Cross-lingual Lexical Sememe Prediction

Fanchao Qi[1*], **Yankai Lin**[1*], **Maosong Sun**[1,2†], **Hao Zhu**[1], **Ruobing Xie**[3], **Zhiyuan Liu**[1]
[1]Department of Computer Science and Technology, Tsinghua University
Institute for Artificial Intelligence, Tsinghua University
State Key Lab on Intelligent Technology and Systems, Tsinghua University
[2]Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University
[3]Search Product Center, WeChat Search Application Department, Tencent, China
`{qfc17, linyk14}@mails.tsinghua.edu.cn, sms@tsinghua.edu.cn`
`zhuhao15@mails.tsinghua.edu.cn`
`xrbsnowing@163.com, liuzy@tsinghua.edu.cn`

Abstract

Sememes are defined as the minimum semantic units of human languages. As important knowledge sources, sememe-based linguistic knowledge bases have been widely used in many NLP tasks. However, most languages still do not have sememe-based linguistic knowledge bases. Thus we present a task of cross-lingual lexical sememe prediction, aiming to automatically predict sememes for words in other languages. We propose a novel framework to model correlations between sememes and multi-lingual words in low-dimensional semantic space for sememe prediction. Experimental results on real-world datasets show that our proposed model achieves consistent and significant improvements as compared to baseline methods in cross-lingual sememe prediction. The codes and data of this paper are available at `https://github.com/thunlp/CL-SP`.

1 Introduction

Words are regarded as the smallest meaningful unit of speech or writing that can stand by themselves in human languages, but not the smallest indivisible semantic unit of meaning. That is, the meaning of a word can be represented as a set of semantic components. For example, "*Man* = *human* + *male* + *adult*" and "*Boy* = *human* + *male* + *child*". In linguistics, the minimum semantic unit of meaning is named sememe (Bloomfield, 1926). Some people believe that semantic meanings of concepts such as words can be composed of a limited closed set of sememes. And sememes can help us comprehend human languages better.

Unfortunately, the lexical sememes of words are not explicit in most human languages. Hence, people construct sememe-based linguistic knowledge bases (KBs) via manually annotating every words with a pre-defined closed set of sememes. HowNet (Dong and Dong, 2003) is one of the most well-known sememe-based linguistic KBs. Different from WordNet (Miller, 1995) which focuses on the relations between senses, it annotates each word with one or more relevant sememes. As illustrated in Fig. 1, the word *apple* has two senses including *apple (fruit)* and *apple (brand)* in HowNet. The sense *apple (fruit)* has one sememe *fruit*, and the sense *apple (brand)* has five sememes including *computer*, *PatternValue*, *able*, *bring* and *SpecificBrand*. There exist about 2,000 sememes and over 100 thousand labeled Chinese and English words in HowNet. HowNet has been widely used in various NLP applications such as word similarity computation (Liu and Li, 2002), word sense disambiguation (Zhang et al., 2005), question classification (Sun et al., 2007) and sentiment classification (Dang and Zhang, 2010).

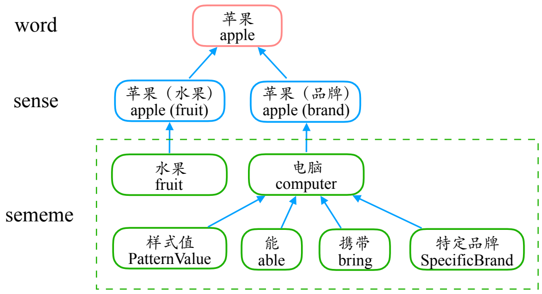

Figure 1: An example of HowNet.

However, most languages do not have such sememe-based linguistic KBs, which prevents us understanding and utilizing human languages to a greater extent. Therefore, it is important to build sememe-based linguistic KBs for various languages. Manual construction for sememe-based linguistic KBs requires efforts of many linguistic experts, which is time-consuming and

[*] Indicates equal contribution
[†] Corresponding author

labor-intensive. For example, the construction of HowNet has cost lots of Chinese linguistic experts more than 10 years.

To address the issue of the high labor cost of manual annotation, we propose a new task, cross-lingual lexical sememe prediction (CLSP) which aims to automatically predict lexical sememes for words in other languages. CLSP aims to assist in the annotation of linguistic experts. There are two critical challenges for CLSP: (1) There is not a consistent one-to-one match between words in different languages. For example, English word "beautiful" can refer to Chinese words of either "美丽" or "漂亮". Hence, we cannot simply translate HowNet into another language. And how to recognize the semantic meaning of a word in other languages becomes a critical problem. (2) Since there is a gap between the semantic meanings of words and sememes, we need to build semantic representations for words and sememes to capture the semantic relatedness between them.

To tackle these challenges, in this paper, we propose a novel model for CLSP, which aims to transfer sememe-based linguistic KBs from source language to target language. Our model contains three modules including (1) monolingual word embedding learning which is intended for learning semantic representations of words for source and target languages respectively; (2) cross-lingual word embedding alignment which aims to bridge the gap between the semantic representations of words in two languages; (3) sememe-based word embedding learning whose objective is to incorporate sememe information into word representations. For simplicity, we do not consider the hierarchy information in HowNet in this paper.

In experiments, we take Chinese as source language and English as target language to show the effectiveness of our model. Experimental results show that our proposed model could effectively predict lexical sememes for words with different frequencies in other languages. Our model also has consistent improvements on two auxiliary experiments including bilingual lexicon induction and monolingual word similarity computation by jointly learning the representations of sememes, words in source and target languages.

2 Related Work

Since HowNet was published (Dong and Dong, 2003), it has attracted wide attention of researchers. Most of related works focus on applying HowNet to specific NLP tasks (Liu and Li, 2002; Zhang et al., 2005; Sun et al., 2007; Dang and Zhang, 2010; Fu et al., 2013; Niu et al., 2017; Zeng et al., 2018; Gu et al., 2018). To the best of our knowledge, only Xie et al. (2017) and Jin et al. (2018) conduct studies of augmenting HowNet by recommending sememes for new words. However, both of the two works are aimed to recommend sememes for monolingual words and not applicable to cross-lingual circumstance. Accordingly, our work is the first effort to automatically perform cross-lingual sememe prediction to enrich sememe-based linguistic KBs.

Our novel model adopts the method of word representation learning (WRL). Recent years have witnessed great advances in WRL. Models like Skip-gram, CBOW (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) are immensely popular and achieve remarkable performance in many NLP tasks. However, most WRL methods learn distributional information of words from large corpora while the valuable information contained in semantic lexicons are disregarded. Therefore, some works try to inject semantic information of KBs into WRL (Faruqui et al., 2015; Liu et al., 2015; Mrkšic et al., 2016; Bollegala et al., 2016). Nevertheless, these works are all applied to word-based KBs such as WordNet, few works pay attention to how to incorporate the knowledge from sememe-based linguistic KBs.

There also have been plenty of studies working on cross-lingual WRL (Upadhyay et al., 2016; Ruder, 2017). Most of them require parallel corpora (Zou et al., 2013; AP et al., 2014; Hermann and Blunsom, 2014; Kočiskỳ et al., 2014; Gouws et al., 2015; Luong et al., 2015; Coulmance et al., 2015). Some of them adopt unsupervised or weakly supervised methods (Mikolov et al., 2013b; Vulić and Moens, 2015; Conneau et al., 2017; Artetxe et al., 2017). There are also some works using a seed lexicon as the cross-lingual signal (Dinu et al., 2014; Faruqui and Dyer, 2014; Lazaridou et al., 2015; Shi et al., 2015; Lu et al., 2015; Gouws et al., 2015; Wick et al., 2016; Ammar et al., 2016; Duong et al., 2016; Vulić and Korhonen, 2016).

In terms of our cross-lingual sememe prediction task, parallel data-based bilingual WRL methods are unsuitable because most language pairs have no large parallel corpora. Besides, unsupervised methods are not appropriate either as they are generally hard to learn high-quality bilingual word embeddings. Therefore, we choose the seed lexicon method in our model, and further introduce matching mechanism that is inspired by Zhang et al. (2017) to enhance its performance.

3 Methodology

In this section, we introduce our novel model for CLSP. Here we define the language with sememe annotations as source language and the language without sememe annotations as target language. The main idea of our model is to learn word embeddings of source and target languages jointly in a unified semantic space, and then predict sememes for words in target language according to the words with similar semantic meanings in source language.

Our method consists of three parts: monolingual word representation learning, cross-lingual word embedding alignment and sememe-based word representation learning. Hence, we define the objective function of our method corresponding to the three parts:

$$\mathcal{L} = \mathcal{L}_{mono} + \mathcal{L}_{cross} + \mathcal{L}_{sememe}. \tag{1}$$

Here, the monolingual term $\mathcal{L}_{mono}$ is designed for learning monolingual word embeddings from non-parallel corpora for source and target languages respectively. The cross-lingual term $\mathcal{L}_{cross}$ aims to align cross-lingual word embeddings in a unified semantic space. And $\mathcal{L}_{sememe}$ can draw sememe information into word representation learning and conduce to better word embeddings for sememe prediction. In the following subsections, we introduce the three parts in detail.

3.1 Monolingual Word Representation

Monolingual word representation is responsible for explaining regularities in monolingual corpora of source and target languages. Since the two corpora are non-parallel, $\mathcal{L}_{mono}$ comprises two monolingual sub-models that are independent of each other:

$$\mathcal{L}_{mono} = \mathcal{L}^S_{mono} + \mathcal{L}^T_{mono}, \tag{2}$$

where the superscripts S and T denote source and target languages respectively.

As a common practice, we choose the well established Skip-gram model to obtain monolingual word embeddings. Skip-gram model is aimed at maximizing the predictive probability of context words conditioned on the centered word. Formally, taking the source side for example, given a training word sequence $\{w^S_1, \cdots, w^S_n\}$, Skip-gram model intends to minimize:

$$\mathcal{L}^S_{mono} = -\sum_{c=K+1}^{n-K} \sum_{-K\le k\le K, k\ne 0} \log P(w^S_{c+k}|w^S_c), \tag{3}$$

where K is the size of the sliding window. $P(w^S_{c+k}|w^S_c)$ stands for the predictive probability of one of the context words conditioned on the centered word w^S_c, formalized by the following softmax function:

$$P(w^S_{c+k}|w^S_c) = \frac{\exp(\mathbf{w}^S_{c+k} \cdot \mathbf{w}^S_c)}{\sum_{w^S_s \in V^S} \exp(\mathbf{w}^S_s \cdot \mathbf{w}^S_c)}, \tag{4}$$

in which V^s indicates the word vocabulary of source language. $\mathcal{L}^T_{mono}$ can be formulated similarly.

3.2 Cross-lingual Word Embedding Alignment

Cross-lingual word embedding alignment aims to build a unified semantic space for the words in source and target languages. Inspired by Zhang et al. (2017), we align the cross-lingual word embeddings with signals of a seed lexicon and self-matching.

Formally, $\mathcal{L}_{cross}$ is composed of two terms including alignment by seed lexicon $\mathcal{L}_{seed}$ and alignment by matching $\mathcal{L}_{match}$:

$$\mathcal{L}_{cross} = \lambda_s \mathcal{L}_{seed} + \lambda_m \mathcal{L}_{match}, \tag{5}$$

where λ_s and λ_m are hyperparameters for controlling relative weightings of the two terms.

Alignment by Seed Lexicon

The seed lexicon term $\mathcal{L}_{seed}$ encourages word embeddings of translation pairs in a seed lexicon $\mathcal{D}$ to be close, which can be achieved via a L_2 regularizer:

$$\mathcal{L}_{seed} = \sum_{\langle w^S_s, w^T_t \rangle \in \mathcal{D}} \|\mathbf{w}^S_s - \mathbf{w}^T_t\|^2, \tag{6}$$

in which w^S_s and w^T_t indicate the words in source and target languages in the seed lexicon respectively.

Alignment by Matching Mechanism

As for the matching process, it is founded on an assumption that each target word should be matched to a single source word or a special empty word, and vice versa. The goal of the matching process is to find the matched source (target) word for each target (source) word and maximize the matching probabilities for all the matched word pairs. The loss of this part can be formulated as:

$$\mathcal{L}_{match} = \mathcal{L}_{match}^{T2S} + \mathcal{L}_{match}^{S2T}, \tag{7}$$

where $\mathcal{L}_{match}^{T2S}$ is the term for target-to-source matching and $\mathcal{L}_{match}^{S2T}$ is the term for source-to-target matching.

Next, we give a detailed explanation of target-to-source matching, and the source-to-target matching is defined in the same way. We first introduce a latent variable $m_t \in \{0, 1, \cdots, |V^S|\}$ $(t = 1, 2, \cdots, |V^T|)$ for each target word w_t^T, where $|V^S|$ and $|V^T|$ indicate the vocabulary size of source and target languages respectively. Here, m_t specifies the index of the source word that w_t^T matches with, and $m_t = 0$ signifies the empty word is matched. Then we have $\mathbf{m} = \{m_1, m_2, \cdots, m_{|V^T|}\}$, and can formalize the target-to-source matching term:

$$\begin{aligned}\mathcal{L}_{match}^{T2S} &= -\log P(\mathcal{C}^T|\mathcal{C}^S)\\ &= -\log \sum_{\mathbf{m}} P(\mathcal{C}^T, \mathbf{m}|\mathcal{C}^S),\end{aligned} \tag{8}$$

where $\mathcal{C}^T$ and $\mathcal{C}^S$ denote the target and source corpus respectively. Here, we simply assume that the matching processes of target words are independent of each other. Therefore, we have:

$$\begin{aligned}P(\mathcal{C}^T, \mathbf{m}|\mathcal{C}^S) &= \prod_{w^T \in \mathcal{C}^T} P(w^T, \mathbf{m}|\mathcal{C}^S)\\ &= \prod_{t=1}^{|V^T|} P(w_t^T|w_{m_t}^S)^{c(w_t^T)},\end{aligned} \tag{9}$$

where $w_{m_t}^S$ is the source word that w_t^T matches with, and $c(w_t^T)$ is the number of times w_t^T occurs in the target corpus.

3.3 Sememe-based Word Representation

Sememe-based word representation is intended for improving word embeddings for sememe prediction by introducing the information of sememe-based linguistic KBs of source language. In this section, we present two methods of sememe-based word representation.

Word Relation-based Approach

A simple and intuitive method is to let words with similar sememe annotations tend to have similar word embeddings, which we name word relation-based approach. To begin with, we construct a synonym list from sememe-based linguistic KBs of source language, where we regard words sharing a certain number of sememes as synonyms. Next, we force synonyms to have closer word embeddings.

Formally, we let $\mathbf{w}_i^S$ be original word embedding of w_i^S and $\hat{\mathbf{w}}_i^S$ be its adjusted word embedding. And let $\text{Syn}(w_i^S)$ denote the synonym set of word w_i^S. Then the loss function is:

$$\begin{aligned}\mathcal{L}_{sememe} = \sum_{w_i^S \in V^S} \Big[\alpha_i \|\mathbf{w}_i^S - \hat{\mathbf{w}}_i^S\|^2 +\\ \sum_{w_j^S \in \text{Syn}(w_i^S)} \beta_{ij} \|\hat{\mathbf{w}}_i^S - \hat{\mathbf{w}}_j^S\|^2\Big],\end{aligned} \tag{10}$$

where α and β control the relative strengths of the two terms. It should be noted that the idea of forcing similar words to have close word embeddings is similar to the state-of-the-art retrofitting approach (Faruqui et al., 2015). However, retrofitting approach cannot be applied here because sememe-based linguistic KBs such as HowNet cannot directly provide its needed synonym list.

Sememe Embedding-based Approach

Simple and effective as the word relation-based approach is, it cannot make full use of the information of sememe-based linguistic KBs because it disregards the complicated relations between sememes and words as well as relations between different sememes. To address this limitation, we propose sememe embedding-based approach, which learns both sememe and word embeddings jointly.

In this approach, we represent sememes with distributed vectors as well and place them into the same semantic space as words. Similar to SPSE (Xie et al., 2017), which learns sememe embeddings by decomposing word-sememe matrix and sememe-sememe matrix, our method utilizes sememe embeddings as regularizers to learn better word embeddings. Different from SPSE, we do not use pre-trained word embeddings. Instead, we learn word embeddings and sememe embeddings simultaneously.

More specifically, from HowNet we can extract a source-side word-sememe matrix M^S with $M_{sj}^S = 1$ indicating word w_s^S is annotated with sememe x_j, otherwise $M_{sj}^S = 0$. Hence by factorizing M^S, we can define the loss function as:

$$\mathcal{L}_{sememe} = \sum_{w_s^S \in V^S, x_j \in X} (\mathbf{w}_s^S \cdot \mathbf{x}_j + b_s + b'_j - M_{sj}^S)^2, \tag{11}$$

where b_s and b'_j are the biases of w_s^S and x_j, and X denotes sememe set.

In this approach, we obtain word and sememe embeddings in a unified semantic space. The sememe embeddings bear all the information about the relationships between words and sememes, and they inject the information into word embeddings. Therefore, the word embeddings are expected to be more suitable for sememe prediction.

3.4 Training and Prediction

Training

When training monolingual word embeddings, we use negative sampling following Mikolov et al. (2013a). In the optimization of sememe part, we adopt the iterative updating method following Faruqui et al. (2015) for word relation-based approach and stochastic gradient descent (SGD) for sememe embedding-based approach. As for the optimization of the seed lexicon term of cross-lingual part, we also apply SGD.

Nevertheless, due to the existence of the latent variable, optimization of the matching process in cross-lingual part poses a challenge. We settle on Viterbi EM algorithm to address the problem. Next, we still take the target-to-source side as an example and give a detailed description of the training process using Viterbi EM algorithm.

Viterbi EM algorithm alternates between a Viterbi E step and a subsequent M step. The Viterbi E step aims to find the most probable matched word pairs given the current parameters. Considering the independence, we can seek the match for each word individually:

$$\hat{m}_t = \mathop{\arg\max}_{s\in\{0,1,\cdots,|V^S|\}} P(w_t^T|w_s^S). \tag{12}$$

As for the parametrization of the matching probability, there are various choices. For computational simplicity, we select cosine similarity:

$$P(w_t^T|w_s^S) = \begin{cases} \epsilon & \text{if } s = 0, \\ cos(\mathbf{w}_t^T, \mathbf{w}_s^S) & \text{otherwise,} \end{cases} \tag{13}$$

where ϵ is a hyperparameter indicating the probability of matching the empty word. Therefore, the Viterbi E step computes matching by:

$$\tilde{m}_t = \mathop{\arg\max}_{s\in\{1,\cdots,|V^S|\}} cos(\mathbf{w}_t^T, \mathbf{w}_s^S), \tag{14}$$

$$\hat{m}_t = \begin{cases} \tilde{m}_t & \text{if } cos(\mathbf{w}_t^T, \mathbf{w}_{\tilde{m}_t}^S) > \epsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

From this, we can see that ϵ serves as a threshold to keep out unreliable matched pairs.

The Viterbi M step performs maximization as if the latent variable has been observed in the Viterbi E step. Thus, we can treat the matched pairs as correct translations, and use a L_2 regularizer as well. Consequently, the M step computes:

$$(\hat{\mathbf{w}}^S, \hat{\mathbf{w}}^T) = \mathop{\arg\max}_{\mathbf{w}^S, \mathbf{w}^T} \mathcal{M}(\mathbf{w}^S, \mathbf{w}^T), \tag{16}$$

where $\mathcal{M}(\mathbf{w}^S, \mathbf{w}^T)$ is defined as:

$$\mathcal{M}(\mathbf{w}^S, \mathbf{w}^T) = -\sum_{t=1}^{|V^T|} \mathbb{I}[\tilde{m}_t \neq 0] \frac{c(w_t^T)}{|\mathcal{C}^T|} \|\mathbf{w}_t^T - \mathbf{w}_{\tilde{m}_t}^S\|^2. \tag{17}$$

Prediction

Since we assume that words with similar sememe annotations are similar and similar words should have similar sememes, which resembles collaborative filtering in personalized recommendation, we can recommend sememes for target words according to their most similar source words.

Formally, we define the score function $P(x_j|w_t^T)$ of sememes x_j given a target word w_t^T as:

$$P(x_j|w_t^T) = \sum_{w_s^S \in V^S} cos(\mathbf{w}_s^S, \mathbf{w}_t^T) \cdot M_{sj}^S \cdot c^{r_s}, \tag{18}$$

where r_s is the descending rank of word similarity $cos(\mathbf{w}_s^S, \mathbf{w}_t^T)$ for the source word w_s^S, and $c \in (0, 1)$ is a hyperparameter. Thus, c^{r_s} is a declined confidence factor which can eliminate the noise from irrelevant source words and concentrate on the most similar source words when predicting sememes for target words.

4 Experiments

In this section, we first introduce the dataset used in the experiments and then describe the experimental settings of both baseline method and our model. Next, we present the experimental results of different methods on the task of cross-lingual lexical sememe prediction. And then we conduct detailed analysis and exhaustive case studies. Following this, we investigate the effect of word frequency on cross-lingual sememe prediction results. Finally, we perform further quantitative analysis through two sub-tasks including bilingual lexicon induction and word similarity computation.

4.1 Dataset

We use sememe annotations in HowNet for sememe prediction. HowNet annotates sememes for $118,346$ Chinese words and $104,025$ English words. The number of sememes in total is $1,983$. Since some sememes only appear few times in HowNet, which are expected to be unimportant, we filter out those low-frequency sememes. Specifically, the frequency threshold is 5, and the final number of distinct sememes used in our experiments is $1,400$.

In our experiments, Chinese is source language and English is target language. To learn Chinese and English monolingual word embeddings, we extract about 2.0G text from Sogou-T[1] and Wikipedia[2] respectively. And we use THULAC[3] (Li and Sun, 2009) for Chinese word segmentation.

As for seed lexicon, we build it in a similar way to Zhang et al. (2017). First, we employ Google Translation API[4] to translate the source side (Chinese) vocabulary. Then the translations in the target language (English) are queried again in the reverse direction to translate back to the source language (Chinese). And we only keep the translation pairs whose back translated words match with the original source words.

In the task of bilingual lexicon induction, we opt for Chinese-English Translation Lexicon Version 3.0[5] to be the gold standard. In the task of word similarity computation, we choose WordSim-240 and WordSim-297 (Jin and Wu, 2012) datasets for Chinese, and WordSim-353 (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2015) datasets for English to evaluate the performance of our model. These datasets contain word pairs as well as human-assigned similarity scores. The word vectors are evaluated by ranking the word pairs according to their cosine similarities, and measuring Spearman's rank correlation coefficient with the human ratings.

4.2 Experimental Settings

We empirically set the dimension of word and sememe embeddings to 200. And the embeddings are all randomly initialized. In monolingual word embedding learning, we follow the optimal parameter settings in Mikolov et al. (2013a). We set the window size K to 5, down-sampling rate for high-frequency words to 10^{-5}, learning rate to 0.025 and the number of negative samples to 5. In cross-lingual word embedding alignment, the seed lexicon term weight λ_s is 0.01, and the matching term weight λ_m is $1,000$. In sememe-based word representation, the number of shared sememes for synonyms in the word relation-based approach is 2. In the training of matching process, we set ϵ to 0.5 empirically. When predicting sememes for words in target language, we only consider 100 most similar source words for each target word and the attenuation parameter c is 0.8. The testing set for cross-lingual lexical sememe prediction contains $2,000$ randomly selected English words from the vocabulary.

4.3 Cross-lingual Lexical Sememe Prediction

We evaluate our model by recommending sememes for English words. In HowNet, many words have multiple sememes, so that sememe prediction can be regarded as a multi-label classification task. We use mean average precision (MAP) and F$_1$ score to evaluate the sememe prediction results.

We compare our model that incorporates sememe information with word relation-based approach (named CLSP-WR) and our model which jointly trains word and sememe embeddings (named CLSP-SE) with a baseline method BiLex (Zhang et al., 2017), a bilingual WRL model without incorporation of sememe information. For BiLex, we use its trained bilingual word embeddings to predict sememes for the words in target language with our sememe prediction approach.

Table 1 exhibits the evaluation results of cross-lingual lexical sememe prediction with different

[1] Sogou-T is a corpus of web pages provided by a Chinese commercial search engine. https://www.sogou.com/labs/resource/t.php

[2] https://dumps.wikimedia.org/

[3] http://thulac.thunlp.org/

[4] https://cloud.google.com/translate/

[5] https://catalog.ldc.upenn.edu/LDC2002L27

Method	Seed Lexicon	Sememe Prediction	
		MAP	F$_1$ Score
BiLex	1000	27.57	16.08
	2000	33.79	22.33
	4000	35.78	25.74
	6000	38.29	28.71
CLSP-WR	1000	28.12	18.55
	2000	33.78	23.64
	4000	38.30	27.74
	6000	41.23	30.64
CLSP-SE	1000	31.78	18.22
	2000	37.70	24.31
	4000	40.77	29.33
	6000	**43.16**	**32.49**

Table 1: Evaluation results of cross-lingual lexical sememe prediction with different seed lexicon sizes.

seed lexicon sizes in {1000, 2000, 4000, 6000[6]}. From the table, we can clearly see that:

(1) Our two models perform much better compared with BiLex in all the seed lexicon size settings. It indicates that incorporating sememe information into word embeddings can effectively improve the performance of predicting sememes for target words. The reason is that both of our models make words with similar sememe annotations have similar embeddings, and as a result, we can recommend better sememes for target words according to its related source words.

(2) CLSP-SE model achieves better results than CLSP-WR model. The reason is that by representing sememes in a latent semantic space, CLSP-SE model can further capture the relatedness between sememes as well as the relatedness between words and sememes, which is helpful for modeling the representations of those words with similar sememes.

4.4 Case Study

In case study, we conduct qualitative analysis to explain the effectiveness of our models with detailed cases. We show two examples of cross-lingual word sememe prediction, in which we predict sememes for *handcuffs* and *canoeist*. Fig. 2 shows the embeddings of five closest Chinese and English words to *handcuffs* and *canoeist*, and the vector of each word is projected down to two dimensions using t-SNE (Maaten and Hinton, 2008).

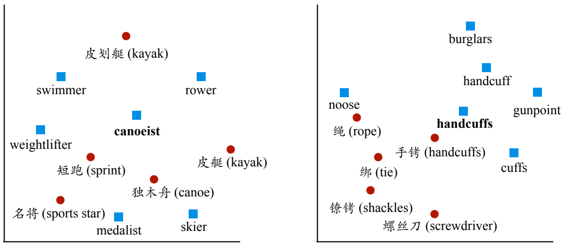

Figure 2: Two examples of nearest English and Chinese words.

Table 2 lists top-5 sememes we predict for the two words and the sememes annotated for each word in HowNet are in boldface. In the table, we also exhibit the annotated sememes of the five closest Chinese words.

In the first example, our model finds the best translated word for *handcuffs* in Chinese 手铐 "handcuffs", whose sememe annotations are exactly the same as those of *handcuffs*. In addition, the second closest Chinese word 镣铐 "shackles" is a synonym for 手铐 "handcuffs" and also has the same sememe annotations. Therefore, our model predicts all the correct sememes successfully. From the prediction results of this example, we notice that our model can accurately predict general sememes like 用具 "tool" and 人 "human", which are supposed to be difficult to predict.

In the second example, accurate Chinese translated counterpart for *canoeist* does not exist, but our model still hits all the three annotated sememes in the top-5 predicted sememes. By observing the most similar Chinese words, we can find that although these words do not have the same meaning as *canoeist*, they are related to *canoeist* in different aspects. For example, 短跑 "sprint" and *canoeist* are both in the sports domain so that they share the sememes 锻炼 "exercise" and 体育 "sport". 名将 "sports star" has the meaning of sports star and it can provide the sememe 人 "human" in sememe prediction. Furthermore, it is noteworthy that our model predicts 船 "ship" due to the nearest Chinese words 独木舟 "canoe" and 皮艇 "kayak", whereas 船 "ship" is not annotated for *canoeist* in HowNet. It is obvious that 船 "ship" is an appropriate sememe for *canoeist*. Since HowNet is manually annotated by experts, misannotated words always exist inevitably, which in some cases underestimates our models.

[6] The largest seed lexicon size is 6000 because that is the maximum number of translation word pairs that we can obtain from the bilingual corpora.

Type	Words	Sememes
English Word	handcuffs	**用具 "tool"**, **警 "police"**, **扣住 "detain"**, **人 "human"**, **有罪 "guilty"**
5 Nearest Chinese Words	手铐 "handcuffs" 镣铐 "shackles" 绑 "tie" 螺丝刀 "screwdriver" 绳 "rope"	**有罪 "guilty"**, **警 "police"**, **人 "human"**, **扣住 "detain"**, **用具 "tool"** **有罪 "guilty"**, **警 "police"**, **人 "human"**, **扣住 "detain"**, **用具 "tool"** 包扎 "wrap" **用具 "tool"**, 放松 "loosen", 勒紧 "tighten" 线 "linear", 材料 "material", 拴连 "fasten"
English Word	canoeist	**锻炼 "exercise"**, **人 "human"**, **体育 "sport"**, 事情 "fact", 船 "ship"
5 Nearest Chinese Words	短跑 "sprint" 独木舟 "canoe" 皮艇 "kayak" 名将 "sports star" 皮划艇 "kayak"	事情 "fact" **锻炼 "exercise"** **体育 "sport"** 船 "ship" 船 "ship" 著名 "famous", **人 "human"**, 官 "official", 军 "military" 事情 "fact", **锻炼 "exercise"**, **体育 "sport"**

Table 2: Two examples of cross-lingual lexical sememe prediction.

4.5 Effect of Word Frequency

To explore how frequencies of target words affect cross-lingual sememe prediction results, we split the testing set into four subsets according to word frequency and then calculate the sememe prediction MAP and F_1 score for each subset. The results are shown in Table 3.

Method	Word Frequency	Sememe Prediction	
		MAP	F_1 Score
BiLex	<200	30.35	21.83
	200 - 500	34.83	25.95
	501 - 2500	40.21	28.62
	>2500	47.56	35.80
CLSP-WR	<200	34.73	24.41
	200 - 500	39.50	29.49
	501 - 2500	43.92	33.87
	>2500	47.33	34.99
CLSP-SE	<200	36.54	27.49
	200 - 500	41.46	30.09
	501 - 2500	45.35	35.01
	>2500	**49.34**	**37.16**

Table 3: Evaluation results of cross-lingual lexical sememe prediction with different word frequencies. The number of words in each frequency range is 497, 458, 522 and 523 respectively.

From the table we can see that: (1) The more frequently a target word appears in the corpus, the better its predicted sememes are. It is because high-frequency words normally have better word embeddings, which are crucial to sememe prediction. (2) Our models evidently perform better than BiLex in different word frequencies, especially in low frequency. It indicates that by considering external information of HowNet, our models are more robust and can competently handle sparse scenarios.

4.6 Further Quantitative Analysis

In this section, we conduct two typical auxiliary experiments to further analyze the superiority of our models quantitatively.

Bilingual Lexicon Induction

Our models learn bilingual word embeddings in one unified semantic space. Here we use translation top-1 and top-5 average precision (P@1 and P@5) to evaluate bilingual lexicon induction performance of our models and BiLex. The seed lexicon size also varies in {1000, 2000, 4000, 6000}.

Method	Seed Lexicon	Lexicon Induction	
		P@1	P@5
BiLex	1000	6.48	10.78
	2000	10.84	15.84
	4000	19.48	23.96
	6000	25.89	29.59
CLSP-WR	1000	6.89	11.28
	2000	11.96	18.08
	4000	19.50	25.78
	6000	25.83	31.03
CLSP-SE	1000	6.60	11.04
	2000	11.90	18.62
	4000	19.26	25.11
	6000	**26.91**	**32.17**

Table 4: Bilingual lexicon induction performance with different seed lexicon sizes.

The results are shown in Table 4. From this table, we observe that our models, especially CLSP-SE model, enhance the performance of word translation compared to BiLex no matter how large the seed lexicon is. It indicates that our models can bind bilingual word embeddings better.

Word Similarity Computation

We also evaluate the task of monolingual word similarity computation on WordSim-240 (WS-240) and WordSim-297 (WS-297) datasets for Chinese, and WordSim-353 (WS-353) and SimLex-999 (SL-999) datasets for English.

Method	Chinese (source)		English (target)	
	WS-240	WS-297	WS-353	SL-999
BiLex	60.36	62.17	60.46	27.22
CLSP-WR	61.27	65.25	60.46	27.22
CLSP-SE	60.84	65.62	62.47	28.79

Table 5: Performance on monolingual word similarity computation with seed lexicon size 6000.

Table 5 shows the results of monolingual word similarity computation on four datasets. From the table, we find that: (1) Our models perform better than BiLex on both Chinese word similarity datasets. It signifies incorporating sememe information helps learn better monolingual embeddings; (2) CLSP-WR model does not enhance English word similarity results but CLSP-SE model does. It is because CLSP-WR model only post-processes Chinese word embeddings and keeps English word embeddings unchanged while CLSP-SE model undertakes bilingual alignment and sememe information incorporation together, which makes English word embeddings improve with Chinese word embeddings.

5 Conclusion and Future Work

In this paper, we introduce a new task of cross-lingual sememe prediction. This task is very important because the construction of sememe-based linguistic knowledge bases in various languages is beneficial to better understanding these languages. We propose a simple and effective model for this task, including monolingual word representation learning, cross-lingual word representation alignment and sememe-based word representation learning. Experimental results on real-world datasets show that our model achieves consistent and significant improvements compared to baseline method in cross-lingual sememe prediction.

In the future, we will explore the following research directions: (1) In this paper, for simplification, we ignore the rich hierarchy information in HowNet and also ignore the fact that a word may have multiple senses. We will extend our models to consider the structure information of sememe and multiple senses of words; (2) In fact, our framework for cross-lingual lexical sememe prediction can be transferred to other cross-lingual tasks. We will explore the effectiveness of our model in these tasks such as cross-lingual information retrieval.

Acknowledgments

This research is funded by the National 973 project (No. 2014CB340501). It is also partially supported by the NExT++ project, the National Research Foundation, Prime Minister’s Office, Singapore under its IRC@Singapore Funding Initiative. Hao Zhu is supported by Tsinghua University Initiative Scientific Research Program. We also thank the anonymous reviewers for their valuable comments and suggestions.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of NIPS*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*.

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

Danushka Bollegala, Mohammed Alsuhaibani, Takanori Maehara, and Ken-ichi Kawarabayashi. 2016. Joint word representation learning using a corpus and a semantic lexicon. In *Proceedings of AAAI*.

Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of EMNLP*.

Lei Dang and Lei Zhang. 2010. Method of discriminant for chinese sentence sentiment orientation based on hownet. *Application Research of Computers*, 4:43.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Zhendong Dong and Qiang Dong. 2003. Hownet-a hybrid language and knowledge resource. In *Proceedings of NLP-KE*.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of EMNLP*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the EACL*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Xianghua Fu, Guo Liu, Yanyan Guo, and Zhiqiang Wang. 2013. Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon. *Knowledge-Based Systems*, 37:186–195.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: fast bilingual distributed representations without word alignments. In *Proceedings of ICML*.

Yihong Gu, Jun Yan, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. 2018. Language modeling with sparse product of sememe experts. In *Proceedings of EMNLP*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual distributed representations without word alignment. In *Proceedings of ICLR*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. 2018. Incorporating chinese characters of words for lexical sememe prediction. In *Proceedings of ACL*.

Peng Jin and Yunfang Wu. 2012. SemEval-2012 Task 4: Evaluating chinese word similarity. In *Proceddings of *SEM*.

Tomáš Kočiskỳ, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proceedings of ACL*.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of ACL-IJCNLP*.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics*, 35(4):505–512.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of ACL-IJCNLP*.

Qun Liu and Sujian Li. 2002. Word similarity computing based on hownet. *International Journal of Computational Linguistics & Chinese Language Processing*, 7(2):59–76.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of NAANL-HLT*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*.

Laurens van der Maaten and Geoffrey E Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

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

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

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

Nikola Mrkšic, Diarmuid OSéaghdha, Blaise Thomson, Milica Gašic, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*.

Yilin Niu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Improved word representation learning with sememes. In *Proceedings of ACL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of ACL-IJCNLP*.

Jingguang Sun, Dongfeng Cai, Dexin Lv, and Yanju Dong. 2007. Hownet based chinese question automatic classification. *Journal of Chinese Information Processing*, 21(1):90–95.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of ACL*.

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of ACL*.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of ACL-IJCNLP*.

Michael Wick, Pallika Kanani, and Adam Craig Pocock. 2016. Minimally-constrained multilingual embeddings via artificial code-switching. In *Proceedings of AAAI*.

Ruobing Xie, Xingchi Yuan, Zhiyuan Liu, and Maosong Sun. 2017. Lexical sememe prediction via word embeddings and matrix factorization. In *Proceedings of AAAI*.

Xiangkai Zeng, Cheng Yang, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Chinese liwc lexicon expansion via hierarchical classification of word embeddings with sememe attention. In *Proceedings of AAAI*.

Meng Zhang, Haoruo Peng, Yang Liu, Huan-Bo Luan, and Maosong Sun. 2017. Bilingual lexicon induction from non-parallel data with minimal supervision. In *Proceedings of AAAI*.

Yuntao Zhang, Ling Gong, and Yongcheng Wang. 2005. Chinese word sense disambiguation using hownet. In *Proceedings of International Conference on Natural Computation*.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*.