Towards Building a Multilingual Sememe Knowledge Base: Predicting Sememes for BabelNet Synsets

Fanchao Qi,[1*] **Liang Chang,**[2*†] **Maosong Sun,**[1,3‡] **Sicong Ouyang,**[2†] **Zhiyuan Liu**[1]

[1]Department of Computer Science and Technology, Tsinghua University
Institute for Artificial Intelligence, Tsinghua University
Beijing National Research Center for Information Science and Technology
[2]Beijing University of Posts and Telecommunications
[3]Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University
qfc17@mails.tsinghua.edu.cn, changliang@bupt.edu.cn
{sms, liuzy}@tsinghua.edu.cn, scouyang4354@gmail.com

Abstract

A sememe is defined as the minimum semantic unit of human languages. Sememe knowledge bases (KBs), which contain words annotated with sememes, have been successfully applied to many NLP tasks. However, existing sememe KBs are built on only a few languages, which hinders their widespread utilization. To address the issue, we propose to build a unified sememe KB for multiple languages based on BabelNet, a multilingual encyclopedic dictionary. We first build a dataset serving as the seed of the multilingual sememe KB. It manually annotates sememes for over 15 thousand synsets (the entries of BabelNet). Then, we present a novel task of automatic sememe prediction for synsets, aiming to expand the seed dataset into a usable KB. We also propose two simple and effective models, which exploit different information of synsets. Finally, we conduct quantitative and qualitative analyses to explore important factors and difficulties in the task. All the source code and data of this work can be obtained on https://github.com/thunlp/BabelNet-Sememe-Prediction.

Introduction

A word is the smallest element of human languages that can stand by itself, but not the smallest indivisible semantic unit. In fact, the meaning of a word can be divided into smaller components. For example, one of the meanings of "man" can be represented as the composition of the meanings of "human", "male" and "adult". In linguistics, a *sememe* (Bloomfield 1926) is defined as the minimum semantic unit of human languages. Some linguists believe that meanings of all the words in any language can be decomposed of a limited set of predefined sememes, which is related to the idea of universal semantic primitives (Wierzbicka 1996).

Sememes are implicit in words. To utilize them in practical applications, people manually annotate words with predefined sememes to construct sememe knowledge bases (KBs). HowNet (Dong and Dong 2003) is the most famous one, which uses about $2,000$ language-independent
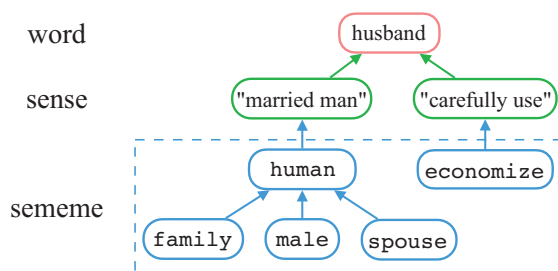

Figure 1: Sememe annotation of the word "husband" in HowNet.

sememes to annotate senses of over 100 thousand Chinese and English words. Figure 1 illustrates an example of how words are annotated with sememes in HowNet.

Different from most linguistic KBs like WordNet (Miller 1995), which explain meanings of words by word-level relations, sememe KBs like HowNet provide intensional definitions for words using infra-word sememes. Sememe KBs have two unique strengths. The first one is their sememe-to-word semantic compositionality, which endows them with special suitability for integration into neural networks (Gu et al. 2018; Qi et al. 2019a). The second one is their nature that limited sememes can represent unlimited meanings, which makes sememes very useful in low data regimes, e.g., improving embeddings of low-frequency words (Sun and Chen 2016; Niu et al. 2017). In fact, sememe KBs have been proven beneficial to various NLP tasks such as word sense disambiguation (Duan, Zhao, and Xu 2007) and sentiment analysis (Fu et al. 2013).

Most languages have no sememe KBs, which prevents NLP applications of these languages benefiting from sememe knowledge. However, building a sememe KB for a new language from scratch is time-consuming and labor-intensive — the construction of HowNet takes several linguistic experts more than two decades. To tackle this challenge, Qi et al. (2018) present the task of cross-lingual lexical sememe prediction (CLSP), aiming to facilitate the construction of a new language's sememe KB by predicting sememes for words in that language. However, CLSP can pre-

---

[*]Indicates equal contribution

[†]Work done during internship at Tsinghua University
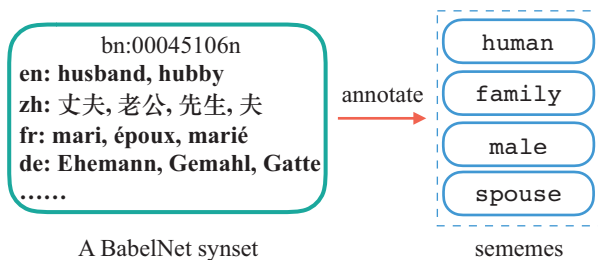
[‡]Corresponding author

Figure 2: Annotating sememes for the BabelNet synset whose ID is *bn:00045106n*. The synset comprises words in different languages (multilingual synonyms) having the same meaning "the man a woman is married to", and they share the four sememes on the right part.

dict sememes for only one language at a time, which means repetitive efforts, including manual correctness checking conducted by native speakers, are required when constructing sememe KBs for multiple languages.

To solve this problem, we propose to build a unified sememe KB for multiple languages, namely a multilingual sememe KB, based on BabelNet (Navigli and Ponzetto 2012a), which is a more economical and efficient way to transfer sememe knowledge to other languages. BabelNet is a multilingual encyclopedic dictionary and comprises over 15 million entries called *BabelNet synsets*. Each BabelNet synset contains words in multiple languages with the same meaning (multilingual synonyms), and they should have the same sememe annotation. Therefore, building a multilingual sememe KB by annotating sememes for BabelNet synsets can actually provide sememe annotation for words in multiple languages simultaneously (Figure 2 shows an example).

To advance the creation of a multilingual sememe KB, we build a seed dataset named **BabelSememe**, which contains about 15 thousand BabelNet synsets manually annotated with sememes. Also, we present a novel task of automatic sememe prediction for BabelNet synsets (SPBS), aiming to gradually expand the seed dataset into a usable multilingual sememe KB. In addition, we exploratively put forward two simple and effective models for SPBS, which utilize different information incorporated in BabelNet synsets. The first model exploits semantic information and recommends similar sememes to semantically close BabelNet synsets, while the second uses relational information and directly predicts relations between sememes and BabelNet synsets. In experiments, we evaluate sememe prediction performance of the two models on BabelSememe, finding they achieve satisfying results. Moreover, the ensemble of the two models yield obvious performance enhancement. Finally, we conduct detailed quantitative and qualitative analyses to explore the factors influencing sememe prediction results, aiming to point out the characteristics and difficulties of the SPBS task.

In conclusion, our contributions are threefold: (1) first proposing to construct a multilingual sememe KB based on BabelNet and presenting a novel task SPBS; (2) building the BabelSememe dataset containing over 15 thousand BabelNet synsets manually annotated with sememes; and (3)

proposing two simple and effective models and conducting detailed analyses of factors in SPBS.

Dataset and Task Formalization

BabelSememe Dataset

We build the BabelSememe dataset, which is expected to be the seed of a multilingual sememe KB and expanded steadily by automatic sememe prediction together with examination of humans. Sememe annotation in HowNet embodies hierarchical structures of sememes, as shown in Figure 1. Nevertheless, considering the structures of sememes are seldom used in existing applications (Niu et al. 2017; Qi et al. 2019a) and it is very hard for ordinary people to make structured sememe annotation, we ignore them in BabelSememe. Thus, each BabelNet synset in BabelSememe is annotated with a set of sememes (as shown in Figure 2). In the following part, we elaborate how we build the dataset and provide its statistics.

Selecting Target Synsets   We first select 20 thousand synsets as target synsets. Each of them includes English and Chinese synonyms annotated with sememes in HowNet,[1] so that we can generate candidate sememes for them.

Generating Candidate Sememes   We generate candidate sememes for each target synset using the sememes annotated to its synonyms. Some sememes of the synonyms in a target synset should be annotated to the target synset. For example, the word "husband" is annotated with five sememes in HowNet (Figure 1) in total, and the four sememes annotated to the sense "married man" should be annotated to the BabelNet synset *bn:00045106n* (Figure 2). Thus, we group the sememes of all the synonyms in a target synset together to form the candidate sememe set of the target synset.

Annotating Appropriate Sememes   We ask more than 100 annotation participants to select appropriate sememes from corresponding candidate sememe set for each target synset. All the participants have a good grasp of both Chinese and English. We show them Chinese and English synonyms as well as definitions of each synset, making sure its meaning is fully understood. When annotation is finished, each target synset has been annotated by at least three participants. We remove the synsets whose inter-annotator agreement (IAA) is poor, where we use Krippendorff's alpha coefficient (Krippendorff 2013) to measure IAA. Finally, 15, 756 BabelNet synsets are retained, which are annotated with 43, 154 sememes selected from 104, 938 candidate sememes, and their average Krippendorff's alpha coefficient is **0.702**.

Dataset Statistics   Detailed statistics of BabelNet synsets with different POS tags in BabelSememe are shown in Table 1. In addition, we show the distribution of BabelNet synsets over sememe numbers in Figure 3.

---

[1]We use OpenHowNet (Qi et al. 2019b), the open-source data accessing API of HowNet to obtain the sememes of a word.

POS Tag	noun	verb	adj	adv	total
#synset	10,417	2,290	2,507	542	15,756
average #sememe	2.95	2.49	2.29	1.78	2.74

Table 1: Statistics of BabelNet synsets with different POS tags in BabelSememe.
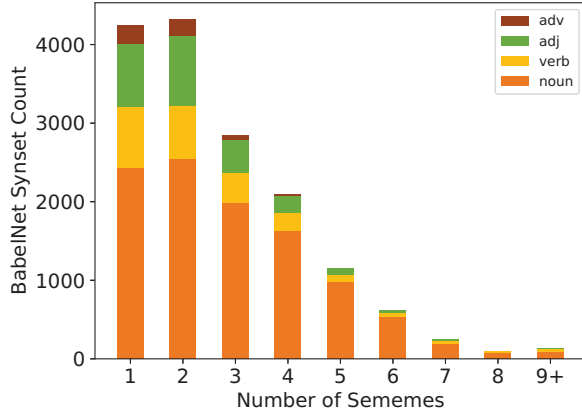

Figure 3: Distribution of BabelNet synsets over sememe numbers in BabelSememe.

SPBS Task Formalization

SPBS is aimed at predicting appropriate sememes for unannotated BabelNet synsets. It can be modeled as a multi-label classification problem, where sememes are regarded as labels to be attached to BabelNet synsets. Formally, we define B as the set of all the BabelNet synsets, and S as the set of all the sememes. For a given target BabelNet synset $b \in B$, we intend to predict its sememe set $S_b = \{s_1, \cdots, s_{|S_b|}\} \subset S$, where $|S_b|$ is the number of b's sememes.

Previous methods of sememe prediction for *words* usually compute an association score for each sememe and select the sememes with scores higher than a threshold to form the predicted sememe set (Xie et al. 2017; Jin et al. 2018). Following this formulation, we have

$$\hat{S}_b = \{s \in S | P(s|b) > \delta\}, \quad (1)$$

where $\hat{S}_b$ is the predicted sememe set of b, $P(s|b)$ is the association score of sememe s, and δ is the association score threshold. To compute the association score, existing methods of sememe prediction for *words* capitalize on semantic similarity between target words and sememe-annotated words or directly model the relations between target words and sememes (Xie et al. 2017; Jin et al. 2018). No matter which way to choose, representations of target words are of vital importance. Similarly, representations of BabelNet synsets are crucial to SPBS. In following section of Methodology, we make an preliminary attempt to utilize two kinds of representations of BabelNet synsets in the SPBS task.

Methodology

As mentioned in task formalization, learning representations of BabelNet synsets[2] is very important to SPBS. BabelNet merges various resources such as WordNet (Miller 1995) and Wikipedia, which provide abundant information for learning synset representations. Correspondingly, we summarize two kinds of synset representations according to the information used for representation learning: (1) **semantic representation**, which bears the meaning of a synset. Much information can be used for learning the semantic representation of a synset, e.g., textual definitions from WordNet and related Wikipedia articles; (2) **relational representation**, which captures the relations between different synsets. Most relations are the semantic relations transferred from WordNet (e.g., "antonym").

Next, we introduce two preliminary models, namely SPBS-SR and SPBS-RR, which utilize semantic and relational representations respectively to predict sememes for synsets. We also present an ensemble model which combines the two models' prediction results to obtain better performance.

SPBS-SR Model

Inspired by Xie et al. (2017), the idea of SPBS-SR (SPBS with Semantic Representations) is to compute the association score $P(s|b)$ by measuring the similarity between the target synset b and the other synsets annotated with the target sememe s. In other words, if a synset with known sememe annotation is very similar to the target synset, its sememes should have high association scores. In fact, this idea is similar to collaborative filtering (Sarwar et al. 2001) in recommendation systems.

Formally, following the notations in task formalization, we can calculate $P(s|b)$ by

$$P(s|b) \sim \sum_{b' \in B'} \cos(\mathbf{b}, \mathbf{b}') \cdot I_{S_{b'}}(s) \cdot c^{r_{b'}}, \quad (2)$$

where B' is the set of synsets with known sememe annotation, $\mathbf{b}$ and $\mathbf{b}'$ represent the semantic representations of b and b' respectively, $I_{S_{b'}}(s)$ is a indicator function indicating whether s is in $S_{b'}$, $r_{b'}$ is the descending rank of $\cos(\mathbf{b}, \mathbf{b}')$ and $c \in (0, 1)$ is a hyper-parameter. Here $c^{r_{b'}}$ is a declined confidence factor used to diminish the influence of irrelevant words on predicting sememes. The irrelevant words may have totally different sememes which are actually noises.

We choose the embedded vector representations of synsets from NASARI (Camacho-Collados, Pilehvar, and Navigli 2016) as required semantic representations. NASARI utilizes the content of Wikipedia pages linked to synsets to learn vector representations, and only nominal synsets have NASARI representations because non-nominal synsets have no linked Wikipedia pages. Correspondingly, SPBS-SR can be used in sememe prediction for nominal synsets only.

---

[2]From now on, we use "synset" as the abbreviation of "BabelNet synset" for succinctness.
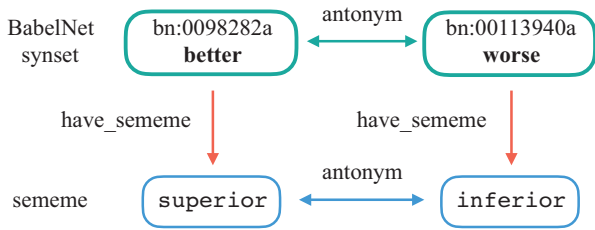
Figure 4: An example of how relations between BabelNet synsets are consistent with the relations between respective sememes. Notice that we only show the English synonyms in the BabelNet synsets.

SPBS-RR Model

SPBS-RR (SPBS with Relational Representations) aims to use relational representations of target synsets to compute $P(s|b)$. As mentioned before, there are many relations between synsets, and most of them are semantic relations from WordNet. As for sememes, they also have four semantic relations in HowNet, namely "hypernym", "hyponym", "antonym" and "converse". The semantic relation between a pair of synsets should be consistent with the relation between their respective sememes. Taking Figure 4 as an example, the synset "better" is the antonym of "worse", and their respective sememes superior and inferior are also a pair of antonyms. Naturally, this property can be used to predict sememes when synsets and sememes as well as all their relations are considered together — if we know "better"-"worse" and superior-inferior are both antonym pairs, and "better" has the sememe superior, then we should undoubtedly predict inferior for "worse".

To this end, we introduce an artificial relation "have_sememe" between a synset and any of its sememes, aiming to build a bridge between synsets and sememes. Now by considering all the synset-synset, sememe-sememe and synset-sememe relations, synsets and sememes form a semantic graph, and as a result, knowledge graph embedding methods can be used to learn relational representations of synsets.

Here we borrow the translation idea from the well-established TransE model (Bordes et al. 2013). Formally, for the above-mentioned semantic graph, all its relation triplets form a set G. Each triplet in G can be represented as (h, r, t), where $h, t \in S \cup B$ are nodes and $r \in R_B \cup R_S \cup \{r_h\}$ is a relation. R_B and R_S are the sets of relations between synsets and between sememes respectively, and r_h refers to the "have_sememe" relation. Then we can learn representations of both nodes and relations by minimizing:

$$\mathcal{L}_1 = \sum_{(h,r,t)\in G} [\tau + d(\mathbf{h} + \mathbf{r}, \mathbf{t}) - d(\mathbf{h} + \mathbf{r}, \mathbf{t}')]_+, \quad (3)$$

where $[x]_+ = \max(0, x)$, scalar τ is a hyper-parameter, $(h, r, t'), t' \in S \cup B$ is a corrupted triplet, boldface symbols represent corresponding vectors, and $d(\mathbf{x}, \mathbf{y})$ is L_2 distance function:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2. \quad (4)$$

However, the synset-sememe semantic graph is not exactly the same as general knowledge graphs, because it embraces two kinds of nodes, namely synsets and sememes. Furthermore, according to the definition of sememes, the meaning of a synset should be equal to the sum of its sememes' meanings. In other words, there exists a special semantic equivalence relation between a synset and all its sememes. We formalize this relation and design a corresponding constraint loss:

$$\mathcal{L}_2 = \sum_{b\in B} \|\mathbf{b} + \mathbf{r}_s - \sum_{s\in S_b} \mathbf{s}\|^2, \quad (5)$$

where r_s denotes the semantic equivalence relation. Therefore, the overall training loss is as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2, \quad (6)$$

where λ_1 and λ_2 are hyper-parameters controlling relative weights of the two losses. By optimizing the loss function, we can obtain relational representations of synsets, sememes and relations. Sememe prediction for the target synset b can be regarded as finding the tail node of the incomplete triplet $(b, r_h, ?)$. Hence, the association score $P(s|b)$ can be computed by:

$$P(s|b) \sim d(\mathbf{b} + \mathbf{r}_h, \mathbf{s}). \quad (7)$$

Ensemble Model

SPBS-SR depends on semantic representations of synsets while SPBS-RR utilizes relational representations. It is obvious that the two models employ different kinds of information and exhibit different features of a synset. Accordingly, combining the two models together is expected to improve sememe prediction performance. Considering sememe association scores yielded by the two models are not comparable, we redefine the sememe prediction score by making use of the reciprocal sememe ranks:

$$P(s|b) = \lambda_c \frac{1}{\text{rank}_s^c} + \lambda_r \frac{1}{\text{rank}_s^r}, \quad (8)$$

where rank_s^c and rank_s^r are descending ranks of sememe s according to the association scores computed by SPBS-SR and SPBS-RR respectively, and λ_c and λ_r are hyper-parameters controlling relative weights of the two items.

Experiments

In this section, we evaluate our two SPBS models. Furthermore, we conduct both quantitative and qualitative analyses to investigate the factors in SPBS results, aiming to reveal characteristics and difficulties of the SPBS task.

Dataset

We use BabelSememe as the source of sememe annotation for synsets. We extract all the relations between the synsets in BabelSememe from BabelNet. In addition, there are four semantic relations between sememes, and the "have_sememe" relation between a synset and any of its sememes. Since our SPBS-RR model is graph-based, follow-

Model	noun		verb		adj		adv		avg.	
	MAP	F1	MAP	F1	MAP	F1	MAP	F1	MAP	F1
LR	54.4	40.0	—	—	—	—	—	—	—	—
TransE	60.2	46.6	33.2	24.6	31.1	24.3	30.0	**21.3**	51.3	39.5
SPBS-SR	65.0	50.0	—	—	—	—	—	—	—	—
SPBS-RR	62.5	47.9	**34.8**	**25.3**	**32.7**	**24.5**	**30.9**	20.0	53.3	40.5
Ensemble	**69.0**	**55.4**	**34.8**	**25.3**	**32.7**	**24.5**	**30.9**	20.0	**57.6**	**45.6**

Table 2: Overall and POS tag-specific SPBS results of all the models on the test set.

ing previous knowledge graph embedding work, we filter the low-frequency synsets, sememes, and relations out.[3]

The final dataset we use contains $15,461$ synsets, $2,106$ sememes, 196 synset-synset relations, 4 sememe-sememe relations and 1 synset-sememe relation ("have_sememe"). And there are $171,147$ triplets in total, including $125,114$ synset-synset, $3,317$ sememe-sememe and $42,716$ synset-sememe triplets. We randomly divide the synsets into three subsets in the ratio of 8:1:1. Since only the tail nodes of synset-sememe triplets need to be predicted in the SPBS-RR model, we select all the synset-sememe triplets comprising the synsets in the two 10% subsets to form the validation and test sets. All the other triplets compose the training set.

Notice that our first model SPBS-SR only works on nominal synsets and needs no synset-synset or sememe-sememe triplets, which means only the synset-sememe triplets comprising nominal synsets in the training set are utilized and only nominal synsets have sememe prediction results.

Experimental Settings

Baseline Methods SPBS is a brand new task, and there are no previous methods specifically designed for it. Hence, we simply choose logistic regression (LR) and TransE[4] as baseline methods. Similar to SPBS-SR, LR also takes NASARI embeddings of synsets (semantic representations) as input and only works on nominal synsets. TransE learns relational representations of synsets and differs from SPBS-RR in the constraint of semantic equivalence relation.

Hyper-parameters and Training For both SPBS-SR and LR, the dimension of used NASARI synset embeddings is 300. For SPBS-SR, c in Equation (2) is 0.8. For both TransE and SPBS-RR, the embedding dimension of synsets, sememes and relations is empirically set to 800, and the margin τ in Equation (3) is set to 4. For SPBS-RR, the relative weight $\lambda_1 = 0.95$ and $\lambda_2 = 0.05$. For the ensemble model, $\lambda_c = 0.45$ and $\lambda_r = 0.55$. We adopt SGD as the optimizer whose learning rate is set to 0.01. All these hyper-parameters have been tuned to the best on the validation test.

---

[3]Although our SPBS-SR model can handle the low-frequency synsets well, we use the same filtered dataset to evaluate both models for fair comparison.

[4]Numerous knowledge graph embedding methods have been proposed recently, but we find that TransE performs substantially better than all other popular models on this task by experiment.

POS Tag	noun	verb	adj	adv
#synset	10,360	2,240	2,419	442
#triplet	210,127	20,657	23,490	4,952
avg. #triplet	20.28	9.22	9.71	11.20

Table 3: Numbers of POS tag-specific synsets and triplets and their average triplet numbers.

Evaluation Metrics Following previous sememe prediction work, we choose mean average precision (MAP) and the F1 score as evaluation metrics. And the sememe association score threshold, i.e., δ in Equation (1), is set to 0.32.

Overall SPBS Results

The overall and POS tag-specific SPBS results of our models as well as baseline methods on the test set are shown in Table 2. Note that the ensemble model has the same results as SPBS-RR on non-nominal synsets because SPBS-SR works on nominal synsets only. From the table, we can see that:

(1) SPBS-RR performs markedly better than TransE on synsets with whichever POS tag. This can prove the effectiveness of the constraint of semantic equivalence relation, which we propose specifically for the SPBS task by taking advantage of the nature of sememes.

(2) On the nominal synsets, SPBS-SR achieves the best performance among all of the four single models, which demonstrates that recommending identical sememes to meaning-similar synsets is effectual. In addition, the ensemble model produces substantial performance improvement as compared to its two submodels, which manifests the success of our ensemble strategy.

(3) SPBS-RR yields much better results on nominal synsets than non-nominal synsets. To explore the reason, we count the number of synsets with different POS tags as well as the numbers of triplets comprising POS tag-specific synsets and calculate their average triplet numbers. The statistics are listed in Table 3. We find that the number of nominal synsets and their average triplet number are significantly bigger than those of the non-nominal synsets. Consequently, less relational information of non-nominal synsets is captured, and it is hard to learn good relational representations for them, which explains their bad performance.

model	MAP	F1
LR	59.5	45.3
TransE	65.2	50.4
SPBS-SR	65.0	50.0
SPBS-RR	66.0	51.0
Ensemble	**70.3**	**56.7**

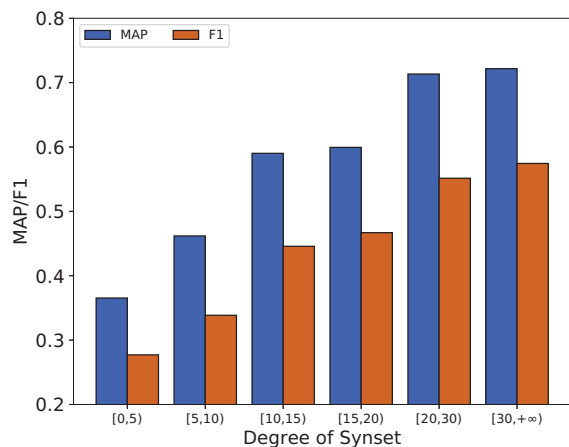Table 4: Sememe prediction results of all the models on nominal synsets.

Figure 5: SPBS results of synsets within different degree ranges. The numbers of synsets in the six ranges are 72, 340, 231, 110, 84 and 131 respectively.

SPBS for Nominal Synsets

According to the statistics in Table 3, we speculate those non-nominal synsets have a negative influence on sememe prediction for nominal synsets. To prove this, we remove all the non-nominal synsets as well as related triplets from the dataset, and then re-evaluate all the models on nominal synsets. The results are shown in Table 4.

We observe that both TransE and SPBS-RR receive considerable performance boost in predicting sememes for nominal synsets, and they even produce better results than SPBS-SR. In addition, the ensemble model performs correspondingly better. These results confirm our previous conjecture, and point out that we should notice the effect of other low-frequency synsets on the target synsets in SPBS.

Effect of the Synset's Degree

In this subsection, we investigate the effect of the synset's degree on SPBS results, where the degree of a synset is the number of the triplets comprising the synset. This experiment as well as following ones is conducted on the nominal synsets of the test set using the ensemble model.

Figure 5 exhibits the sememe prediction results of synsets within different degree ranges. We find that the degree of a synset has great impact on its sememe prediction results. It is easier to predict sememes for the synsets with larger degrees because more relation information of these synsets is captured and better relational representations are learned.
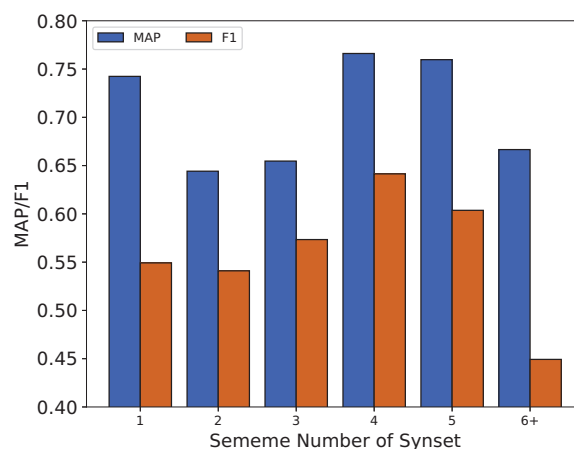

Figure 6: SPBS results of synsets whose sememe numbers are in different ranges. The numbers of synsets in the six ranges are 218, 239, 179, 179, 88 and 65 respectively.

Effect of the Synset's Sememe Number

In this subsection, we explore whether the number of a synset's annotated sememes affects SPBS results. Figure 6 exhibits the sememe prediction results of synsets whose sememe numbers are in different ranges. We find that sememe prediction performance increases first and then decreases with sememe numbers basically, which shows the synsets with too few or too many sememes are hard to cope with. The anomaly that MAP is high for the single-sememe synsets is because of the characteristic of MAP.[5]

Effect of the Sememe's Degree

To investigate what sememes are easy or hard to predict, we first focus on the degree of a sememe. Similar to the degree of a synset, the degree of a sememe is the number of the triplets comprising the sememe, and it is the only quantitative feature of sememes. Figure 7 shows the experimental results, where sememe degree is on the x-axis and the average prediction performance of the synsets having sememes within corresponding degree ranges is on the y-axis. We can observe that it is harder to predict the sememes with lower degrees, whose reason is similar to that of low-degree synsets' bad performance, i.e., low-degree nodes in graph normally have poor representations.

Effect of Sememe's Qualitative Features

In this subsection, we intend to observe some typical sememes to find qualitative features of sememes influencing SPBS results. Table 5 lists top 10 easiest and top 10 hardest sememes to predict as well as the average sememe prediction results of the synsets having them. We find most of the easiest sememes are concrete and normally annotated to the tangible entities. For example, the easiest sememe `capital` is always annotated to the synsets of capital cities

---

[5]A single-sememe synset's sememe prediction MAP would be 1 as long as the correct sememe is ranked first, no matter how many sememes are selected as predicted results.
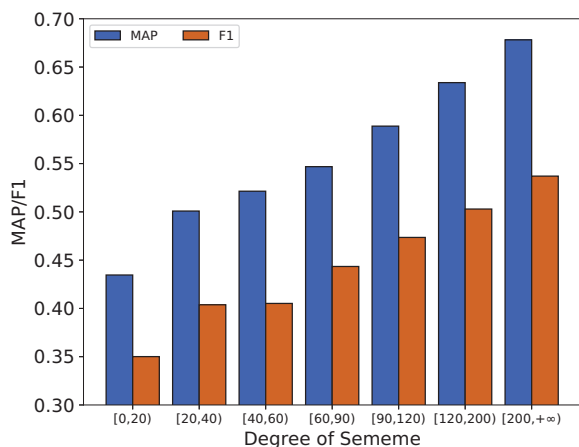
Figure 7: Average SPBS results of the synsets having sememes with degrees within different ranges. The numbers of sememes in the seven ranges are 1186, 235, 68, 47, 32, 26 and 28 respectively.

10 Easiest Sememes		10 Hardest Sememes	
Sememe	MAP/F1	Sememe	MAP/F1
capital	96.4/81.3	shape	13.3/5.0
metal	95.1/81.1	document	32.7/37.8
chemistry	92.8/77.8	expression	36.5/30.1
city	92.1/79.3	artifact	37.6/36.9
physics	89.6/72.6	protect	37.9/31.6
provincial	89.4/75.5	animate	38.6/27.8
PutOn	87.8/56.9	route	40.4/32.3
place	87.3/73.4	implement	43.8/45.1
ProperName	86.6/71.5	kind	45.7/30.6
country	85.7/63.3	own	47.3/56.0

Table 5: Top 10 sememes which are easiest to predict and top 10 sememes which are hardest to predict, and their corresponding sememe prediction results.

like "Beijing". As for the hardest sememes, they are more abstract, e.g., expression and protect, and usually annotated to the intangible concepts or non-nominal synsets. Therefore, we speculate that the concreteness of sememes is also an important factor in SPBS results.

Related Work

HowNet As the most well-known sememe KB, HowNet has attracted considerable research attention. Most related work employs HowNet in specific NLP tasks (Liu and Li 2002; Duan, Zhao, and Xu 2007; Fu et al. 2013; Niu et al. 2017; Gu et al. 2018; Qi et al. 2019a; Zang et al. 2019; Qin et al. 2019; Zhang et al. 2020). Some work tries to expand HowNet by predicting sememes for new words (Xie et al. 2017; Jin et al. 2018). To the best of our knowledge, only Qi et al. (2018) make an attempt to build a sememe KB for another language by cross-lingual lexical sememe prediction (CLSP). They learn bilingual word embeddings in a unified semantic space, and then predict sememes for target words according to their meaning-similar words in the sememe-annotated language. However, CLSP can predict sememes for only one language at one time and cannot work on low-resource languages or polysemous words.

BabelNet BabelNet (Navigli and Ponzetto 2012a) is a multilingual encyclopedic dictionary which amalgamates WordNet (Miller 1995) with Wikipedia as well as many other KBs, such as Wikidata (Vrandečić and Krötzsch 2014) and FrameNet (Baker, Fillmore, and Lowe 1998). It has been successfully utilized in all kinds of tasks (Moro and Navigli 2013; Iacobacci, Pilehvar, and Navigli 2015; De Gemmis et al. 2015), especially the cross-lingual or multilingual tasks (Navigli and Ponzetto 2012b; Vyas and Carpuat 2016). BabelNet has many advantages in terms of serving as the base of a multilingual sememe KB including (1) covering all the commonly used languages (284 languages); (2) incorporat-

ing polysemous words into multiple BabelNet synsets which enables sememe annotation for senses of polysemous words; and (3) amalgamating various resources, including dictionary definitions, semantic relations from WordNet and the content of Wikipedia pages, all of which can assist sememe annotation.

Knowledge Graph Embedding There are innumerable methods of knowledge graph embedding (KGE) towards knowledge base completion or link prediction (Wang et al. 2017). For example, translational distance models (Bordes et al. 2013; Lin et al. 2015; Xiao, Huang, and Zhu 2016), semantic matching models (Nickel, Tresp, and Kriegel 2011; Yang et al. 2015; Trouillon et al. 2016) and neural network models (Dettmers et al. 2018; Nguyen et al. 2018). However, none of these models consider heterogeneous knowledge graphs like the synset-sememe semantic graph. To our best knowledge, we are the first to model synsets and sememes in a semantic graph and propose a specifically modified KGE model for it.

Conclusion and Future Work

In this paper, we propose and formalize a novel task of Sememe Prediction for BabelNet Synsets (SPBS), which is aimed at facilitating the construction of a multilingual sememe KB based on BabelNet. We also build a dataset named BabelSememe which serves as the seed of the multilingual sememe KB. In addition, we preliminarily propose two simple and effective SPBS models. They utilize different kinds of information and can be combined to achieve better performance. Finally, we conduct quantitative and qualitative analyses, aiming to point out characteristics and difficulties of the SPBS task.

In the future, we will try to use more useful information of BabelNet synsets, e.g., WordNet definitions, in the SPBS task to improve performance. We also consider predicting the hierarchical structures of sememes for BabelNet sysnets. Moreover, we will conduct extrinsic evaluations of the predicted sememes when there are enough annotated synsets.

Acknowledgements

References

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The berkeley framenet project. In *Proceedings of COLING*.

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

Bordes, A.; Usunier, N.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-Relational Data. In *Proceedings of NIPS*.

Camacho-Collados, J.; Pilehvar, M. T.; and Navigli, R. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence* 240:36–64.

De Gemmis, M.; Lops, P.; Musto, C.; Narducci, F.; and Semeraro, G. 2015. Semantics-aware content-based recommender systems. In *Recommender Systems Handbook*. Springer. 119–159.

Dettmers, T.; Minervini, P.; Stenetorp, P.; and Riedel, S. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of AAAI*.

Dong, Z., and Dong, Q. 2003. HowNet-a hybrid language and knowledge resource. In *Proceedings of NLP-KE*.

Duan, X.; Zhao, J.; and Xu, B. 2007. Word sense disambiguation through sememe labeling. In *Proceedings of IJCAI*.

Fu, X.; Liu, G.; Guo, Y.; and Wang, Z. 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems* 37:186–195.

Gu, Y.; Yan, J.; Zhu, H.; Liu, Z.; Xie, R.; Sun, M.; Lin, F.; and Lin, L. 2018. Language modeling with sparse product of sememe experts. In *Proceedings of EMNLP*.

Iacobacci, I.; Pilehvar, M. T.; and Navigli, R. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of ACL*.

Jin, H.; Zhu, H.; Liu, Z.; Xie, R.; Sun, M.; Lin, F.; and Lin, L. 2018. Incorporating Chinese characters of words for lexical sememe prediction. In *Proceedings of ACL*.

Krippendorff, K. 2013. *Content Analysis: An Introduction to Its Methodology*. SAGE.

Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *Proceedings of AAAI*.

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

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

Moro, A., and Navigli, R. 2013. Integrating syntactic and semantic analysis into the open information extraction paradigm. In *Proceedings of IJCAI*.

Navigli, R., and Ponzetto, S. P. 2012a. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193:217–250.

Navigli, R., and Ponzetto, S. P. 2012b. Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of EMNLP*.

Nguyen, D. Q.; Nguyen, T. D.; Nguyen, D. Q.; and Phung, D. 2018. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. In *Proceedings of NAACL-HLT*.

Nickel, M.; Tresp, V.; and Kriegel, H.-P. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of ICML*.

Niu, Y.; Xie, R.; Liu, Z.; and Sun, M. 2017. Improved word representation learning with sememes. In *Proceedings of ACL*.

Qi, F.; Lin, Y.; Sun, M.; Zhu, H.; Xie, R.; and Liu, Z. 2018. Cross-lingual lexical sememe prediction. In *Proceedings of EMNLP*.

Qi, F.; Huang, J.; Yang, C.; Liu, Z.; Chen, X.; Liu, Q.; and Sun, M. 2019a. Modeling semantic compositionality with sememe knowledge. In *Proceedings of ACL*.

Qi, F.; Yang, C.; Liu, Z.; Dong, Q.; Sun, M.; and Dong, Z. 2019b. Openhownet: An open sememe-based lexical knowledge base. *arXiv preprint arXiv:1901.09957*.

Qin, Y.; Qi, F.; Ouyang, S.; Liu, Z.; Yang, C.; Wang, Y.; Liu, Q.; and Sun, M. 2019. Enhancing recurrent neural networks with sememes. *arXiv preprint arXiv:1910.08910*.

Sarwar, B. M.; Karypis, G.; Konstan, J. A.; Riedl, J.; et al. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of WWW*.

Sun, M., and Chen, X. 2016. Embedding for words and word senses based on human annotated knowledge base: Use HowNet as a case study. *Journal of Chinese information processing* 30(6).

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *Proceedings of ICML*.

Vrandečić, D., and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57(10):78–85.

Vyas, Y., and Carpuat, M. 2016. Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of NAACL-HLT*.

Wang, Q.; Mao, Z.; Wang, B.; and Guo, L. 2017. Knowledge graph embedding: A survey of approaches and applications. *TKDE* 29(12):2724–2743.

Wierzbicka, A. 1996. *Semantics: Primes and universals: Primes and universals*. Oxford University Press, UK.

Xiao, H.; Huang, M.; and Zhu, X. 2016. TransG: A generative model for knowledge graph embedding. In *Proceedings of ACL*.

Xie, R.; Yuan, X.; Liu, Z.; and Sun, M. 2017. Lexical sememe prediction via word embeddings and matrix factorization. In *Proceedings of AAAI*.

Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of ICLR*.

Zang, Y.; Yang, C.; Qi, F.; Liu, Z.; Zhang, M.; Liu, Q.; and Sun, M. 2019. Textual adversarial attack as combinatorial optimization. *arXiv preprint arXiv:1910.12196*.

Zhang, L.; Qi, F.; Liu, Z.; Wang, Y.; Liu, Q.; and Sun, M. 2020. Multi-channel reverse dictionary model. In *Proceedings of AAAI*.