Lexical Sememe Prediction via Word Embeddings and Matrix Factorization

Ruobing Xie[1*], **Xingchi Yuan**[1*], **Zhiyuan Liu**[1,2†], **Maosong Sun**[1,2]
[1] Department of Computer Science and Technology,
State Key Lab on Intelligent Technology and Systems,
National Lab for Information Science and Technology, Tsinghua University, Beijing, China
[2] Jiangsu Collaborative Innovation Center for Language Ability,
Jiangsu Normal University, Xuzhou 221009 China

Abstract

Sememes are defined as the minimum semantic units of human languages. People have manually annotated lexical sememes for words and form linguistic knowledge bases. However, manual construction is time-consuming and labor-intensive, with significant annotation inconsistency and noise. In this paper, we for the first time explore to automatically predict lexical sememes based on semantic meanings of words encoded by word embeddings. Moreover, we apply matrix factorization to learn semantic relations between sememes and words. In experiments, we take a real-world sememe knowledge base HowNet for training and evaluation, and the results reveal the effectiveness of our method for lexical sememe prediction. Our method will be of great use for annotation verification of existing noisy sememe knowledge bases and annotation suggestion of new words and phrases.

1 Introduction

Words are distinct meaningful elements of speech or writing in human languages, but not the indivisible semantic units. Linguists define sememes as the minimum units of semantic meanings of human languages [Bloomfield, 1926], and the semantic meanings of concepts (e.g., senses, words and phrases) can be semantically composed by a limited close set of sememes. And the idea of sememes is closely related to the idea of language universals [Goddard and Wierzbicka, 1994].

Since the lexical sememes of words are not explicit in languages, people build linguistic knowledge bases (KBs) by annotating words with a pre-defined set of sememes. One of the most well-known sememe KBs is HowNet [Dong and Dong, 2006]. In HowNet, experts construct an ontology of $2,000$ sememes, and manually annotate more than 100 thousand words and phrases in Chinese with sememes in hierarchical structures, which will be introduced in details in Section 3.1. HowNet has been widely used in various NLP tasks such as word similarity computation [Liu and Li, 2002], word sense disambiguation [Duan *et al.*, 2007], and sentiment analysis [Huang *et al.*, 2014].

The manual construction is actually time-consuming and labor-intensive, e.g., HowNet is built for more than 10 years by several linguistic experts. As the development of communications and techniques, new words and phrases are emerging and the semantic meanings of existing words are also dynamically evolving. In this case, sustained manual annotation and update are becoming much more overwhelmed. Moreover, due to the high complexity of sememe ontology and word meanings, it is also challenging to maintain annotation consistency among experts when they collaboratively annotate lexical sememes.

To address the issues of inflexibility and inconsistency of manual annotation, we propose to automatically predict lexical sememes for words, which is expected to assist expert annotation and reduce manual workload. For simplicity, we do not consider the complicated hierarchies of word sememes, and simply group all annotated sememes of each word as the sememe set for learning and prediction.

The basic idea of sememe prediction is that those words of similar semantic meanings may share overlapped sememes. Hence, the key challenge of sememe prediction is how to represent semantic meanings of words and sememes to model the semantic relatedness between them. In this paper, we propose to model the semantics of words and sememes using distributed representation learning [Hinton, 1986]. Distributed representation learning aims to encode objects into a low-dimensional semantic space, which has shown its impressive capability of modeling semantics of human languages, e.g., word embeddings [Mikolov *et al.*, 2013] have been widely studied and utilized in various tasks of NLP.

As shown in previous works [Mikolov *et al.*, 2013], it is effective to measure word similarities using cosine similarity or Euclidean distance of their word embeddings learned from large-scale text corpus. Hence, a straightforward method for sememe prediction is, given an unlabeled word, we find its most related words in HowNet according to their word embeddings, and recommend the annotated sememes of these related words to the given word. The method is intrinsically similar to collaborative filtering [Sarwar *et al.*, 2001] in recommender systems, capable of capturing semantic relatedness between words and sememes based on their annotation co-occurrences.

* indicates equal contribution
† Corresponding author: Z. Liu (liuzy@tsinghua.edu.cn)

Word embeddings can also be learned with techniques of matrix factorization [Levy and Goldberg, 2014]. Inspired by successful practice of matrix factorization for personalized recommendation [Koren *et al.*, 2009], we propose to factorize word-sememe matrix from HowNet and obtain sememe embeddings. In this way, we can directly measure the relatedness of words and sememes using dot products of their embeddings, according to which we could recommend the most related sememes to an unlabelled word.

The two methods are named as Sememe Prediction with Word Embeddings (SPWE) and with Sememe Embeddings (SPSE/SPASE) respectively. We also explore the ensemble of both sememe prediction methods. In experiment, we evaluate our models on sememe prediction in HowNet compared with baselines, and experiment results show our method can effectively identify related sememes for unlabelled words. We extensively investigate the characteristics of various methods and analyze typical errors, which is expected be useful for further improving prediction performance. We demonstrate the main contributions of this work as follows:

- This work is the first attempt to automatically predict sememes for words, and we propose two methods for solving sememe prediction.
- We evaluate our models on a real-world dataset HowNet and achieve promising results. We also conduct detailed exploration for deep understanding of the relationships between words and sememes.

2 Related Work

Many works have been done to automatically extract knowledge to build knowledge bases. For example, knowledge graphs, such as Freebase [Bollacker *et al.*, 2008], DBpedia [Auer *et al.*, 2007] and YAGO [Hoffart *et al.*, 2013], rely the task of relation extraction to identify relational facts between entities from plain text [Mintz *et al.*, 2009]. Typical linguistic KBs, such as WordNet [Miller, 1995] and BabelNet [Navigli and Ponzetto, 2012], usually have to identify those words of similar meanings to build thesaurus [Nastase *et al.*, 2013]. Apart from other linguistic KBs, sememe KBs like HowNet [Dong and Dong, 2006] are built following a philosophy of reductionism, emphasizing the *parts* and *attributes* of words represented by sememes. Sememe KBs are significant for understanding the nature of semantics in human languages, which have been widely used in various NLP tasks like information structure annotation [Gan and Wong, 2000] and word sense disambiguation [Gan *et al.*, 2002].

To the best of our knowledge, automatic sememe prediction has not been addressed by previous works. As aforementioned, the task is similar to personalized recommendation, which has been extensively studied for years [Bobadilla *et al.*, 2013]. Our proposed methods in this paper are partially inspired by two representative methods in recommendation system, namely collaborative filtering [Sarwar *et al.*, 2001] and matrix factorization [Koren *et al.*, 2009]. The difference is that, in order to model semantic meanings of words, we learn word embeddings from large-scale text corpus, which are further fed to sememe prediction according to semantic relatedness between words and sememes. As will shown in experiments, our proposed methods are simple and effective. In future, we can explore more effective recommendation models such as Factorization Machines [Rendle, 2010] for sememe prediction.

3 Methodology

We propose our models for a novel task sememe prediction, which aims to recommend the most appropriate sememes for each unlabelled word. In this section, we first introduce how words and sememes are organized in HowNet. Next, we show the details of three sememe prediction models, namely Sememe Prediction with Word Embeddings (SPWE), with Sememe Embeddings (SPSE) and with Aggregated Sememe Embeddings (SPASE). Finally, we further improve the performance with ensemble strategy.

3.1 Sememes and Words in HowNet

First, we introduce how words, senses and sememes are organized in HowNet[1]. In HowNet, a word may have various senses, and each sense has several sememes describing the exact meaning of sense. We denote W and S as the overall set of words and sememes. For each word $w \in W$, we denote its sememe set in HowNet as $S_w = \{s_1, \ldots, s_{n_w}\}$. Fig. 1 shows the sememe annotations of the word *apple*. In HowNet, *apple* has two senses, namely *apple (brand)* and *apple (fruit)*. The former sense has several sememes including *computer*, *PatternValue*, *able*, *bring*, *speBrand (specific brand)* to describe the meaning of *apple (brand)*, and the latter one has the sememe *fruit*. In HowNet, there are about $2,000$ sememes to describe all words with different combinations.

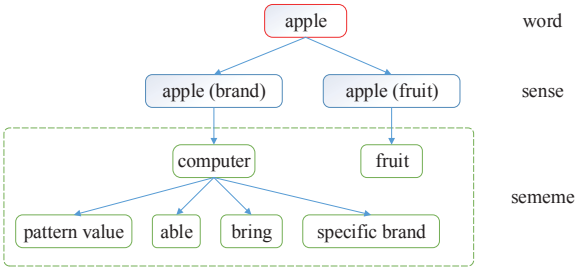

Figure 1: Word, sense and sememe in HowNet.

Fig. 2 gives another example to reveal the capability of sememes to describe semantic meanings of words. The word *antique shop* in Chinese has a list of sememes including *InstitutePlace*, *commerce*, *buy*, *treasure*, *past*, *sell*, which is exactly equivalent to the definition of antique shop that "antique shop is a commercial place to sell and buy past treasure". Therefore, it is intuitive that sememes could flexibly and precisely represent the essential meanings of words.

In this paper, we focus on modeling semantic relatedness between words and sememes for sememe prediction. For simplicity, we ignore the hierarchical structure of sememes, and group all sememes of a word in various senses together to form the corresponding sememe set of each word.

[1] http://www.keenage.com/

古董店(antique store) :
DEF={InstitutePlace|场所 : domain = {commerce|商业},
{buy|买 : location={~}, possession = {treasure|珍宝 :
modifier = {past|过去}}}, {sell|卖 : agent={~}, possession =
{treasure|珍宝 : modifier = {past|过去}}}}

Figure 2: An example of detailed sememes for a word.

3.2 Sememe Prediction with Word Embeddings

Given an unlabelled word, it is straightforward to recommend sememes according to its most related words, since we assume that similar words should have similar sememes. This idea is similar to collaborative filtering in personalized recommendation, for in the scenario of sememe prediction words can be regarded as users, and sememes as the items/products to be recommended. Inspired by this, we propose Sememe Prediction with Word Embeddings (SPWE), using similarities of word embeddings to judge user distances.

Formally, we define the score function $P(s_j, w)$ of sememes s_j given a word w as:

$$P(s_j, w) = \sum_{w_i \in W} \cos(\mathbf{w}, \mathbf{w_i}) \cdot \mathbf{M}_{ij} \cdot c^{r_i}, \tag{1}$$

where $\cos(\mathbf{w}, \mathbf{w_i})$ is the cosine similarity between word embeddings of w and w_i pre-trained by GloVe. $\mathbf{M}_{ij}$ indicates the annotation of sememe s_j on word w_i, where $\mathbf{M}_{ij} = 1$ indicates the word w_i has the sememe s_j in HowNet and otherwise has not. Higher the score function $P(s_j, w)$ is, more possible the word w should be recommended with s_j.

Differing from classical collaborative filtering in recommendation systems, we should only concentrate on the most similar words when predicting sememes for new words, since irrelevant words have totally different sememes which may be noises for sememe prediction. To address this problem, we assign a declined confidence factor c^{r_i} for each word w_i, where r_i is the descend rank of word similarity $\cos(\mathbf{w}, \mathbf{w_i})$, and $c \in (0, 1)$ is a hyper-parameter. In this way, only a few top words that are similar to w has strong influences on predicting sememes.

SPWE only uses word embeddings for word similarities, and is simple and effective for sememe prediction. It is because that differing from noisy and incomplete user-item matrix in most recommender systems, HowNet is carefully annotated by human experts, and thus the word-sememe matrix is with high confidence. Therefore, we can confidently apply the word-sememe matrix to collaboratively recommend reliable sememes based on similar words.

3.3 Sememe Prediction with Sememe Embeddings

Sememe Prediction with Word Embeddings model follows the assumption that the sememes of a word can be predicted according to its related words' sememes. However, simply considering sememes as discrete labels may inevitably neglect the latent relations between sememes. To take the latent relations of sememes into consideration, we propose Sememe Prediction with Sememe Embeddings (SPSE), which projects both words and sememes into the same semantic vector space, learning sememe embeddings according to the co-occurrences of words and sememes in HowNet.

Inspired by GloVe [Pennington *et al.*, 2014] decomposing co-occurrence matrix of words to learn word embeddings, we propose to learn sememe embeddings by factorizing word-sememe matrix and sememe-sememe matrix simultaneously. These two matrices are both constructed from HowNet. As for word embeddings, similar to SPWE, we use word embeddings pre-trained from large-scale corpus and fix them during factorizing word-sememe matrix. With matrix factorization, we can encode both sememe and word embeddings into the same low-dimensional semantic space, and compute the cosine similarity between normalized embeddings of words and sememes for sememe prediction.

More specifically, from HowNet we can extract a word-sememe matrix $\mathbf{M}$ with $\mathbf{M}_{ij} = 1$ indicating word w_i is annotated with sememe s_j, otherwise $\mathbf{M}_{ij} = 0$. We can also extract a sememe-sememe matrix $\mathbf{C}$, where $\mathbf{C}_{jk}$ indicates the correlations between two sememes s_j and s_k, which is defined as point-wise mutual information that $\mathbf{C}_{jk} = \text{PMI}(s_j, s_k)$. Note that, by factorizing $\mathbf{C}$, we will obtain two distinct embeddings for each sememe s, denoted as $\mathbf{s}$ and $\bar{\mathbf{s}}$ respectively. The loss function of learning sememe embeddings is defined as follows:

$$\begin{aligned}\mathcal{L} = &\sum_{w_i \in W, s_j \in S} \left(\mathbf{w}_i \cdot (\mathbf{s}_j + \bar{\mathbf{s}}_j) + \mathbf{b}_i + \mathbf{b}'_j - \mathbf{M}_{ij}\right)^2 \\ &+ \lambda \sum_{s_j, s_k \in S} \left(\mathbf{s}_j \cdot \bar{\mathbf{s}}_k - \mathbf{C}_{jk}\right)^2,\end{aligned} \tag{2}$$

where $\mathbf{b}_i$ and $\mathbf{b}'_j$ denote the bias of w_i and s_j. These two parts correspond to the losses of factorizing matrices $\mathbf{M}$ and $\mathbf{C}$, adjusted by the hyper-parameter λ. Since the sememe embeddings are shared by the both factorizations, our SPSE model enables jointly encoding both words and sememes into a unified semantic space.

Since each word is typically annotated with 2 to 5 sememes in HowNet. Hence, most elements in the word-sememe matrix are zeros. If we treat all zero elements and non-zero elements equally during factorization, the performance will be much worse. To address this issue, we assign different factorization strategies for zero and non-zero elements. For each zero elements, we choose to factorize them with a small probability like 0.5%, and otherwise we choose to ignore. While for non-zero elements, we always choose to factorize them. With the help of this strategy, we can pay more attention to those annotated word-sememe pairs.

In SPSE, we learn sememe embeddings accompanying with word embeddings via matrix factorization into the unified low-dimensional semantic space. Matrix factorization has been verified as an effective approach in personalized recommendation, because it can accurately model relatedness between users and items, and is highly robust to noises in user-item matrices. Using this model, we can flexibly compute semantic relatedness of words and sememes, which provides us an effective tool to manipulate and manage sememes, including but not limited to sememe prediction.

3.4 Sememe Prediction with Aggregated Sememe Embeddings

In HowNet, sememes are annotated as the tiny semantic components of words. Inspired by the characteristics of sememes, we assume that the word embeddings are semantically composed of sememe embeddings. In the word-sememe joint space, we can simply implement semantic composition as additive operations that each word embedding is expected to be the sum of its all sememes' embeddings. Following this assumption, we propose Sememe Prediction with Aggregated Sememe Embeddings (SPASE). SPASE is also based on matrix factorization, and is formally denoted as:

$$\mathbf{w}_i = \sum_{s_j \in S_{w_i}} \mathbf{M}'_{ij} \cdot \mathbf{s}_j, \tag{3}$$

where S_{w_i} is the sememe set of the word w_i, and $\mathbf{M}'_{ij}$ represents the weight of sememe s_j for word w_i, which only has value on non-zero elements of word-sememe labelled matrix $\mathbf{M}$. To learn sememe embeddings, we attempt to decompose the word embedding matrix $\mathbf{W}$ into $\mathbf{M}'$ and sememe embedding matrix $\mathbf{S}$, with word embeddings are pre-trained and fixed during training, which could also be written as $\mathbf{W} = \mathbf{M}' \times \mathbf{S}$.

The contribution of SPASE is that it complies with the definition of sememes in HowNet that sememes are the semantic components of words. In SPASE, each sememe can be regarded as a tiny semantic unit, and all words can be represented by composing several semantic units, i.e., sememes, which makes up an interesting semantic regularity. However, SPASE is difficult to train because word embeddings are fixed and the number of words is much larger than the number of sememes. In the case of modeling complex semantic compositions of sememes into words, the representation capability of SPASE may be strongly constrained by limited parameters of sememe embeddings and excessive simplification of additive assumption.

3.5 Sememe Prediction with Ensemble Model

We propose three models including SPWE, SPSE and SPASE for sememe prediction. SPWE and SPSE/SPASE take two different approaches and have different characteristics. SPWE is inspired by collaborative filtering and recommend sememes which belong to similar words, while SPSE/SPASE are instructed by matrix factorization which predict sememes according to both word and sememe embeddings in the joint semantic space. These two approaches are complementary and should be integrated into an ensemble model for sememe prediction.

For example, in SPSE/SPASE, all color sememes such as *white* and *blue* tend to be learned close to each other in sememe embeddings. When predicting sememes for the word *black*, all these color sememes will get high ranks since their embeddings are similar to *black*, which is obviously not true. Meanwhile, SPWE can well address this issue for it can learn discriminative annotation structures of word sememes. We also find that SPSE/SPASE works better when dealing with those words having unique sememes, while SPWE performs well with complicated sememes. In this paper, we integrate two models by simple weighted addition of their recommendation scores, resulting in improvements for sememe prediction as will shown in experiments.

4 Experiment

In experiment, we evaluate our models mainly on the task of sememe prediction. Moreover, we also conduct detailed case study for further intuitive comparisons. In the following sections, we first introduce the dataset we utilize for sememe prediction, and then introduce the experimental settings of both baselines and our models. Next, we demonstrate the experimental results in sememe prediction with different evaluation metrics, and give detailed analysis on these results. Finally, we conduct extensive case studies on the analysis of recommended sememes for unlabelled words, the performance variance of words with different part-of-speech tags and frequencies.

4.1 Dataset

We utilize the sememe annotations in HowNet for sememe prediction. HowNet contains $212,539$ senses with annotations belonging to $103,843$ words. The number of sememes in total is approximately $2,000$. Since many sememes only appear quite few times in HowNet, which are expected to be unimportant sememes. We wipe out all these sememes with low frequencies, and the number of distinct sememes finally used in our dataset is $1,400$. We use the Sogou-T corpus [2] as the text corpus to learn Chinese word embeddings. Sogou-T is provided by a Chinese commercial search engine, which contains 2.7 billion words in total.

4.2 Experimental Settings

We evaluate our sememe prediction models including Sememe Prediction with Word Embedding (SPWE), Sememe Prediction with Sememe Embedding (SPSE) and Sememe Prediction with Aggregated Sememe Embedding (SPASE) on sememe prediction. Moreover, we also implement two enhanced ensemble models with different combination strategies, i.e., SPWE+SPSE and SPWE+SPASE. In SPWE, we predict sememes for a word according to the most common sememes of those related words in HowNet. In SPSE and SPASE, we predict sememes according to the cosine similarity between word embeddings and sememe embeddings. In the ensemble models, we merge the scores of both integrated methods together with a pre-defined fixed weights to predict sememes.

As for baselines, since there have been few previous works on sememe prediction, we choose some conventional and straightforward methods as our baselines. Specifically, we utilize the word embeddings learned by GloVe [Pennington *et al.*, 2014] as word feature vectors, and then directly use logistic regression for sememe prediction with the learned word embeddings taken as inputs, considering sememe prediction as a multi-label classification task. The sememes in HowNet are regarded as classification labels to be predicted.

[2] https://www.sogou.com/labs/resource/t.php

For fair comparisons, the corpus for learning word embeddings as well as the embedding dimension of words and sememes are the same including all baselines and our sememe prediction models.

We empirically set the dimension of word and sememe embeddings to be 200. In SPSE, we set the probability of zero elements to be decomposed in word-sememe matrix as 0.5%, and select the initial learning rate to be 0.01, which will descend through iterations. We set the ratio λ in Equation (2) to be 0.5. In SPWE, we set the hyper-parameter p to be 0.2 and the number of most related words $K = 100$. In ensemble models, we have tested on different weights and choose $\lambda_1/\lambda_2 = 2.1$. In HowNet, we find $66,126$ words with labelled sememes which appear at least 50 times in the Sogou-T corpus for learning word embeddings, and we divide $60,000$ of the words into train set and the rest $6,216$ of them into test set. We empirically select our parameters with the best performance on sememe prediction.

4.3 Sememe Prediction

Evaluation Protocol

Since a large amount of words have multiple sememes, the task of sememe prediction could be viewed as a multi-label classification task. In evaluation, we utilize mean average precision (MAP) as evaluation metric.

Experimental Results

Table 1 shows the evaluation results of these models on sememe prediction. From the table we can observe that:

(1) The ensemble models perform better as compared to those single sememe prediction methods, in which SPWE+SPSE achieves the best performance. It indicates that the ensemble model can combine the advantages in both SPWE and SPSE models. It is because that SPWE can learn the elaborate structures of sememes according to related words, while SPSE can provide latent relationships between words and sememes. The two methods are complementary and combining these two kinds of methods will actually improve the prediction performance.

(2) SPWE seems to be better than SPSE and SPASE models. It is because that the SPWE model predicts appropriate sememes according to the related words, which exactly matches the real-world situation in sememe prediction. Unlike conventional recommender systems where most user-item matrices are typically noisy and incomplete, HowNet is annotated carefully by human experts. In this case, the word-sememe co-occurrences are much more accurate as compared to user-item matrices in recommender systems. Hence we can obtain good performance simply using the idea of collaborative filtering in SPWE. Logistic regression is similar to SPWE because it also utilizes word embeddings as certain kind of features to extract discriminative patterns for classification. However, the improvement introduced by ensemble models also indicates the significance in SPSE by modeling latent relationships of words and sememes. Nevertheless, the expert annotation in HowNet doesn't cover all appropriate sememes and collaborative filtering will capture the preferences of expert annotations, which causes the SPWE performs better though SPSE and SPASE.

(3) SPSE performs better as compared to SPASE. It is because that word embeddings are fixed during training of SPASE, so it is very difficult to learn effective sememe embeddings to fit the assumption that word embeddings are the sum of sememe embeddings. It is fair to say that, although the assumption in SPASE fits well with the original definition of sememes in HowNet, the limited sememe parameters are still hard to represent complex semantic meanings of words in real world. It also suggests that simplified additive assumption of semantic composition from sememes to words also leads to the decrease in prediction performance.

(4) In all, the absolute score of MAP that our ensemble model achieves is quite high and better than the baseline achieved by conventional multi-label classification methods like logistic regression, which means the sememe annotations in HowNet are reasonable and effective. It also implies that our models are capable of modeling sememe and word embeddings well for sememe prediction.

Table 1: Evaluation results of sememe prediction.

Method	MAP
SPSE	0.554
SPASE	0.506
GloVe+LR	0.662
SPWE	0.676
SPWE+SPASE	0.683
SPWE+SPSE	**0.713**

4.4 Case Study

In case study, we give some further analysis to explain the effectiveness of our models with detailed cases. Moreover, we also explore the performance variance on sememe prediction of those words with different POS tags and frequencies.

Analysis on Predicted Sememes

As shown in Table 2, we list the top 5 sememes we predict for five words *webaholic*, *express mail*, *film industry*, *rafting* and *ram*. The blackened sememes are the true sememes for each word. From these examples we can conclude that:

(1) In the first three cases, true sememes are all ranked in top positions, which indicates that our models work well and predict sememes accurately for these words. Especially, for *webaholic*, we not only predict *human* and *internet* which are closely related to *webaholic* in top position, but also predict *frequency* and *use* successfully, which are regarded as general sememes and thus are difficult to predict.

(2) For the forth word *rafting*, we do not predict any true sememes in the top 5 predictions. In HowNet, the sememes of *rafting* include *sports*, *exercise*, *float* and *fact*. However, the sememes we predict are also acceptable if we understand *rafting* as "tour by ship for entertainment". In fact, there may exist many appropriate sememes for a word. Since HowNet is manually annotated by experts, some acceptable predictions may not always agree with what annotated in HowNet, which in some cases will under-estimate our models.

(3) For the word *ram*, we predict *livestock* and *male* successfully, but the sememe *female* is also in the top 3 predic-

Table 2: Some examples of sememe prediction.

words	Top 5 sememes prediction
网迷(webaholic)	人**(human)**, 因特网**(internet)**, 经常**(frequency)**, 利用**(use)**, 喜欢(fond of)
专递(express mail)	邮寄**(post)**, 信件**(letter)**, 快**(fast)**, 事情(fact), 车(landvehicle)
电影业(film industry)	事务**(affairs)**, 艺**(entertainment)**, 表演物**(shows)**, 拍摄(take picture), 制造**(produce)**
漂流(rafting)	船(ship), 旅游(tour), 游(swim), 水域(waters), 消闲(whileaway)
公羊(ram)	牲畜**(livestock)**, 男**(male)**, 女(female), 走兽(beast), 饲养(foster)

tions, which can reveal some issues of our models. *male* is relative to *female* and they are close to each other in embedding space. Besides, they happen to be sememes of *ram*'s most related words such as *boar*, *ewe* and *sow* which makes *female* get nearly the same score compared to *male*. Our model cannot distinguish this kind of sememes very well and such situation will affect our prediction results.

Influences of POS Tags on Sememe Prediction

As listed in Table 3, we can observe that the Part of speech (POS) tags of words have great influence on sememe prediction. It is much easier to predict sememes for nouns than other POS of words since nouns are more concrete and unitary. Specifically, the concept of sememe is more reasonable and straightforward for nouns as they are easier to be semantically decomposed to sememes. This situation could be found with the examples in Table 2 compared to those of verbs, adjectives and adverbs. Besides, similar nouns tend to share the same sememes, such like different cities all sharing the sememes of *city*, *place* and *ProperName*. The effectiveness of sememe prediction of nouns makes it available to be applied in real-world sememe-embodied applications.

Table 3: Results of different POS tags on sememe prediction

POS	number of words	MAP
adverb	136	0.568
adjective	808	0.544
verb	1,867	0.583
noun	3,556	0.747

Influences of Word Frequencies on Sememe Prediction

As listed in Table 4, we can observe that word frequency also has great impacts on sememe prediction. The experimental results show that the more a word appears in the corpus, the more difficult for us to predict its sememes. It is because that, on one hand, words with high frequency are widely used in daily life, which are usually common verbs and adverbs. These words have more and boarder senses than those low-frequency words, of which the sememes are even unrelated to each other. Therefore, it is extremely hard to be predicted all these sememes for a common word based on simple similarities. On the other hand, the low-frequency words tend to contain less and simpler sememes as compared to those words of high frequencies, and thus are easier to be predicted as stated above. Moreover, since we have constrained the word frequency used in training to be higher than a threshold, the low-frequency words could also learn relatively good word representations though trained less than high-frequency words.

Table 4: Results of different word frequencies on sememe prediction

word frequency	number of words	MAP
$<$800	1,659	0.817
800 - 3,000	1,494	0.736
3,001 - 15,000	1,672	0.690
$>$15,000	1,311	0.596

5 Conclusion and future work

In this paper, we propose a novel task of sememe prediction, and propose several prediction models based on word embeddings and sememe embeddings inspired by collaborative filtering and matrix factorization. We evaluate our sememe prediction models on a real-world database HowNet. From the experimental results, we can find that our models are effective and achieve promising results, which also confirms the significance of internal relations between words and sememes. The source code of this paper can be obtained from `https://github.com/thunlp/Sememe_prediction`.

We will explore the following research directions in future: (1) We will explore better models to learn sememe and word embeddings simultaneously. In this paper, we fix word embeddings to learn sememe embeddings, which makes it hard to flexibly learn semantic relations between words and sememes in semantic space. Joint learning of word and sememe embeddings will enable our models to better encode semantic relations in HowNet and large-scale corpus. (2) HowNet contains rich structured information for words and sememes as shown in our paper, which is not utilized in the current version. Besides, some words have multiple senses to represent distinct meanings, while we regard as there are no differences between different senses. In future, we will extend our models to consider the sememe structures as well as the sense information. (3) The sememes are considered as the minimum semantic units of human languages, which are believed to be universal for all languages. We will explore the effectiveness of sememes in different languages. Moreover, our models achieve better performance on sememe prediction for noun words, and the sememes usually describe the essential attributes and properties of these noun words. In future, we will explore to utilize sememes to enhance the construction of real-world knowledge graphs, where sememes could be used to give further information of entities.

Acknowledgments

This work is supported by the 973 Program (No. 2014CB340501), the National Natural Science Foundation of China (NSFC No. 61661146007, 61532010), and Tsinghua University Initiative Scientific Research Program (20151080406).

References

[Auer *et al.*, 2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[Bloomfield, 1926] Leonard Bloomfield. A set of postulates for the science of language. *Language*, 2(3):153–164, 1926.

[Bobadilla *et al.*, 2013] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.

[Bollacker *et al.*, 2008] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of KDD*, pages 1247–1250, 2008.

[Dong and Dong, 2006] Zhendong Dong and Qiang Dong. *HowNet and The Computation of Meaning*. World Scientific, 2006.

[Duan *et al.*, 2007] Xiangyu Duan, Jun Zhao, and Bo Xu. Word sense disambiguation through sememe labeling. In *Proceedings of IJCAI*, pages 1594–1599, 2007.

[Gan and Wong, 2000] Kok Wee Gan and Ping Wai Wong. Annotating information structures in chinese texts using hownet. In *Proceedings of ACL*, pages 85–92, 2000.

[Gan *et al.*, 2002] Kok-Wee Gan, Chi-Yung Wang, and Brian Mak. Knowledge-based sense pruning using the hownet: an alternative to word sense disambiguation. In *International Symposium on Chinese Spoken Language Processing*, 2002.

[Goddard and Wierzbicka, 1994] Cliff Goddard and Anna Wierzbicka. *Semantic and lexical universals: Theory and empirical findings*, volume 25. John Benjamins Publishing, 1994.

[Hinton, 1986] Geoffrey E Hinton. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA, 1986.

[Hoffart *et al.*, 2013] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61, 2013.

[Huang *et al.*, 2014] Minlie Huang, Borui Ye, Yichen Wang, Haiqiang Chen, Junjun Cheng, and Xiaoyan Zhu. New word detection for sentiment analysis. In *Proceedings of ACL*, pages 531–541, 2014.

[Koren *et al.*, 2009] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

[Levy and Goldberg, 2014] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Proceedings of NIPS*, pages 2177–2185, 2014.

[Liu and Li, 2002] Qun Liu and Sujian Li. Word similarity computing based on how-net. *Computational Linguistics and Chinese Language Processing*, 7(2):59–76, 2002.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, 2013.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Mintz *et al.*, 2009] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP*, pages 1003–1011. Association for Computational Linguistics, 2009.

[Nastase *et al.*, 2013] Vivi Nastase, Preslav Nakov, Diarmuid O Seaghdha, and Stan Szpakowicz. Semantic relations between nominals. *Synthesis Lectures on Human Language Technologies*, 6(1):1–119, 2013.

[Navigli and Ponzetto, 2012] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014.

[Rendle, 2010] Steffen Rendle. Factorization machines. In *Proceedings of ICDM*, pages 995–1000. IEEE, 2010.

[Sarwar *et al.*, 2001] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of World Wide Web*, pages 285–295. ACM, 2001.