Augmenting Chinese WordNet semantic relations with contextualized embeddings

Yu-Hsiang Tseng
Graduate Institute of Linguistics
National Taiwan University
seantyh@gmail.com

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

Constructing semantic relations in WordNet has been a labour-intensive task, especially in a dynamic and fast-changing language environment. Combined with recent advancements of contextualized embeddings, this paper proposes the concept of morphology-guided sense vectors, which can be used to semi-automatically augment semantic relations in Chinese Wordnet (CWN). This paper (1) built sense vectors with pre-trained contextualized embedding models; (2) demonstrated the sense vectors computed were consistent with the sense distinctions made in CWN; and (3) predicted the potential semantically-related sense pairs with high accuracy by sense vectors model.

1 Introduction

Chinese Wordnet(CWN) (Huang et al., 2010) has been one of the most important lexical resources in Chinese. Through years of rigorous works from linguists and lexicographers, CWN covers large amount of Chinese words, senses distinctions, and various lexical semantic relations. However, the linguistic knowledge CWN tries to incorporate is far more than a static snapshot of the language usage from a given time. As a lexical resource which aims to facilitate better NLP applications, the current version of CWN has intended to incorporate the complicated and dynamic relations that language implicitly encodes. This is a challenging task for resource maintainer, for they have to manually edit the database, in order to keep up the the neologisms and ever-changing novel word usage.

Recent algorithmic advancements shed lights on how we can augment lexical resources, at least semi-automatically. Thanks to the bloom of internet and social media, voluminous textual data are easily available, where emergent concepts and their relations could be discovered from the real-world and most updated data. This process is further facilitated by recent development of deep learning and machine learning models, such as pre-trained language model (Howard and Ruder, 2018), word embeddings (Joulin et al., 2017), or contextualized embeddings. These computation resources allows us to leverage the ample data, without going through considerable efforts to actually collect, and store the vast amount of data, and setup a model training infrastructure.

In this paper, we took advantage the recent development on contextualized embeddings. Specifically, we used a pre-trained bidirectional encoder representations from transformer (BERT) (Devlin et al., 2018), basing on which we semi-automatically predicted new related senses in CWN. The predictions were only possible with the constraints encoded in Chinese morphology, where the semantic relationship between the whole word and its composing sub-word were suggested (Hsieh and Chang, 2014). We introduced how we applied BERT to construct sense vectors from existing example sentences in each CWN senses, and how to use sense vectors and heuristics rule s regarding Chinese word morphology to semi-automatically generate new relationships (hyponymy/troponymy pairs) among CWN senses. We evaluated these sense vectors with a simulation study and conducted an experiment on the model-predicted sense relation pairs. The procedures described in this paper was shown in Figure 1.

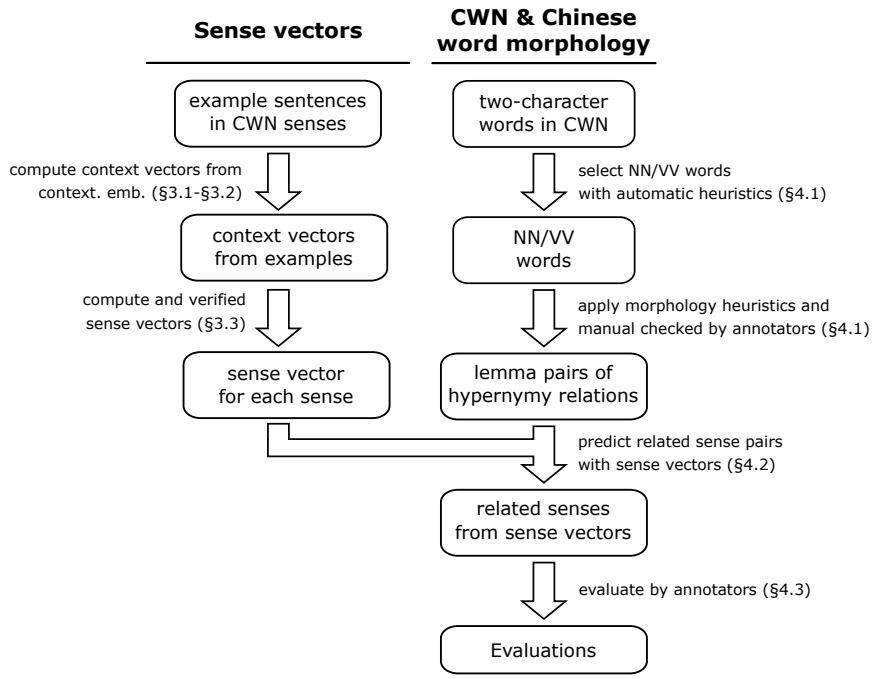

Figure 1: An workflow overview of predicting related senses with sense vectors and Chinese word morphology.

2 Related Works

2.1 Chinese morphology

The concept of word seems to be robust in many language, but remains elusive to languages such as Chinese (Hoosain, 1992). Chinese words were written as a series of Chinese characters, and there is no orthographic cues (such as spaces in English) delineating word boundaries. Therefore, words are instead defined by different theories, focusing on different linguistic aspects, such as their morphological, syntactical or semantic properties. In CWN, words were defined as characters with independent meaning and play a specific syntactic role (CKIP, 1996), and 7 guidelines were introduced to ensure a consistent and meaningful criterion of words.

Most Chinese words are composed of two characters. Characters are the writing units in Chinese, each are written within a square block. Arguably treated as morpheme as its linguistic property by definition, some characters can be used alone, some characters need to combine other characters to form a word, and most of them bring their original meanings into the composition process. For example, 泉水（quán shuǐ, 'spring'）is a word composing of two characters. The second character 水 (shuǐ) can itself be used independently to indicate the meaning of 'water.' For words like 語言 (yǔ yán, 'language'), though the second character 言 (yán) cannot be used independently in contemporary Mandarin Chinese, it still nonetheless contributes an *etymological* meaning of 'speech, speak.'

Unlike inflectional languages, Chinese words do not undergo morphological alternations, such as eat, eats, eaten, eating or eater in English. There are only a few affix-like morphemes in Chinese that account for a small portion of Chinese words. For instance, 師 (-shī) can be attached to a noun as a suffix, indicating a profession, such as 工程 (gōng chéng) means engineering, and attaching the suffix 工程師 (gōng chéng shī) means engineers. However, Chinese do have intricate word morphology, which reflects knowledge about the structure and use of words. For example, 直升機 (zhí shēng jī) means helicopter, and the three characters of which the word are composed literally means vertically(直, zhí) arised(升, shēng) machine(機, jī). Likewise, 汽車 (qì chē) means automobile, the two composing characters could be loosely translated as "gas(汽, qì)-

car(車, chē)." The fact that meanings of word and its composing characters match suggests that Chinese words, through their morphology, reflect systematic knowledge that a native speaker have toward the world. (Packard, 2000)

In order to leverage the copious knowledge encodes within Chinese morphology, previous studies devised heuristic rules to decode the semantic relationships between word and their composing characters (Hsieh and Chang, 2014). The relations decoded provided useful hints for semantic relations, that can be used to expand semantic relations in CWN. Specifically, for some (two-character) words following a *modifier-head* structure, the second component (serving as the head) is the hypernym of the whole word. For example, 書店 (shū diàn) means 'book store', the second component 店 (diàn, 'shop, store') is then inferred to be the hypernym of the whole word (書店). The heuristic rule in application is very effective, for it provide a clear guidance of possible hypernym relations a concept could link to. However, these rules only apply on the lemma level. That is, after the potential hypernyms were identified, the rule cannot provide further guidance on the senses upon which the hypernymy relation should be created.

2.2 Contextualized Embeddings

Vector semantics are models in which researcher use a formal mathematical structure (i.e. vectors) to represent how lexical meanings of words reside in a vector space. The vectors representing each words also encode, to some extent, their mutual semantic relations in that space. This general approach, while being a heated topic in recent years (Landauer and Dumais, 1997; Griffiths et al., 2007; Mikolov et al., 2013; Peters et al., 2018), could be traced back to mid-20th century (Firth, 1957). The idea was to explore the co-occurrence of the words in context (sentences, or a groups of preceding and following words), and use the context to determine the *location* of a word vector in semantic space, where thus location could best reflect the relationships with other words.

While models of vector semantics enjoyed great successes in various NLP tasks, even were indispensable constructs in virtually all deep learning models, challenges emerged when they came to WordNet. WordNet, as a lexical resource of word senses and linguistic knowledge, make intricate distinctions on word senses and the synsets among them. However, vector semantics models had a major limitation of meaning conflation deficiency (Camacho-Collados and Pilehvar, 2018), namely they conflate multiple meanings of a word (lemma) into one representation. For example, in word2vec model (Mikolov et al., 2013), vectors of target word were constructed through the task of predicting the target word with surrounding word vectors (continuous bag of words, CBOW), or, conversely, predicting surrounding words with the target word vectors (skip gram). Different word contexts were independent samples in training, they are not explicitly used by the model. The resulting word vectors were therefore undifferentiated representations of word senses.

Other models have the potential to accommodate, or even represent, word senses information, but not without caveats. For example, latent Dirichlet allocation (LDA) (Griffiths et al., 2007), representing meanings of each word as a probability distribution over different topics, could describe each word sense as a mixture of different topic components. But the problems remains on how to relate latent topics with the word senses. Other endeavors relies on a sense-disambiguated corpus (Iacobacci et al., 2015), and inferred the sense vectors through the disambiguated context. But this approach required a mature word sense disambiguation (WSD) algorithm or sense-tagged corpus with given sets of word sense distinctions. Chinese WSD is an active and productive research topic, but the word sense disambiguation on CWN word senses remains a challenging task.

Instead of relying on sense-disambiguated corpus, recent models tried to incorporate word context into deep learning models and construct contextualized vectors (Peters et al., 2018; McCann et al., 2017). Inspired by the deep learning models in computer vision, these models represent word contexts as an abstract information built upon the basic word embeddings in a language modeling task. Specifically, a model was trained to predict the

next word in a sentence based on the words previously seen. The models used word vectors as input, but the embeddings layers (i.e. word vectors) stacked upon were deep layers tried to encode the contextual information. The outputs of these deep layers were used to complete the prediction task in training; and additionally, they represented the context vectors the words occurred in. Recent deep learning researches provided multiple choices of such layers, like bidirectional LSTM used in ELMo (Peters et al., 2018) and decoder transformer used in OpenAI transformer (Radford et al., 2018). These models, instead of treating each word as a static vector, could generate a contextualized vector for each word in any given contexts. However, as these models were trained on language modeling tasks, only either preceding or succeeding word contexts were exploited to build context vectors.

Bidirectional-encoder representation (BERT) (Devlin et al., 2018) employed different task to train models making use of surrounding word contexts to generate context vectors. As other contextualized vector models, BERT also uses word vector as its input, but the deep layers stacked upon them were layers of encoder transformers (Vaswani et al., 2017). In order to allow encoder to consider the surrounding word contexts without peeking into the predicting targets in the same time, BERT used a cloze task in its training stage. In the cloze task, each word in the whole sentence was available to model, with only the clozed word (the target) masked out. The model then learned to construct a context vector with the surrounding words, and predicts the clozed word with the context vectors. The contextualized vectors trained on this model had wide range of applicability. It had been shown that without substantial modification, the model achieved superior performance on NLP tasks, such as question answering and language inferences.

This paper aims to investigate whether the model of contextualized embeddings can help researchers to identify the semantic relations between word senses defined in CWN or not. The goals of present paper are as follows: (1) Examine how the sense vectors computed by contextualized vector models (i.e. BERT) differentiated the word sense distinctions made in CWN. (2) Predict possible hypernymy-hyponymy relations among sense pairs from sense vectors, guided by Chinese morphology. (3) Evaluate the predictions made by the model with human annotations.

3 Building sense vectors

Word sense is closely related to the context the word resides in, and the contextualize embeddings is meant to encode the context. If we can characterized the context through contextualized embeddings, the context vector was then a formal representation of a word sense. We therefore computed sense vectors from the contextualized embeddings of the target word located in an disambiguated context.

In this section, we first identified the lemmas (and their senses) to be included in current analysis and the experiment in following section. Secondly, we built sense vectors from example sentences of each sense. Thirdly, in order to explore the nature of the sense vectors, we conducted a simulation study over the computed sense vectors.

3.1 Extracting example sentences

We first selected 1,815 lemmas from CWN. [1] These lemmas satisfied following criterion: (1) they are two-character lemmas; (2) each of the composing character is itself a lemma in CWN; (3) all senses of each lemma (both two-character lemma and one-character lemma) must have at least two example sentences. The complete lemmas hence included 2,897 lemmas, which were comprised of two-character lemmas, and their 1,082 unique composing characters as one-character lemma.

These lemmas were related to 11,521 senses (40.0% of all CWN senses) in CWN, and 37,976 example sentences were extracted from these sentences.

3.2 Computing sense vectors

We used BERT (pre-trained on Chinese Wikipedia data dump) as the model of contextualized embeddings. The model had 12

[1] Note that *homonyms* are treated as separate words in CWN, e.g., 打 ('punch' and other derived senses) and 打 ('dozen') are the same lemma used as two words. In this experiment, homonyms are considered as different word senses.

layers, each having 768 hidden states. In this analysis, we concatenated the hidden states of the last 4 layers as the contextualized embeddings. The resulting contextualized embedding dimensions ($\text{CE}_{\text{dimension}}$) was 3,072. The context vector of target lemma in the sentence was then selected from the contextualized embeddings obtained from BERT model. The context vector of example j of sense i, denoted by s_{ij}, can be written as:

$$s_{ij} = \underbrace{\mathbf{1}_{\text{target}}}_{1\times T} \underbrace{\mathbf{CEs}([w_{ij}^{(1)}, \ldots, w_{ij}^{(t)}, \ldots, w_{ij}^{(T)}])}_{T\times \text{CE}_{\text{dimension}}} \tag{1}$$

where T denoted the number of tokens in the example sentences, $w_{ij}^{(t)}$ was the t^{th} token in the example sentences, and $\mathbf{1}_{\text{target}}$ is a vector with each of its element an indicator function:

$$\mathbf{1}_{\text{target}} = \begin{cases} 1, & w_{ij}^{(t)} \text{ is the target lemma} \\ 0, & \text{otherwise} \end{cases}$$

The sense vector, μ_j, of sense j for a given lemma μ, was computed as the centroid of context vectors in all $n_e^{(j)}$ example sentences:

$$\mu_j = \sum_{i}^{n_e^{(j)}} s_{ij}/n_e^{(j)} \tag{2}$$

The sense vectors were computed for respective senses in selected CWN lemmas. However, these sense vectors were only a linear combinations of the context vectors, which were generated by an intricate deep learning model. The possibility exists that the context BERT trying to represent might be an abstract concept independent from the word context referred in language usage. In order to further investigate the nature of these sense vectors, we carried out following simulation study.

3.3 Sense vector simulation

The purpose of the simulation study was to verify the sense vectors came from groups respecting sense distinctions made in CWN. We compared the grouping patterns of sense vectors and two others from simulated conditions, to demonstrate the sense vectors reflected different contexts of word senses, instead of coming from random patterns.

We first devised a statistic to quantify how clear-cut the groups context vectors formed into, where the sense vectors were computed from. For a given lemma μ, to describe how well the context vectors, s_{ij} were "grouped together" within different senses, we calculated two scores, $\text{MS}_k^{(\text{senses})}$ and $\text{MS}_k^{(\text{error})}$, based on the euclidean distance between s_{ij}, their sense vector μ_j, and the centroid of all sense vector, $\overline{s_{..}}$:

$$\text{MS}_k^{(\text{senses})} = \frac{\sum_j n_j \|\mu_j - \overline{s_{..}}\|^2}{m-1} \tag{3}$$

$$\text{MS}_k^{(\text{error})} = \frac{\sum_i \sum_j \|s_{ij} - \mu_j\|^2}{N_k - m} \tag{4}$$

The ratio of these two scores measured the extent to which the sense vectors distanced from each other, by comparing with the sense vectors distanced from their respective context vectors. This ratio, ζ_k, was computed as:

$$\zeta_k = \frac{\text{MS}_k^{(\text{senses})}}{\text{MS}_k^{(\text{error})}}$$

Intuitively, a small ζ_k indicated the sense vectors themselves were not clearly grouped, since the distance between the sense vectors was similar with the distance between the context vectors used to calculate the sense vectors. This ratio was closely related to F statistic, which was often in comparing two sample variances. However, two caveats existed kept us from directly proceeding to hypothesis testing with F statistic. (1). The explicit distribution of sense vectors as a random variable was not readily available, it is unclear if ζ_k still followed F-distribution under null hypothesis. (2). The simulation was to compare all lemmas in CWN. That is, each lemma was itself a sample in the simulation. However,each lemma has different number of sense vectors and number of examples, a normalized index was then needed to describe ζ_k from different lemmas.

To normalize ζ_k from different lemmas with different senses and examples, we defined π_k, which was the area under the right-tail of ζ_k

in the probability density function of F distribution.

$$\pi_k = 1 - \int_0^{\zeta_k} F_{pdf}(x; df_1, df_2)dx \quad (5)$$

$$df_1 = m - 1 \quad (6)$$

$$df_2 = N_k - m \quad (7)$$

$$(8)$$

where F_{pdf} denoted the probability density function of a given F distribution, N_k denoted total number of examples in lemma$_k$.

Since ζ_k may not follow F distribution, the value of π_k was just a score indicating the "well-groupness" of the senses in lemma k. Smaller π_k signified more clear-cut grouping. The resulting π_k from actual sense vectors had mean of 0.14, standard deviation of 0.10 (Figure 2).

In order to better interpreted the π_k from actual sense vectors, we compare the π_k with two other simulated conditions: (1) random Gaussian vectors and (2) permuted vectors. The first simulated condition was to replace all context vectors with random standard Gaussian vectors of the same length. This condition provided a random baseline of how π_k distributed if context vectors were random noises. The second simulated conditions permuted the actual context vectors. The context vectors were randomly shuffled, and randomly assigned to each word senses, while the sense number and the number of examples of each sense remained the same. The underlying rationale was if the context vectors from the same sense were closer together, then a permuted version of which would destroy the patterns.

Figure 2 showed the results of simulations and the sense vectors. Patterns of π_k in random condition ($M = 0.41$, $SD = 0.05$) was similar to those in permuted condition ($M = 0.42$, $SD = 0.09$). Importantly, the distribution of π_k of actual sense vectors were smaller than any of the simulated conditions. These patterns showed the computed sense vectors had clear grouping structures and the groupings were consistent with sense distinctions in CWN.

4 Experiment

With sense vectors as a computable representation of word senses, we aimed to semi-automatically discover potential hypernymy-hyponymy sense relations in CWN, guided by Chinese morphology. Previous study argued that Chinese two-character words with inner structure of two nouns and two verbs, were likely a hyponymy of the second character (when used as a one-character word). That is, at lemma level, we could discover semantic relations leveraging Chinese word morphology. However, semantic relations in WordNet are relationship among word senses. Given there are multiple word senses in each lemma, manually found them would be a daunting task. With help of sense vectors, we could try to find senses among which relations existed.

4.1 Selecting candidate lemma

To find out candidate hypernymy-hyponymy lemma pairs, we first used heuristic rules to automatically select words composing of two nouns (NN) or two verbs (VV). The heuristic rules were to determine the part-of-speech of composing character, basing on the dictionary data compiled by the Ministry of Education of Taiwan. Three criterion were applied consecutively: (1) excluding senses from classical Chinese, compare the number of senses a POS have, the POS with more sense count was the POS of the character; (2) if sense counts of different POS were equal, compare the frequency sum of the example words (as listed in sense entries) of that sense in a corpus; (3) if the frequency sum were equal, compare the sense counts of POS in CWN. These three criterion labeled 99% words in 1,815 two-character words (the same set of words in analyzing sense vectors). POS of the remaining words were assigned manually. There were respectively 824 and 362 words of NN and VV structures selected.

Three graduate students in Graduate Institute of Linguistics, National Taiwan University examined these N_1N_2 and V_1V_2 words, labeling words (W) with hyponymy relations (W is a kind of N_2) or troponymy relations (V is a way of doing V_2). Since determining the relations were relatively straightforward given the words and composing character , each item

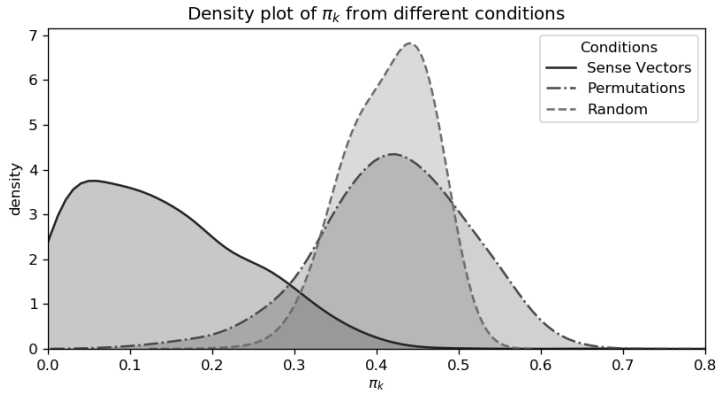

Figure 2: Distribution of sense vectors statistics, π_k.

was only annotated by one annotator. The resulting word list comprised 337 NN words and 150 VV words.

4.2 Predicting related senses

We used sense vectors computed in previous section to predict which sense were related in the lemma pair (i.e. the whole word lemma and the N_2/V_2 lemma). Given a pair of lemmas, μ_j was the sense vector computed of lemma μ and ν_j were of lemma ν. We predicted the related senses as the nearest sense vectors between two set of lemma senses. The distance measure, $d_{i,j}$, was the euclidean distance between the sense vectors:

$$d_{i,j} = \|\mu_i - \nu_j\|^2$$

All distances between the sense pairs in lemma μ and lemma ν formed a distance matrix $\mathbb{D}$:

$$\mathbb{D}_{(\mu,\nu)} = \begin{array}{c} \\ \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_m \end{array} \begin{array}{c} \begin{array}{cccc} \nu_1 & \nu_2 & \cdots & \nu_n \end{array} \\ \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ d_{3,1} & d_{3,2} & \cdots & d_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m,1} & d_{m,2} & \cdots & d_{m,n} \end{pmatrix} \end{array}$$

The predicted sense pairs were the senses pairs of smallest d_{ij}:

$$\text{Related sense pair } (\mu_i, \nu_j) = \underset{i,j}{\operatorname{argmin}} \left\{ d_{i,j} \,\middle|\, d_{i,j} \in \mathbb{D}_{(\mu,\nu)} \right\}$$

The calculations were performed on all 487 lemma pairs. Two of the lemmas had format errors in the example sentences, and had no sense vectors. Therefore 485 sense pairs predictions were made.

Word Structure	NN n=337	VV n=148	Overall n=485
Baseline			
Random	0.12	0.21	0.15
First Sense	0.40	0.46	0.42
Model Prediction			
Top 1	0.81	0.83	0.82
First 5	0.96	0.94	0.96
First 10	0.99	0.97	0.98

Table 1: Accuracy of related sense pairs predicted by model and baseline performance.

4.3 Evaluation

Model-predicted related sense pairs were equally divided into three parts and each part was evaluated by an annotators. Annotators marked whether the predicted sense pairs were actually hyponymy/troponymy pairs. If they found erroneous predictions, correct sense pairs would be added, and these data were further used in evaluation. The results were shown in Table 1.

The overall accuracy of model predictions was 0.82, with similar performance on either NN or VV constructions. To better illustrate the nature of the predicting task, two baseline performance were provided: (1) a random baseline was the performance the model was random guessing; (2) 'first sense strategy' was the model always picked the first sense listed in CWN. Compared with the accuracy of

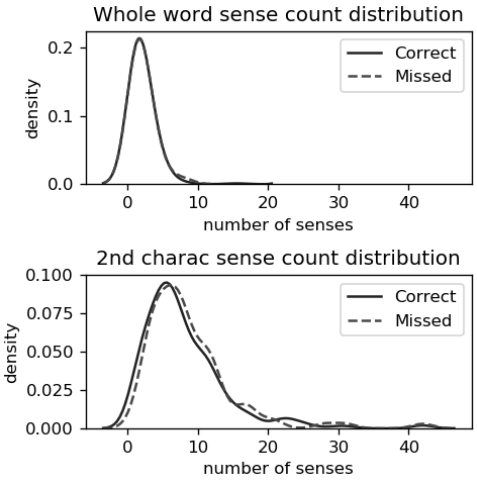

Figure 3: Sense counts on correct and missed-linked senses

these baselines, present model provides valuable suggestions on potential sense pairs.

Table 1 also shows the prediction ranks of the correct sense pairs. That is, if the correct sense pairs were not the nearest one in the distance matrix, would the correct pairs rank in first 5 or 10 pairs in the distance matrix. The results indicated there were 96% of correct pairs were ranked within the first 5 pairs.

To further investigate the errors made by the model, Figure 3 shows the sense counts distribution of the whole word and the second composing character (N_2 / V_2), on correct and missed predictions. From Figure 3, the distribution of the second character when missed predicted, was marginally more than the correct ones; while the distribution was virtually the same in whole word. The latter pattern was expected since the Chinese two-character words generally had fewer word senses.

This experiment demonstrated how to leverage Chinese word morphology and sense vectors to discover potential hypernymy or troponymy relations in CWN. The evaluation also showed this semi-automatically procedure suggest valuable sense pairs.

5 Conclusion

This paper combines recent advancements of contextualized embeddings and existing CWN resources to build sense vectors. We have demonstrated these sense vectors followed the sense distinctions made in CWN, and showed sense vectors, guided by Chinese morphology, provided valuable suggestion discovering hypernymy/troponymy. The semi-automatic procedures greatly facilitate the on-going development of CWN in the fast-paced language environment.

Acknowledgements

This work was supported by Ministry of Science and Technology (MOST), Taiawan. Grant Number MOST 108-2634-F-001-006. We thank Yong-Fu Chao, Ying-Yu Chen, Chiung-Yu Chiang, and Yi-Ju Lin (National Taiwan University) for their assistance on data preprocessing and annotations.

References

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

CKIP. 1996. *Study on Chinese word boundaries and computational standard in segmentation.* CKIP Technical Reports. Institute of Information Science, Acaemia Sinica.

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

J. R. Firth. 1957. Applications of general linguistics. *Transactions of the Philological Society*, 56(1):1–14, November.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Rumjahn Hoosain. 1992. Psychological reality of the word in chinese. In *Language Processing in Chinese*, pages 111–130. Elsevier.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Shu-Kai Hsieh and Yu-Yun Chang. 2014. Leveraging morpho-semantics for the discovery of relations in chinese wordnet. In *Proceedings of the Seventh Global Wordnet Conference*, pages 283–289, Tartu, Estonia.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and

Shen-Wei Huang. 2010. Constructing chinese wordnet: Design principles and implementation. (in chinese). *Zhong-Guo-Yu-Wen*, 24:2:169–186.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China, July. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *CoRR*, abs/1708.00107.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *NAACL-HLT*, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.