Bridging the Gap Between BabelNet and HowNet: Unsupervised Sense Alignment and Sememe Prediction

Xiang Zhang, Ning Shi, Bradley Hauer, Grzegorz Kondrak
Alberta Machine Intelligence Institute (Amii), Department of Computing Science
University of Alberta, Edmonton, Canada
{xzhang23,ning.shi,bmhauer,gkondrak}@ualberta.ca

Abstract

As the minimum semantic units of natural languages, sememes can provide interpretable representations of concepts. Despite the widespread utilization of lexical resources for semantic tasks, the use of sememes is limited by a lack of available sememe knowledge bases. Recent efforts have been made to connect BabelNet with HowNet by automating sememe prediction. However, these methods depend on large manually annotated datasets. Instead, we propose to use sense alignment via a novel unsupervised and explainable method. Our method consists of four stages, each relaxing predefined constraints until a complete alignment of BabelNet synsets to HowNet senses is achieved. Experimental results demonstrate the superiority of our unsupervised method over previous supervised ones by an improvement of 12% overall F1 score, setting a new state of the art. Our work is grounded in an interpretable propagation of sememe information between lexical resources, and may benefit downstream applications which can incorporate sememe information.

1 Introduction

Sememes are the minimum semantic units of human languages (Bloomfield, 1926). The theory of semantic primitives (Wierzbicka, 1996a) hypothesizes that the meaning of a word in any language can be decomposed into a finite set of language-independent sememes. For example, the English noun *plant* has distinct senses that correspond to the "factory" and "vegetation" concepts, respectively. The former can be represented by the sememes "industrial", "produce", and "institute/place", and the latter by "crop", "tree", and "flower/grass". (See Figure 2 for additional examples.) Although sememes provide a way of representing concepts, their incorporation into natural language processing (NLP) has been limited by a lack of available sememe resources for commonly used sense inventories. Tremendous efforts have been made to construct sememe knowledge bases (KBs) manually. One of the most widely used is HowNet (Dong and Dong, 2003), which unfortunately is limited to only two languages: English and Chinese.

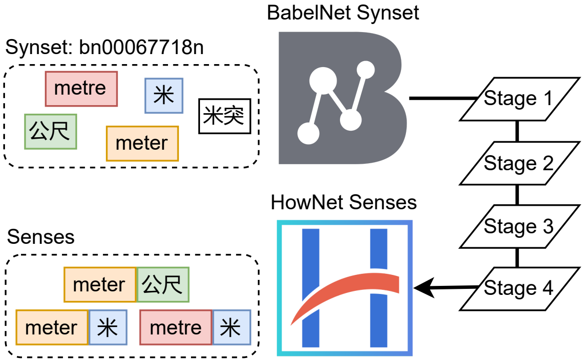
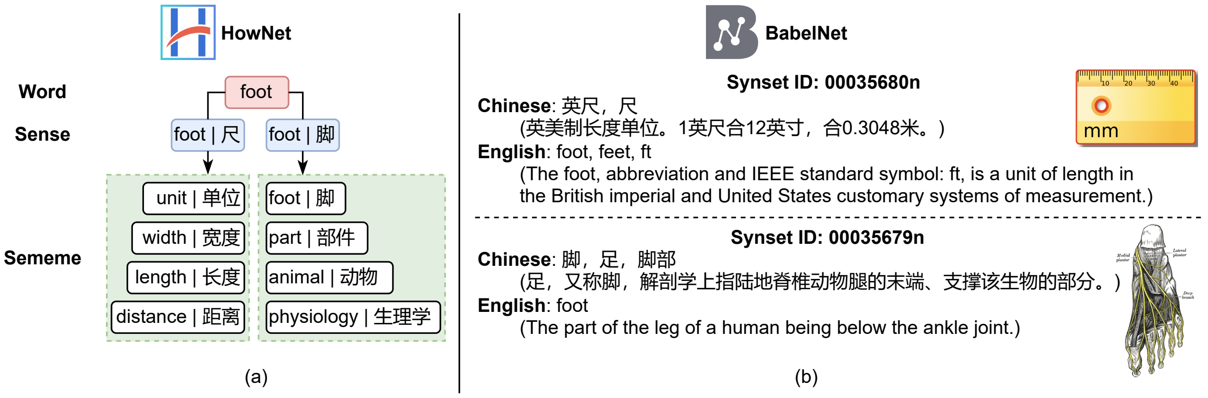


Figure 1: Given a BabelNet synset as input, our alignment algorithm identifies a set of corresponding HowNet senses. The algorithm consists of four stages which progressively relax alignment constrains.

A related problem is linking lexical resources to one another. As manually creating semantic knowledge bases remains a challenging and costly process, BabelNet (Navigli and Ponzetto, 2012) was instead created by combining WordNet (Miller, 1995), Wikipedia, and other resources. While BabelNet covers hundreds of languages, it does not include sememe information. Previous work on automatically predicting sememes for BabelNet concepts has depended on large human labeled data (Qi et al., 2020, 2022). Various systems for automatically aligning word senses across heterogeneous resources have been proposed to mitigate this issue (Meyer and Gurevych, 2011; Pilehvar and Navigli, 2014; Bao et al., 2022), but those methods do not leverage the unique structure of HowNet, which differs from other lexical databases.

In this work, rather than attempting to predict sememes directly, as in prior work, we instead attempt to align BabelNet concepts and HowNet senses. Since each HowNet sense is annotated with

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2789–2798
May 2-6, 2023 ©2023 Association for Computational Linguistics



Chinese: 英尺，尺
(英美制长度单位。1英尺合12英寸，合0.3048米。)
English: foot, feet, ft
(The foot, abbreviation and IEEE standard symbol: ft, is a unit of length in the British imperial and United States customary systems of measurement.)

Chinese: 脚，足，脚部
(足，又称脚，解剖学上指陆地脊椎动物腿的末端、支撑该生物的部分。)
English: foot
(The part of the leg of a human being below the ankle joint.)

Figure 2: A comparison of the representations of HowNet senses and BabelNet concepts.

sememes (c.f., Figure 2), correctly mapping a BabelNet concept to a HowNet sense will effectively add sememe information to the corresponding BabelNet synset, as well as associate the HowNet sense with information unique to BabelNet, such as synonymy information. We propose a fully unsupervised algorithm which achieves reliable and stable results, and produces fully explainable mappings. The ability to identify precisely why a particular sememe was associated with a given concept is a unique strength of our method, which we believe will facilitate analysis and development of downstream applications of our work.

The results of our experimental evaluation provide evidence that our unsupervised approach substantially outperforms previous supervised methods on the BabelNet sememe prediction task by up to 12% F-score. This also indirectly demonstrates the high accuracy of the generated alignment between BabelNet and HowNet, which we make available to facilitate further research.[1]

Our contributions are as follows:

- We propose the first method for aligning BabelNet synsets to HowNet senses, adding new information to both resources. Our method is unsupervised and explainable.
- We set a new state of the art for sememe prediction task on the BabelSememe dataset.
- We provide an API for identifying the HowNet senses and their sememes for any BabelNet synset that contains English or Chinese words.
- We perform a detailed analysis and ablation study of the results, focusing on the impact of the differences between HowNet and BabelNet.

2 Related Work

In this section, we first describe the two multilingual knowledge bases, and then discuss the tasks of sense alignment and sememe prediction.

2.1 Multilingual knowledge bases

Our main focus is on the sense alignment between BabelNet and HowNet (Figure 2).

BabelNet (Navigli and Ponzetto, 2012) is a multilingual sense inventory created by automatically combining various knowledge bases including WordNet and Wikipedia. The most recent version covers over 500 languages and contains 1.4 billion word senses. Following the WordNet model, words which are interchangeable in some context are grouped into synonym sets, or *synsets*. Each synset is associated with a unique lexical concept or named entity, and contains all the words which can express that concept or refer to that entity. In BabelNet, the *senses* of a word correspond to the concepts that it can express; that is, each sense of a word corresponds to a synset which contains the word, and shares its meaning with the other words in that synset. As shown in Figure 2, each synset has a unique ID and consists of all word senses that share the same meaning across various languages. Each synset is also associated with a gloss, and, optionally, example sentences or images.

HowNet is a sememe-based knowledge base for both Chinese and English. Each HowNet sense contains a unique Chinese-English word pair, associated with one or more sememes, as shown in Figure 2. Note that the meaning of the term *sense* is different in HowNet and BabelNet, and thus multiple HowNet senses may correspond to a single BabelNet concept. HowNet contains more than 2K

[1] Our code and used data are available at: https://github.com/senseAlign/BabelNet_2_HowNet

sememes, created by human experts, and more than 100K Chinese and 950K English words. Over its more than 20 years of development, HowNet has become one of the most popular knowledge base in the Chinese NLP community (Dong and Dong, 2003), and has been applied to downstream tasks such as word sense disambiguation (Zhang et al., 2005; Duan et al., 2007; Zhang et al., 2022), word representation learning (Niu et al., 2017), language modeling (Gu et al., 2018), textual adversarial attack (Zang et al., 2020; Qi et al., 2021), text matching (Lyu et al., 2021), and sememe prediction (Qi et al., 2020, 2022).

2.2 Sense aligning

The task of aligning lexical knowledge bases consists of associating entries in one resource with one or more entries in another. As described above, BabelNet was created by associating Wikipedia articles with WordNet senses. As another example, Open Multilingual WordNet (Bond and Foster, 2013) was created by linking wordnets covering various languages. Gurevych et al. (2012) create a sense similarity measure to align WordNet senses with entries in the German OmegaWiki. McCrae and Cillessen (2021) conduct an alignment between WordNet synsets and entities in Wikidata. Bao et al. (2022) present a translation-based method for aligning BabelNet concepts to entries in CLICS and OmegaWiki. Combining knowledge bases in this way provides additional information, and allows knowledge-based methods to be applied to more languages and tasks.

Various approaches to sense alignment between HowNet and the other KBs have also been investigated in prior work. Carpuat et al. (2002) use the tf-idf scores to align the senses in HowNet and synsets in WordNet. Chen and Fung (2004) present a method for aligning HowNet and FrameNet mapping by applying word sense disambiguation. Sornlertlamvanich et al. (2005) propose another algorithm for aligning HowNet and FrameNet, which is based on constructing feature vectors.

Besides, more general methods to link two or more lexical resources (ontologies) have been examined in prior work. McCrae and Cillessen (2021) proposed the *Lexicon Model for Ontologies* (Lemon) which aims to align any amount of lexical resources by modeling a universal semantic representation. Chiarcos et al. (2013) further combine multiple linking methods including Lemon and Graph construction, based on the theory of *Resource Description Framework*, and provide the insight in unifying lexical resources. Our work focuses solely on the alignment between HowNet and BabelNet as they represent two most widely used lexical knowledge bases for English and Chinese. However, similar ideas used in the work can also be applied for the general mapping for multilingual lexical resources.

2.3 Sememe prediction

The lack of sememe information in existing multilingual KBs was highlighted by Qi et al. (2020), motivating the sememe prediction task, which aims at bridging the gap between HowNet and BabelNet by predicting a set of most related sememes for a given BabelNet synset.

Qi et al. (2020) propose a model named *Sememe Prediction for BabelNet Synsets* (SPBS), which aims to learn a representation of the input synsets. This supervised model compares the representation of the input synset B to those in the training data, which are labeled with sememes. The input synset is then assigned sememes which are most frequently associated with the synsets that are most similar to it. Intuitively, if B has a similar representation to training synsets associated with some sememe s, then SPBS will be likely to predict sememe s for B. A variant of this method, SPBS-SS, also uses synset relation information from BabelNet to improve synset representations. An ensemble method combines both SPBS and SPBS-SS representations and yields better results.

Qi et al. (2022) propose another method: *Multilingual Synonyms and Glosses as well as Images* (MSGI). This approach uses multimodal information provided by BabelNet synsets, such as images. The model combines a text encoder, image encoder and multi-label classifier.

2.4 Sememes vs. Semantic Primes

While sememes have some similarities to the concept of semantic primitives (or semantic primes), they differ in their goals and representation. The purpose of semantic primes (Wierzbicka, 1996b) is to facilitate comprehensive linguistic analysis by formulating a set of universal basic concepts that can be used to convey the meaning of any linguistic expression (Goddard, 1998). Contrariwise, the purpose of sememes is to provide information about word senses by linking them to a subset of key terms. Sememes need not convey all of the informa-

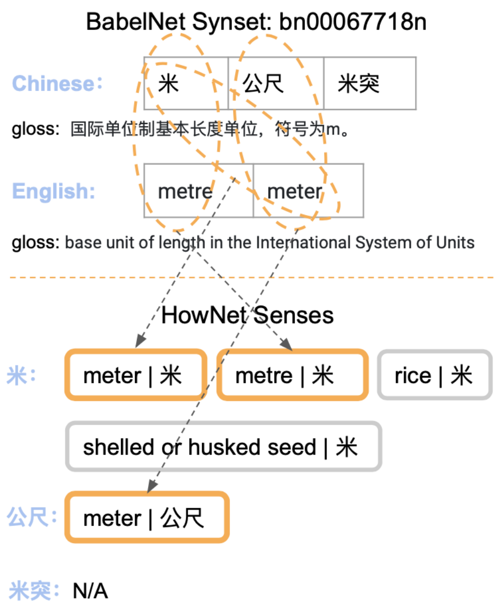

BabelNet Synset: bn00067718n
Chinese:
米
公尺
米突
gloss: 国际单位制基本长度单位，符号为m。
English:
metre
meter
gloss: base unit of length in the International System of Units
HowNet Senses
米:
meter | 米
metre | 米
rice | 米
shelled or husked seed | 米
公尺:
meter | 公尺
米突:
N/A

Figure 3: An illustration of the exact match algorithm (Stage 1), in which a BN synset is aligned to three different HowNet senses by comparing translation pairs.

tion about each sense. It may happen, for example, that distinct senses of distinct words may actually be associated with the same set of sememes, as in the case of apple (苹果) and banana (香蕉). While not as comprehensive as semantic primes, sememes are markedly less controversial (Fodor and Garrett, 1975), and are considered sufficient for a resource such as HowNet.

3 Methods

In this section, we introduce our method for the task of alignment between BabelNet and HowNet. Our method has four stages, each relaxing some constraint until an alignment is found. Such an alignment implicitly associates a BabelNet synset with a set of sememes from the aligned HowNet senses. We describe each of the four stages in a separate subsection.

Formally, our task is as follows: Let B be a BabelNet synset. Let $B_Z = \{z_1, \ldots, z_n\}$ be the set of Chinese words in B, and let $B_E = \{e_1, \ldots, e_m\}$ be the set of English words in B. We assume that at least one of B_Z and B_E is non-empty. The output is a set H of one or more HowNet senses that express the same concept as B.

3.1 Stage 1: Exact match

Our first stage uses a strict criterion for aligning a BabelNet synset B with HowNet senses, which is aimed at high precision rather than high coverage (later stages are aimed at improving coverage). It exploits the fact that each HowNet sense is annotated with a lexicalization in both Chinese and English, as well as the well-known observation that distinct senses of a word may translate differently (Gale et al., 1992). For each Chinese-English pair in B, that is, each $(z_i, e_j) \in B_Z \times B_E$, we check if there is a HowNet sense which has the Chinese lexicalization z_i and the English lexicalization e_j. If there is, we add that sense to H. (Figure 3 shows an overview of this stage.)

Once this stage is completed, if H is non-empty, we return it as the set of senses aligned to B. If H remains empty, we continue to Stage 2.

3.2 Stage 2: Word n-gram partial match

Every Chinese word consists of one to four characters; unlike English, each individual character has some meaning associated with it. Therefore, words that share one or more characters often share parts of their meaning. For example, a BabelNet synset for the word “延伸” (stretch) also contains the synonyms “伸长”, “伸展” and “延长”, each of which shares at least one character with “延伸”. As shown in Figure 4, the exact match approach in Stage 1 may miss the semantic correlation between such words, which reduces its coverage. We therefore propose the approach shown in Algorithm 1 to tackle such problems.

As in Stage 1, we again consider each translation pair in $(z_i, e_j) \in B_Z \times B_E$. For each such pair, we look for the HowNet sense with English word e_j, and with the Chinese word that has the maximum number of characters in common with z_i. Specifically, if z_i contains k characters, we first look for a sense which contains e_j and a Chinese word with k-1-gram in common with z_i. If such a sense is found, we add it to H. If not, we instead look for senses which have a k-2-gram in common with z_i. This partial-matching approach is explainable, unsupervised, and efficient; rather than using an uninterpretable and computationally expensive supervised language model, it instead exploits a useful property of the Chinese script. If H remains empty, we proceed to the next stage.

3.3 Stage 3: Sense information matching

Stages 1 and 2 are sufficient to find at least one HowNet sense for roughly 80% of synsets. Stage 3 attempts to map the remaining synsets by exploiting two key properties: (1) many words are

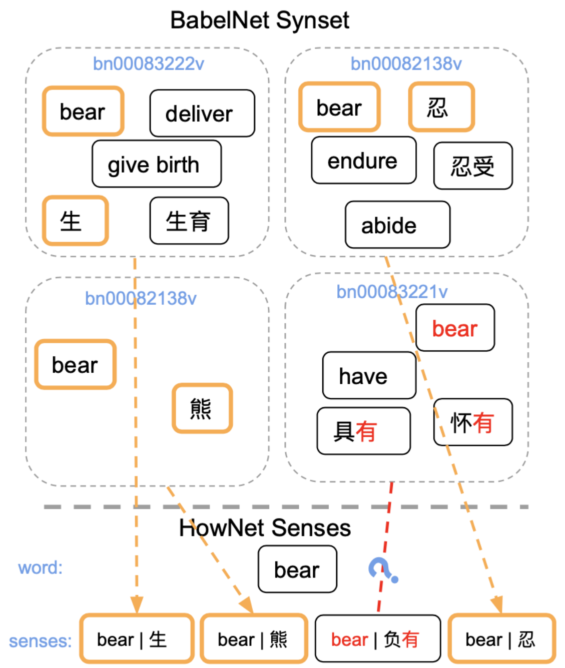


Figure 4: Example of inexact translation pairs during matching. Using exact algorithm, word “bear” has three synsets matched to some HowNet senses, while one synset is not matched despite having the same sense meaning with the unmatched HowNet sense. Note that character “有” means “have” and words “具有”, “怀有” and “负有” express the same meaning in context; Stage 2 will successfully map such a synset to a sense in HowNet, as indicated by red dashed line .

monosemous in HowNet; and (2) sememes are represented as words, which can often be found among the hypernyms of a given BabelNet synset B.

If B contains any Chinese or English words that have only one sense in HowNet, such as “ultraviolet” or “haemolytic”, then we assume the sense of that word in B must align to its only sense in HowNet. Monosemous words tend to be less frequent in text, but make up a substantial proportion of words in HowNet.[2] If even a single English or Chinese word in B is monosemous in HowNet, this sub-stage will add that sense to H.

If B has still not been aligned with any HowNet sense, we exploit the representation of sememes in HowNet to find an appropriate alignment. Each sememe of each sense is represented by an English word and a Chinese word, typically with a meaning that is more general than that of the senses associated with it. For example, one sense of the English word *teacher* has the following sememes: “human”, “occupation”, “education”, and “teach”. Our intuition is that these words will tend to overlap with the hypernyms of the corresponding sense in BabelNet. We therefore add to H any HowNet sense s such that s involves a word in B, and one of the sememes of s is represented by a word in a hypernym synset of B.

BabelNet only provides hypernym relations for nouns and verbs. All other words have no hypernym information, and thus can not be aligned based on sememe and hypernym agreement. We further relax our aligning condition by adding to H the HowNet senses that have any sememe overlap with each other and the English or Chinese words in the given synset. The intuition is that all words in a synset are closely related; if their HowNet senses have similar sets of sememes, we can align them to the given target synset.

3.4 Stage 4: Proper names

In our development experiments, we found that those BabelNet synsets B that could not be aligned with at least one HowNet sense by any of the first three stages were generally proper nouns. Proper nouns or named entities are represented in HowNet by a sense with the English label “ProperName” and an associated set of sememes. Therefore, the fourth and final stage of our method adds this sense to H. As the final result, the algorithm returns the set H, which contains the HowNet senses aligned to a given synset.

4 Experiments

In this section, we describe the experimental evaluation of our sense alignment algorithm. Since we are the first to propose an algorithm for aligning BabelNet and HowNet, there is no gold-annotated dataset for this task that we can use as a test set. To the best of our knowledge, the only dataset that connects BabelNet and HowNet is the BabelSememe dataset, which was constructed for the evaluation of sememe prediction models. Therefore, we evaluate our sense-alignment algorithm extrinsically on the task of sememe prediction.

In the sememe prediction task, given a BabelNet synset B, and a ground set S^* of sememes, we are required to identify a set of sememes $S \subseteq S^*$ which describe the meaning of B. Following prior work, we take the set of all sememes which appear in HowNet as S^*. The output S is then compared to a gold-standard set of sememes for B in the

[2] Due to limitations in the OpenHowNet API, we were not able to get the exact percentage.

Algorithm 1 Stage 2

```
Input  BabelSynsetID
Output  $\mathcal{H}$
$\mathcal{H} \leftarrow \emptyset$
$\mathcal{B} \leftarrow$ synset from BabelSynsetID
$\mathcal{C} \leftarrow$ all English-Chinese word pairs from $\mathcal{B}$
for $\mathrm{r} \in \{1, 2\}$ do
    for each $(\mathrm{e_i}, \mathrm{z_j}) \in \mathcal{C}$ do
        $\mathrm{k} \leftarrow \mathrm{length}(\mathrm{z_j}) - \mathrm{r}$
        $\mathcal{M} \leftarrow$ all k-grams of $\mathrm{z_j}$
        for each $\mathrm{m_i} \in \mathcal{M}$ do
            if $(\mathrm{e_i}, \mathrm{m_j})$ exists in HowNet then
                append HowNet sense to $\mathcal{H}$
            end if
        end for
    end for
    if $\mathcal{H} \neq \emptyset$ then
        return $\mathcal{H}$, end algorithm
    end if
end for
return $\emptyset$
```

BabelSememe dataset.

We compute S by applying our BabelNet-to-HowNet alignment algorithm, and taking the union of all sememes associated with the returned set of HowNet senses. This approach effectively reduces the task of sememe prediction to sense alignment. Since the quality of the alignment will be reflected in the accuracy of sememe prediction, the experimental results can serve as an evaluation of the sense alignment algorithm.

4.1 Experimental setup

In this section, we specify our experimental setup, including data statistics, evaluation metrics, baseline methods, and implementation details.

Data. BabelSememe (Qi et al., 2020) consists of 15,461 BabelNet synsets, each annotated with a set of HowNet sememes which comprise the meaning of the synset. These sememe sets were created manually by more than 100 bilingual (English and Chinese) human annotators. HowNet contains 2,106 sememe types; on average, 2.74 sememes assigned to each BabelNet synset. We use the existing test split, which consists of 10% of the dataset, as our test set. Since our method is unsupervised, we did not use the training or validation splits, instead developing our method on BabelNet alone, without reference to any labeled sememe prediction data.

English	Simplified	Traditional
noodles	面食	麵食
room	房间	房間
weather	天气	天氣
drink	饮	飲

Table 1: Examples in English and Chinese including both simplified and traditional versions.

Evaluation. Following previous work, we adopt mean average precision (MAP) and the F1 score and as our metrics.[3] For a given instance, a classification is positive if the synset is annotated with that sememe, or negative otherwise. MAP takes the weighted mean of the precision of each class, where each sememe is considered a separate class. The F1 score is the harmonic mean of the precision and recall of the predicted labels. Recall is the ratio of sememes correctly predicted to the number of sememes in the gold standard set, while precision is the proportion of sememes correctly predicted to the total number of sememes predicted by the method.

Comparison systems. We compare our sememe prediction results against systems from prior work, including the state of the art. In particular, we compare with three variants of SPBS model proposed by Qi et al. (2020), and five variants of MSGI model proposed by Qi et al. (2022) which we mention in Section 2. Additionally, we compare to *LR-NASARI* (Qi et al., 2020), a logistic regression model trained on NASARI embeddings (Camacho-Collados et al., 2016), and *TransE*, a relational prediction models proposed by (Bordes et al., 2013). In contrast to our unsupervised approach, all of the comparison systems are supervised, and model sememe prediction as a multi-class classification task.

Implementation details. We use the BabelNet Python API[4] with BabelNet 5.0 to retrieve English and Chinese words from synsets. We use OpenHowNet API[5] for retrieving HowNet senses and their corresponding translation pairs (Qi et al., 2019). One technical issue we encountered is the use of both simplified and traditional Chinese characters in BabelNet, likely due to the use of multiple heterogeneous resources in the construction

[3] We use the evaluation script provided by Qi et al. (2022).

[4] https://babelnet.org/

[5] https://github.com/thunlp/OpenHowNet

Methods	**Noun**		**Verb**		**Adj**		**Adv**		**Overall**	
	MAP	**F1**	**MAP**	**F1**	**MAP**	**F1**	**MAP**	**F1**	**MAP**	**F1**
LR-NASARI (Qi et al., 2020)	54.54	39.81	–	–	–	–	–	–	–	–
TransE (Bordes et al., 2013)	61.05	46.78	34.75	26.7	29.11	22.99	30.05	20.69	51.73	39.73
SPBS-SR (Qi et al., 2020)	65.16	49.75	–	–	–	–	–	–	–	–
SPBS-RR (Qi et al., 2020)	62.50	47.92	34.76	25.28	32.68	24.51	30.86	20.07	57.64	45.61
Ensemble (Qi et al., 2020)	68.85	55.35	34.76	25.28	32.68	24.51	30.86	20.07	57.64	45.61
MSGI -Synonym (Qi et al., 2022)	67.40	59.07	35.31	24.99	36.33	26.18	48.33	37.45	57.25	48.54
MSGI -Glosses (Qi et al., 2022)	66.90	56.99	54.22	41.54	53.11	39.20	**68.76**	55.14	62.67	52.21
MSGI -Image (Qi et al., 2022)	71.41	61.58	59.70	44.29	55.86	43.15	63.81	51.63	67.13	56.62
MSGI -MSCP (Qi et al., 2022)	70.58	61.99	57.55	43.27	52.57	40.61	68.49	52.79	65.70	56.05
MSGI (Qi et al., 2022)	71.81	64.36	**59.78**	47.01	55.61	41.02	68.52	55.20	67.23	57.68
Ours	**75.63**	**72.63**	57.70	**56.53**	**66.57**	**69.35**	64.63	**62.98**	**71.49**	**69.69**

Table 2: Main results on BabelSememe test set. Numbers in bold font represent the highest value in the column.

Data	**Noun**		**Verb**		**Adj**		**Adv**		**Overall**	
	MAP	**F1**	**MAP**	**F1**	**MAP**	**F1**	**MAP**	**F1**	**MAP**	**F1**
Train	75.73	72.95	55.33	55.85	64.12	66.08	68.08	69.87	70.71	69.29
Valid	74.75	73.16	57.78	57.21	69.70	71.12	69.33	70.36	71.48	70.59
Test	75.63	72.63	57.70	56.53	66.57	69.35	64.63	62.98	71.49	69.69

Table 3: Results of our method on the training, validation, and test sets.

of BabelNet. Table 1 shows some examples contrasting simplified and traditional characters. Since HowNet contains simplified Chinese characters exclusively, we use the Python `zhconv` library to convert all traditional Chinese characters into their simplified versions.

4.2 Results

The results on the BabelSememe test set are shown in Table 2. Following previous work, we report results for nouns, verbs, adjective, adverbs, as well as for all words. Our overall results outperform all prior work according to both metrics, establishing our method as the new state of the art for sememe prediction. In particular, the improvement in overall F1 score is 12% over the previous supervised state-of-the-art method, which was designed specifically for this task. These results are remarkable considering that our method is not supervised, and demonstrate its efficacy and utility.

In order to assess the generality of our method, we also report the results on the training and validation splits of the BabelSememe dataset, which are otherwise unused by our method. Table 3 shows that our results on these additional splits do not differ substantially from our results on the test set. This provides strong evidence that our method is consistent and reliable.

4.3 Stage analysis

As described in Section 3, our method consists of a sequence of four stages. In this section, we conduct additional experiments on the test set by examining the F1, MAP, and synset coverage rate (percentage of synsets that are mapped to at least one HowNet sense) of our system after every stage.

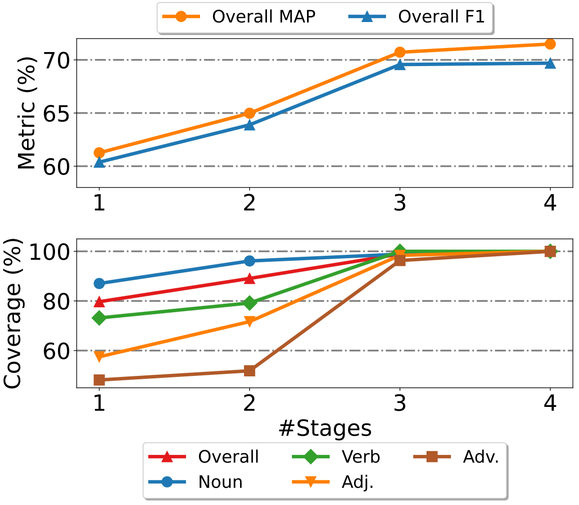


Figure 5: Experimental results for stages 1 to 4.

The results are shown graphically in Figure 5 and numerically in Table 5. We see an increasing trend in F1, MAP, and synset coverage rate with respect to the number of stages used. Stage 1 is

Stage	BableNet			HowNet	
	Synset ID	Synset EN	Synset ZH	Aligned Senses	Aligned Sememes
1	bn:00028111n	doll, dolly	洋娃娃, 玩偶	doll | 洋娃娃 doll | 玩偶	tool | 用具, recreation | 娱乐
2	bn:00002393n	alabaster, alabastar, gypsum	雪花石膏, 汉白玉	gypsum | 石膏	material | 材料, tool | 用具, medical | 医
3	bn:00040267n	geranium, cranesbil	天竺葵, 洋绣球	fish pelargonium | 天竺葵	FlowerGrass | 花草, medicine | 药物
4	bn:00048483n	Judas Iscariot, Judas	加略人犹大	ProperName | 专	ProperName | 专

Table 4: A case study of the alignment in each stage.

sufficient to cover roughly 80% of synsets, and achieve roughly 60% F1, but has particular difficulty finding HowNet senses for adjective and adverb synsets. Adding stages 2 and 3 greatly increases coverage of adjectives and adverbs respectively, with concomitant increases in F1 and MAP. Stage 4 is shown to provide marginal improvement, as almost all synsets can be assigned at least one sememe after stage 3.

4.4 Alignment analysis

To provide additional insight into the relationship between BabelNet and HowNet, we analyzed various properties of the alignment produced by our method. Theoretically, our algorithm could produce many-to-many alignments, with one BabelNet synset aligned to multiple HowNet senses, and vice versa. In practice, the average number of HowNet senses aligned to a given BabelNet synset in the BabelSememe test set is 3.99. This suggests that a single BabelNet concept is often represented by multiple HowNet senses, each of which may be labeled with different sememes. On the other hand, less than 1% of the HowNet senses were aligned to multiple BabelNet synsets (excluding the "ProperName" cases from Stage 4). This suggests that HowNet senses are no more fine grained than BabelNet synsets; in other words, if a sense distinction is made in BabelNet, it is likely made in HowNet as well.

We also analyzed the alignments for each part of speech and for each stage in our method. In Figure 6, we see that verb synsets are aligned to more senses than any other part of speech at every stage. For all parts of speech except adverbs, Stage 2 produces more mappings than any other stage. Since Stage 2 works by allowing partial matches in Chinese words (i.e. an alignment can be made on the bases of a single character shared by two Chinese words with two characters each), this suggests that the two knowledge bases often contain different Chinese words for a given concept. However, our method is still able to align them by identifying shared characters. If a sense reaches Stage 4, it will be assigned exactly one HowNet sense, the "ProperName" sense.

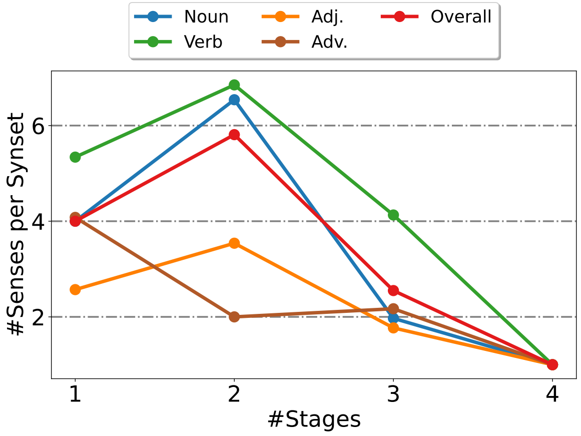


Figure 6: The average number of HowNet senses matched to each BabelNet synsets in each stage on the BabelSememe test set.

Table 4 shows example alignments from each stage. Row 1 shows that Stage 1 of our algorithm is able to correctly align the synset containing the words "doll" and "dolly" to the corresponding HowNet sense of each word. We can then use this alignment to retrieve the sememes "tool" and "recreation" for this synset. In Row 2, the synset containing the word "gypsum" does not contain any Chinese words which have HowNet senses. However, Stage 2 correctly finds a correct alignment between the BabelNet synset containing "雪花石膏" and a HowNet sense for "石膏". In Stage 3, the word "天竺葵" has only one sense in HowNet. We therefore map the BabelNet synset to that HowNet sense. Lastly, the word "Judas" has no senses in HowNet, therefore it is mapped to HowNet's "ProperName" sense in Stage 4.

POS	Stage 1		Stage 2		Stage 3		Stage 4	
	MAP	F1	MAP	F1	MAP	F1	MAP	F1
Noun	71.18	68.89	73.40	70.80	75.14	72.49	75.63	72.63
Verb	44.23	45.62	48.10	48.72	56.16	56.53	57.70	56.53
Adj.	37.30	40.06	47.43	50.95	65.37	69.15	66.57	69.35
Adv.	28.91	31.38	31.38	33.84	63.37	62.98	64.63	62.98
Overall	61.26	60.38	64.98	63.89	70.72	69.56	71.49	69.69

Table 5: Experimental results after each stage.

5 Conclusion

We have presented a novel unsupervised method for aligning two lexical-semantic knowledge bases, BabelNet and HowNet. The results of our experiments on leveraging the sense alignment for the task of sememe prediction demonstrate that our algorithm is highly effective, yielding substantially better results than state-of-the-art supervised systems designed specifically for this task. In the future, we would like to leverage sense alignment for other semantic tasks, including word sense disambiguation.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

Limitations

Our proposed algorithm only works with synsets that contain at least one Chinese or English word. Although this condition is satisfied for a majority of BabelNet synsets, there remain some multilingual synsets that could not be aligned. In addition, a intrinsic evaluation of the produced alignment could not be performed because of the lack of an existing gold data set.

References

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2022. Lexical resource mapping via translations. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7147–7154, Marseille, France. European Language Resources Association.

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

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Marine Carpuat, Grace Ngai, Pascale Fung, and Kenneth Church. 2002. Creating a bilingual ontology: A corpus-based approach for aligning wordnet and hownet. In *Proceedings of the 1st Global WordNet Conference*, pages 284–292. Citeseer.

Benfeng Chen and Pascale Fung. 2004. Automatic construction of an English-Chinese bilingual FrameNet. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 29–32, Boston, Massachusetts, USA. Association for Computational Linguistics.

Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. Towards open data for linguistics: Linguistic linked data. *New trends of research in ontologies and lexical resources: Ideas, projects, systems*, pages 7–25.

Zhendong Dong and Qiang Dong. 2003. Hownet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003*, pages 820–824. IEEE.

Xiangyu Duan, Jun Zhao, and Bo Xu. 2007. Word sense disambiguation through sememe labeling. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1594–1599, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jerold A Fodor and Merrill F Garrett. 1975. The psychological unreality of semantic representations. *Linguistic Inquiry*, 6(4):515–531.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, volume 112. Citeseer.

Cliff Goddard. 1998. Bad arguments against semantic primitives.

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 the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4642–4651, Brussels, Belgium. Association for Computational Linguistics.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY - a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France. Association for Computational Linguistics.

Boer Lyu, Lu Chen, Su Zhu, and Kai Yu. 2021. Let: Linguistic knowledge enhanced graph transformer for chinese short text matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13498–13506.

John P. McCrae and David Cillessen. 2021. Towards a linking between WordNet and Wikidata. In *Proceedings of the 11th Global Wordnet Conference*, pages 252–257, University of South Africa (UNISA). Global Wordnet Association.

Christian M Meyer and Iryna Gurevych. 2011. What psycholinguists know about chemistry: Aligning wiktionary and wordnet for increased domain coverage. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 883–892.

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

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

Yilin Niu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Improved word representation learning with sememes. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2049–2058, Vancouver, Canada. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A robust approach to aligning heterogeneous lexical resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–478.

Fanchao Qi, Liang Chang, Maosong Sun, Sicong Ouyang, and Zhiyuan Liu. 2020. Towards building a multilingual sememe knowledge base: Predicting sememes for babelnet synsets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8624–8631.

Fanchao Qi, Chuancheng Lv, Zhiyuan Liu, Xiaojun Meng, Maosong Sun, and Hai-Tao Zheng. 2022. Sememe prediction for BabelNet synsets using multilingual and multimodal information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. 2019. Openhownet: An open sememe-based lexical knowledge base. *arXiv preprint arXiv:1901.09957*.

Fanchao Qi, Yuan Yao, Sophia Xu, Zhiyuan Liu, and Maosong Sun. 2021. Turn the combination lock: Learnable textual backdoor attacks via word substitution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4873–4883, Online. Association for Computational Linguistics.

Virach Sornlertlamvanich, Canasai Kruengkrai, Shisanu Tongchim, Prapass Srichaivattana, and Hitoshi Isahara. 2005. Term-based ontology alignment. In *Proceedings of the Second International Workshop on UNL, Other Interlinguals and their Applications, Mexico City, Mexico*. Citeseer.

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

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

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

Xiang Zhang, Bradley Hauer, and Grzegorz Kondrak. 2022. Improving HowNet-based Chinese word sense disambiguation with translations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4530–4536, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuntao Zhang, Ling Gong, and Yongcheng Wang. 2005. Chinese word sense disambiguation using hownet. In *Advances in Natural Computation*, pages 925–932, Berlin, Heidelberg. Springer Berlin Heidelberg.