems to evaluate the ranked documents that have been returned for a specific query, and it is designed to assign higher values to the rankings in which the relevant documents are at the top (recall) and the non-relevant ones are at the bottom (precision) [22]. In our case, the semantic relations in the dataset represent our queries: for each target word and each target relation, we create a rank according to the cosine similarity scores with all its related words plus the RANDOM relata, and then we compute the Average Precision values for each semantic relation. Finally, we compute Mean Average Precision (MAP) by averaging the values for each semantic relation (the RANDOM relation being obviously excluded).

The results are shown in Table 6. As it was predictable, the performance of unsupervised models merely based on cosine similarity is very weak when it comes to discriminating between different relation types [43]. Also not surprisingly, for both the Skip-Gram and the FastText model the synonymy relation is by far the highest-scoring one, generally followed by antonymy.

We also computed a Wilcoxon rank sum test to verify whether the models are still able to discriminate between related and random pairs.⁴ The answer was positive: for all our models, the cosines assigned to the related pairs were significantly higher ($p < 0.001$). This can also be observed in the boxplots in Figures 1, 2 and 3, which show, for each model, the cosine similarity values per semantic relation.

B. SUPERVISED EVALUATION

For the supervised evaluation, we concatenated the vectors of each word pair and used them as input features for six supervised classifiers [30]. We implemented a Support Vector Machine, a Logistic Regression and a Random Forest with the Python scripts of the scikit-learn library [44]. We also implemented three deep learning systems with the Keras library [45]: a Feedforward Neural Network and a Convolutional Neural Network and a LSTM. For the sake of comparison and reproducibility, Table 7 contains the description of the classifiers and the parameter settings.

The results of our experiments on the Chinese dataset are

³<https://code.google.com/archive/p/word2vec/>

⁴The p -values have been computed with the R statistical software.

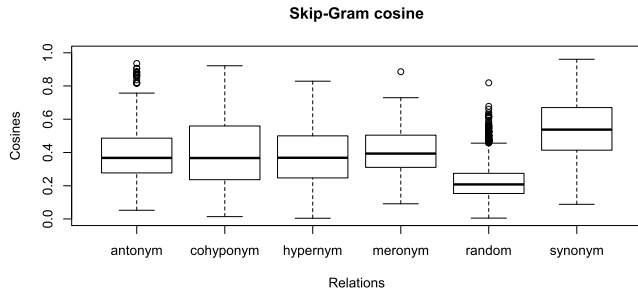


FIGURE 1. Cosine similarity values per relation for the vectors trained with Skip-Gram with negative sampling (Chinese dataset).

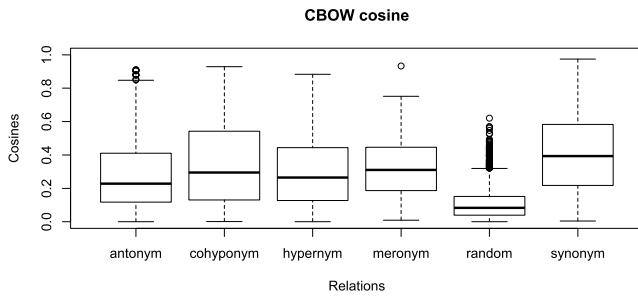


FIGURE 2. Cosine similarity values per relation for the vectors trained with the continuous bag-of-words (Chinese dataset).

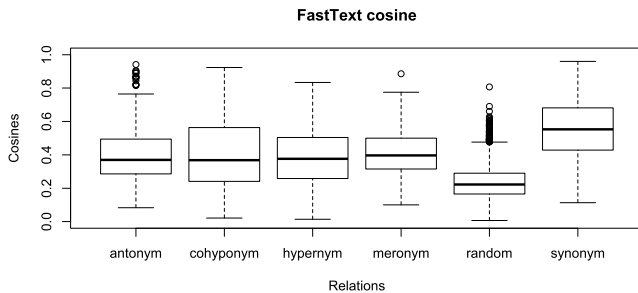


FIGURE 3. Cosine similarity values per relation for the vectors trained with FastText (Chinese dataset).

shown in Table 8: the F1-scores (weighted average) are organized by embedding type. As it can be seen from the table, the performance of the embeddings is generally similar across classifiers. The main exception is the LSTM, which struggles in all settings but achieves a remarkable improvement only when using FastText embeddings. The FastText embeddings are those delivering the most solid performance, since the systems using them as input features always obtain the top score. It should be pointed out that FastText embeddings are trained similarly to traditional Word2Vec vectors, but they incorporate subword information in the word representation. These results suggest that such a feature is useful for the relation classification task: in Chinese, where each character is a morpheme or a quasi morpheme, the training of FastText on character n-grams will allow to access rich information not available to vectors simply trained on word forms. For instance, given the dominance of compounding in Mandarin Chinese, sense-related words often share a compound root,

TABLE 7. Models and Parameter Settings. All the other parameters correspond to the default values in the scikit-learn package, in the case of the first three classifiers, and to the default values of Keras in the case of the last two neural models. The dimensionality of the input features is assumed to be 600 (the length of two concatenated vectors). As some global settings for all neural network classifiers, we set *epochs* = 15 and *batch_size* = 32.

Model	Parameters
Support Vector Machine	C=500
Logistic Regression	random state=0, solver="lbfgs" multi class="multinomial"
Random Forest	number of estimators=10
Feedforward Neural Network	2 Dense layers with 300 units activation=relu, 1 Dense layer with 6 output units activation=softmax loss= sparse categorical crossentropy optimizer=adam
Convolutional Neural Network	1 Flattened convolutional 1D layer with 300 filters and the kernel size is 2 activation=relu, 2 Dense layers with 200 and 300 units, activation=relu and dropouts are 0.2 and 0.5 loss= sparse categorical crossentropy optimizer=adam
LSTM	1 LSTM layer with 120 units input shape = (1,600) and activation=relu 1 LSTM layers with 10 units, activation=sigmoid loss= sparse categorical crossentropy optimizer=adam

TABLE 8. F1-scores (weighted average across relations) for the different systems per type of vector on the Chinese dataset. The top score for metric is in bold.

Model	Skip Gram	CBOW	FastText
Support Vector Machine	0.68	0.6	0.69
Logistic Regression	0.61	0.57	0.62
Random Forest	0.58	0.59	0.6
Feedforward Neural Network	0.72	0.69	0.73
Convolutional Neural Network	0.71	0.7	0.72
LSTM	0.15	0.11	0.48

TABLE 9. Precision, Recall and F1-score for all models with FastText vectors, Chinese dataset (weighted average across relations). The top score for metric is in bold.

Model	Precision	Recall	F1-Score
Support Vector Machine	0.68	0.7	0.69
Logistic Regression	0.63	0.6	0.62
Random Forest	0.7	0.55	0.6
Feedforward Neural Network	0.74	0.72	0.73
Convolutional Neural Network	0.73	0.71	0.72
LSTM	0.65	0.42	0.48

which will be easily captured by a vector trained on character n-grams, but not by one trained on words.

In Table 11, we present a comparison of Precision, Recall and F1-scores achieved by all classifiers on the basis of the FastText vectors. A deep learning system, the Feedforward Neural Network, turns out to be the best performing one,

TABLE 10. Precision, Recall and F1-score for all models with Skip Gram vectors [7], original English dataset (weighted average across relations). The top score for metric is in bold.

Model	Precision	Recall	F1-Score
Support Vector Machine	0.52	0.34	0.39
Logistic Regression	0.44	0.27	0.33
Random Forest	0.43	0.25	0.31
Feedforward Neural Network	0.49	0.45	0.47
Convolutional Neural Network	0.52	0.50	0.50
LSTM	0.52	0.22	0.31

TABLE 11. Precision, Recall and F1-score for all models with FastText vectors [7], original English dataset (weighted average across relations). The top score for metric is in bold.

Model	Precision	Recall	F1-Score
Support Vector Machine	0.45	0.2	0.27
Logistic Regression	0.47	0.23	0.3
Random Forest	0.46	0.27	0.33
Feedforward Neural Network	0.49	0.47	0.48
Convolutional Neural Network	0.48	0.48	0.48
LSTM	0.48	0.15	0.22

closely followed by the Convolutional Neural Network and by the Support Vector Machine, which is the most efficient among the non-deep learning classifiers. These three systems have a strong advantage over the competitors, especially in terms of average recall. We can also observe that, for all metrics, LSTM is the weakest classifier. This is not totally unexpected, since LSTMs are generally used for modeling data with a sequential structure, which is not the case for the word pairs of our relation dataset.

As we anticipated, we ran the same evaluation also on the original Evaluation dataset introduced for English by Santus et al. [16]: we used Skip Gram and FastText vectors as input features, since they turned out to be the best performing models. The scores are quite similar to the ones obtained for the Chinese data: the Convolutional and the Feedforward Network are again the best performing classifiers, with an even larger margin over all the competitors. Interestingly, the results for the English dataset show that FastText vectors do not lead to the same improvements and that the scores obtained with the Skip Gram model are often better. This finding reinforces the idea that subword information is extremely useful for the task on Chinese because of the peculiarity of this language: in the Chinese writing system, the semantic components of a word can arise from both the character and the sub-character level [46]. In other words, subwords in Chinese typically contribute to meaning, while subwords in English typically do not.⁵

To gain more insight in the classification performance,

⁵On the topic of semantics as the orthographically relevant level for the Chinese writing system, see also [47].

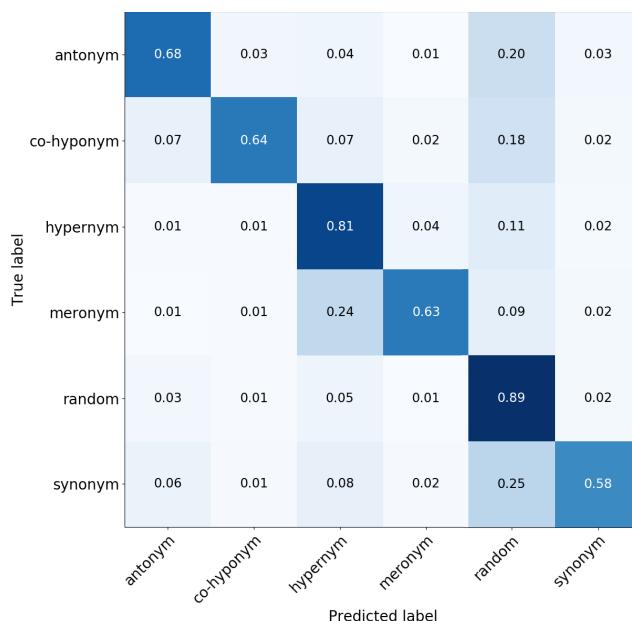


FIGURE 4. Confusion matrix from the classifier with the best performance (trained on fast vectors and using feedforward neural network classifier) on Chinese dataset.

in Figure 4 we report the confusion matrix for the top system for the Chinese dataset, the Feedforward Neural Network. Interestingly, the classifier seems to do best on hypernymy, which is an asymmetric and hierarchical relation, while it has more difficulty in identifying synonyms. For all relations except for meronyms, the random relata, i.e. randomly generated non-relation pairs, are the main source of confusion. Meronyms, on the other hand, are often confused with hypernyms, which is understandable since both of them are a kind of logical inheritance relation.

The above results have two important implications. First, it suggests that the flat relata model might be a more robust model for semantic relation processing than the tree structure of lexical networks (such as WordNet) or ontologies. Recall that networks and ontologies rely on both ‘IS_A’ type relation as well as ‘YIS/EQUAL_TO’ relation to build a tree without cross-branching. Our data show the best results for hypernymy and the worst results for synonymy, suggesting that the synonym relation, as defined in WordNet, might not be strong enough to justify equivalence relations. Without the equivalence relation, it will be difficult to construct a valid tree. A relata model addressing only local binary relations has no such modelling constraints and can accommodate all possible relation pairs. On the other hand, a relata model also has the flexibility of allowing domain or usage specific terms to form a conceptual tree.

Second, given that the random relata represent the major source of confusion, it would be promising to tackle the problem as a two-step classification (as it was done for the CogALex task): a first step to discriminate the related pairs from the random ones, and a second step to discriminate among the different semantic relations. Such an approach is

very likely to improve the classification performance.

VI. CONCLUSION

In this paper, we have introduced Evaluation-MAN 2.0, the first large dataset for the evaluation of the corpus-based identification of semantic relations in Mandarin Chinese. The resource will be publicly available for the development and testing of vector space models for the Chinese language. We have also illustrated the procedures for data collection and reliability checking, which were closely inspired by similar datasets for the English language [16], [17].

We also carried out the first evaluation of this task for Chinese. First, we proposed an unsupervised evaluation, simply based on the cosine similarity between the word pairs and on the mean average precision metric. Our results showed that 1) the cosine similarity metric is quite efficient in discriminating between related and non-related word pairs; 2) all our vector space models tend to assign higher similarity scores to synonyms.

Finally, we ran a supervised evaluation, using the concatenation of the word embeddings as features for different types of classifiers. Deep Learning models such as the Feedforward and the Convolutional Neural Network proved to be the best performing ones in the relation identification task while hypernymy revealed itself as the easiest relation to identify. However, it is better to be cautious in this regard, since previous works dedicated to hypernym detection showed how such results can be inflated by lexical memorization effects [32], [48].⁶

The evaluations proposed in the present contribution are simply aimed at being a proof of concept, but several distributional models and many other classification algorithms could be tested on this new benchmark. Future research directions could explore, for example, the efficiency of the combination of distributional and pattern-based methods on Chinese data [33], or even to adapt the recently-introduced contextualized embeddings to the relation task [49], [50].

REFERENCES

- [1] R. Chaffin and D. J. Herrmann, "The similarity and diversity of semantic relations," *Memory Cognit.*, vol. 12, no. 2, pp. 134–141, Mar. 1984.
- [2] Z. S. Harris, "Distributional structure," *Word*, vol. 10, nos. 2–3, pp. 146–162, 1954.
- [3] K. J. Chen, C. R. Huang, L. P. Chang, and H. L. Hsu, "Sinica corpus: Design methodology for balanced corpora," in *Proc. 11th Pacific Asia Conf. Lang., Inf. Comput.*, Dec. 1996, pp. 167–176.
- [4] W. Y. Ma and C. R. Huang, "Uniform and effective tagging of a heterogeneous Giga-Word corpus," in *Proc. LREC*, Genoa, Italy, May 2006, pp. 24–28.
- [5] A. Lenci, "Distributional semantics in linguistic and cognitive research," *Italian J. Linguistics*, vol. 20, no. 1, pp. 1–31, Jan. 2008.
- [6] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Intell. Res.*, vol. 37, pp. 141–188, Feb. 2010.
- [7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: <https://arxiv.org/abs/1301.3781>
- [8] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, Doha, Qatar, Oct. 2014, pp. 1532–1543.
- [9] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proc. ACL*, Baltimore, MD, USA, Jun. 2014, pp. 238–247.
- [10] P. Mandera, E. Keuleers, and M. Brysbaert, "Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation," in *Journal Memory Language*, vol. 92. Amsterdam, The Netherlands: Elsevier, 2017, pp. 57–78.
- [11] C. R. Huang, R.-Y. Chang, and H.-P. Lee, "Sinica BOW (bilingual ontological wordnet): Integration of bilingual wordnet and SUMO," in *Ontology Lexicon: A Natural Language Processing Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2010, pp. 201–211.
- [12] A. Budanitsky, G., Hirst, "Evaluating Wordnet-based measures of lexical semantic relatedness," in *Computational Linguistics*, vol. 32. Cambridge, MA, USA: MIT Press, 2006, no. 1, pp. 13–47.
- [13] M. Zock, A. Lenci, and S. Evert, in *Proc. 5th Workshop Cogn. Aspects Lexicon (CogALex-V)*, Osaka, Japan, Dec. 2016, pp. III–V.
- [14] E. Santus, "Making sense: From word distribution to meaning," M.S. thesis, Dept. Chin. Bilingual Stud., Polytech. Univ., Hong Kong, 2016.
- [15] H. Liu, K. Neergaard, E. Santus, and C.-R. Huang, "EVALution-MAN: A chinese dataset for the training and evaluation of DSMs," in *Proc. LREC*, Portoroz, Slovenia, 2016, pp. 4583–4587.
- [16] E. Santus, F. Yung, A. Lenci, and C.-R. Huang, "EVALution 1.0: An evolving semantic dataset for training and evaluation of distributional semantic models," in *Proc. ACL Workshop Linked Data Linguistics*, Beijing, China, Jul. 2015, pp. 64–69.
- [17] M. Baroni and A. Lenci, "How we BLESSED distributional semantic evaluation," in *Proc. EMNLP Workshop GEometrical Models Natural Lang. Semantics*, Edinburgh, U.K., Jul. 2011, pp. 1–10.
- [18] J. Weeds, D. Clarke, J. Reffin, D. Weir, and B. Keller, "Learning to distinguish hypernyms and co-hyponyms," in *Proc. COLING*, Dublin, Ireland, 2014, pp. 2249–2259.
- [19] D. A. Cruse, *Lexical Semantics*. Cambridge, U.K.: Cambridge Univ. Press, 1986.
- [20] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.
- [21] C. S. G. Khoo and J. C. Na, "Semantic relations in information science," *Inf. Sci. Technol.*, vol. 40, no. 1, pp. 157–228, 2006.
- [22] E. Santus, Q. Lu, A. Lenci, and C.-R. Huang, "Taking antonymy mask off in Vector Space," in *Proc. PACLIC*, Phuket, Thailand, Dec. 2014, pp. 135–144.
- [23] E. Santus, A. Lenci, Q. Lu, and S. S. I. Walde, "Chasing hypernyms in Vector Spaces with entropy," in *Proc. EACL*, Gothenburg, Sweden, 2014, pp. 38–42.
- [24] V. Shwartz, E. Santus, and D. Schlechtweg, "Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection," in *Proc. EACL*, Valencia, Spain, 2017, pp. 65–75.
- [25] M. Hearst and C. Fellbaum, "Automated discovery of wordnet relations," in *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998, pp. 131–153.
- [26] R. Girju, A. Badulescu, and D. Moldovan, "Automated discovery of part-whole relations," *Comput. Linguistics*, vol. 32, no. 1, pp. 83–135, Jun. 2006.
- [27] M. Zock and D. Tesfaye, "Automatic creation of a semantic network encoding part_of relations," *J. Cognit. Sci.*, vol. 16, no. 4, pp. 431–491, Jan. 2015.
- [28] A. Lenci and G. Benotto, "Identifying hypernyms in distributional semantic spaces," in *Proc. Starssem*, Montreal, QC, Canada, 2012, pp. 75–79.
- [29] E. Santus, T.-S. Chiu, Q. Lu, A. Lenci, and C.-R. Huang, "What a nerd! beating students and vector cosine in the ESL and TOEFL datasets," in *Proc. LREC*, Portoroz, Slovenia, Mar. 2016, pp. 1–8.
- [30] M. Baroni, R. Bernardi, N.-Q. Do, and C.-C. Shan, "Entailment above the word level in distributional semantics," in *Proc. EACL*, Avignon, France, Apr. 2012, pp. 23–32.
- [31] S. Roller, K. Erk, and G. Boleda, "Inclusive yet selective: Supervised distributional hypernymy detection," in *Proc. COLING*, Dublin, Ireland, 2014, pp. 1025–1036.
- [32] E. Santus, A. Lenci, T. S. Chiu, Q. Lu, and C. R. Huang, "Nine features in a random forest to learn taxonomical semantic relations," in *Proc. LREC*, Portoroz, Slovenia, 2016, pp. 4557–4564.

⁶This basically means that the classifier learns that a given word, typically a generic term such as animal, is a typical case of hypernym, independently from its semantic relation with the other word [48].

- [33] V. Shwartz, Y. Goldberg, and I. Dagan, "Improving hypernymy detection with an integrated path-based and distributional method," in *Proc. ACL*, Berlin, Germany, Aug. 2016, pp. 2389–2398.
- [34] S. Roller and K. Erk, "Relations such as hypernymy: Identifying and exploiting Hearst patterns in distributional vectors for lexical entailment," in *Proc. EMNLP*, Austin, TX, USA, Nov. 2016, pp. 2163–2172.
- [35] K. A. Nguyen, S. Schulte im Walde, and N. T. Vu, "Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction," in *Proc. ACL*, Berlin, Germany, Aug. 2016, pp. 454–459.
- [36] K. A. Nguyen, M. Köper, S. S. I. Walde, and N. T. Vu, "Hierarchical embeddings for hypernymy detection and directionality," in *Proc. EMNLP*, Austin, TX, USA, Sep. 2017, pp. 233–243.
- [37] I. Vulic and N. Mrkšić, "Specialising word vectors for lexical entailment," in *Proc. NAACL*, New Orleans, LA, USA, Jun. 2018, pp. 1134–1145.
- [38] H. Liu and C. R. Huang, "EVALution-MAN 2.0: Expand the evaluation dataset for vector space models," in *Workshop on Chinese Lexical Semantics* (Lecture Notes in Computer Science), M. Dong, J. Lin, and X. Tang, Eds. Berlin, Germany: Springer, 2016, pp. 261–268.
- [39] C. R. Huang, S. K. Hsieh, J. F. Hong, Y. Z. Chen, I. L. Su, Y. X. Chen, and S. W. Huang, "Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing," *J. Chin. Inf. Process.*, vol. 24, no. 2, pp. 14–23, 2010.
- [40] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.
- [41] C. R. Huang, *Tagged Chinese Gigaword Version 2.0, LDC2009T14*. Philadelphia, PA, USA: Linguistic Data Consortium, 2009.
- [42] L. Kotlerman, I. Dagan, I. Szpektor, and M. Zhitomirsky-Geffet, "Directional distributional similarity for lexical inference," *J. Natural Lang. Eng.*, vol. 16, no. 4, pp. 359–389, Oct. 2010.
- [43] E. Chersoni, G. Rambelli, and E. Santus, "CogALex-V shared task: ROOT18," in *Proc. COLING Workshop Cognit. Aspects Lexicon*, Osaka, Japan, 2016, pp. 98–103.
- [44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
- [45] F. Chollet. (2015). *Keras GitHub*. [Online]. Available: <https://github.com/fchollet/keras>
- [46] C. R. Huang and S. K. Hsieh, "Chinese lexical semantics," in *The Oxford Handbook of Chinese Linguistics*. Oxford, U.K.: Oxford Univ. Press, 2015 pp. 290–305.
- [47] R. Sproat, *A Computational Theory of Writing Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
- [48] O. Levy, S. Remus, C. Biemann, and I. Dagan, "Do supervised distributional methods really learn lexical inference relations," in *Proc. NAACL*, Denver, CO, USA, 2015, pp. 970–976.
- [49] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. NAACL*, New Orleans, LA, USA, 2018, pp. 2227–2237.
- [50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, Minneapolis, MN, USA, 2019, pp. 4171–4186.



EMMANUELE CHERSONI received the B.A. and M.A. degrees from the University of Pisa, Italy, and the Ph.D. degree from Aix-Marseille University, France. He is currently a Postdoctoral Researcher with the Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University. His main research interests include distributional semantic models, thematic fit modeling, and automatic discovery of semantic relations and sentence processing.



NATALIA KLYUEVA received the Ph.D. degree from Charles University in Prague, where she was working mainly on machine translation for related Slavic languages. She is currently a Postdoctoral Researcher with the Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University. Her research concerns data analysis for various NLP tasks, such as sentiment analysis, biomedical data processing, and related languages.



ENRICO SANTUS is currently a Postdoctoral Researcher with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology. His current research interest includes information extraction. His previous research has investigated lexical distributional semantics, with attention to similarity and relation identification, and thematic fit estimation. He has been co-chair in ESSLLI 2016 and *SEM 2018, and he organized CMCL 2019 and numerous

shared tasks, including SemEval-2018 Task nine on Hypernymy Discovery and the CogALex-V Shared Task on the Corpus-Based Identification of Semantic Relations.



CHU-REN HUANG is currently the Chair Professor with the Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, and a Peking University Guest Professor with the Institute of Computational Linguistics. His recent books include *A Reference Grammar of Chinese* (Cambridge University Press, 2016), and *The Routledge Handbook of Chinese Applied Linguistics* (Cambridge University Press, 2019). He is the Editor of *Studies in Natural*

Language Processing (Cambridge University Press, 2010) and *Frontiers in Chinese Linguistics* (PKU Press/Springer).



HONGCHAO LIU was born in Binzhou, Shandong, China, in 1987. He received the B.A. degree in Chinese linguistics from the Ludong University, the two M.A. degrees in Chinese syntax from the Peking University and National University of Singapore, respectively, and the Ph.D. degree in computational linguistics from The Hong Kong Polytechnic University. Since 2018, he has been an Assistant Professor with the School of Arts, Shandong University. His research interests include

computational linguistics, computational semantics, lexical semantics, and Chinese syntax.

...