OpenHowNet: An Open Sememe-based Lexical Knowledge Base

Fanchao Qi[1], Chenghao Yang[2*], Zhiyuan Liu[1], Qiang Dong[3], Maosong Sun[1], Zhendong Dong[3]

[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]Software College, Beihang University
[3]HowNet Technology Inc.

Abstract

In this paper, we present an open sememe-based lexical knowledge base OpenHowNet. Based on well-known HowNet, OpenHowNet comprises three components: core data which is composed of more than 100 thousand senses annotated with sememes, OpenHowNet Web which gives a brief introduction to Open-HowNet as well as provides online exhibition of OpenHowNet information, and Open-HowNet API which includes several useful APIs such as accessing OpenHowNet core data and drawing sememe tree structures of senses. In the main text, we first give some backgrounds including definition of sememe and details of HowNet. And then we introduce some previous HowNet and sememe-based research works. Last but not least, we detail the constituents of OpenHowNet and their basic features and functionalities. Additionally, we briefly make a summary and list some future works.

1 Introduction

In the field of Natural Language Processing (NLP), words are generally the smallest objects of study because they are considered as the smallest meaningful units that can stand by themselves of human languages. However, the meanings of words can be further divided into smaller parts. For example, the meaning of "man" can be considered as the combination of the meanings of "human","male" and "adult", and the meaning of "boy" is composed of the meanings of "human", "male" and "child". In linguistics, the minimum indivisible units of meaning, i.e. semantic units, are defined as sememes (Bloomfield, 1926). And some linguists believe that meanings of all the words can be composed of a limited closed set of sememes.

---

*Work done while doing internship at Tsinghua University.

However, sememes are implicit and as a result, it is hard to intuitively define the set of sememes and determine which sememes a word have at a glance. Therefore, some researchers spend tens of years sifting sememes from all kinds of dictionaries and linguistic knowledge bases (KBs), and annotating words with these selected sememes to construct sememe-based linguistic KB. HowNet (Dong and Dong, 2003) is the most famous one of such KBs.

2 HowNet

HowNet was initially designed and constructed by Zhendong Dong and his son Qiang Dong (Figure 1) in the 1990s. And it has kept frequent updating since it was published in 1999.

(a) Zhendong Dong   (b) Qiang Dong

Figure 1: Founders of HowNet

The sememe set of HowNet is determined by extracting, analyzing, merging and filtering semantics of thousands of Chinese characters. And the sememe set can also be adjusted or expanded in the subsequent process of annotating words. Each sememe in HowNet is represented by a word or phrase in Chineses and English such as `human|人` and `ProperName|专`.

HowNet also builds a taxonomy for the sememes. All the sememes of HowNet can be classified as one of the following types: Thing, Part,
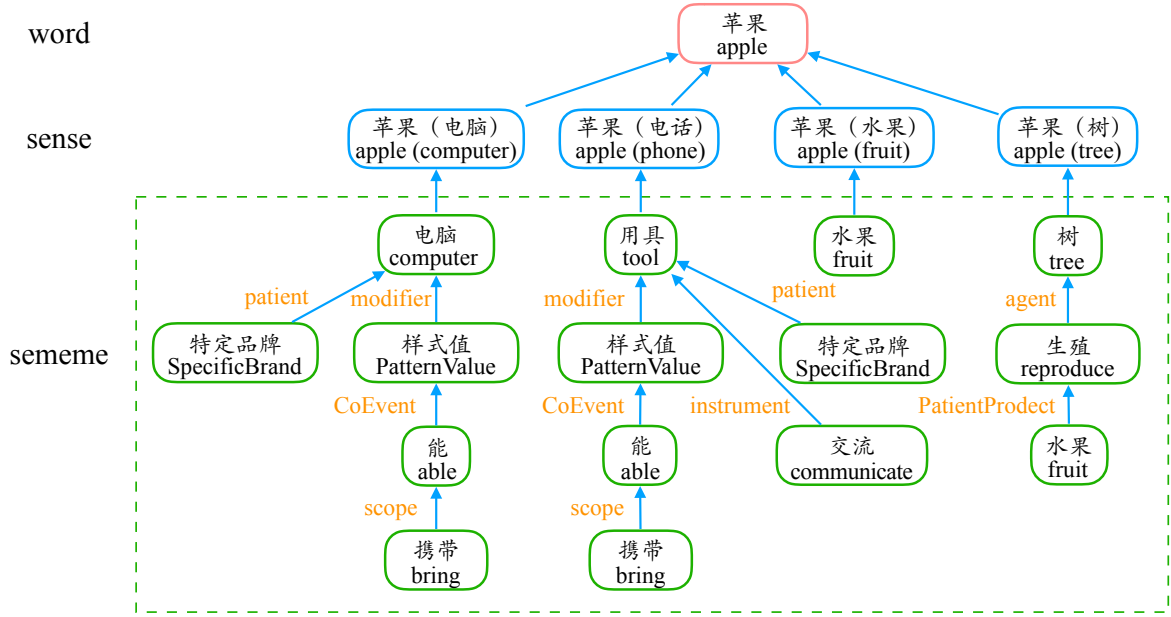
Figure 2: An example of word annotated with sememes in HowNet

Attribute, Time, Space, Attribute Value and Event. In addition, to depict the semantics of words more precisely, HowNet incorporates relations between sememes, which are called "dynamic roles", into the sememe annotations of words.

Considering the polysemy, HowNet differentiates diverse senses of each word in the sememe annotations. And each sense is also expressed in both Chinese and English. An example of sememe annotation for a word is illustrated in Figure 2. We can see from the figure that the word "apple" has four senses including "apple (computer)", "apple (phone)", "apple (fruit)" and "apple (tree)", and each sense is the root node of a "sememe tree" where any pair of father and son sememe node is multi-relational.

Additionally, HowNet annotates POS tag for each sense, and add sentiment category as well as some usage examples for certain senses.

The latest version of HowNet is published in January 2019 and the statistics are shown in Table 1.

Type	Count
Sense	229,767
Distinct Chinese word	127,266
Distinct English word	104,025
Sememe	2,187

Table 1: Statistics of latest version of HowNet

3 HowNet and Sememe-related Researches

Since HowNet was published, it has attracted wide attention. People use HowNet and sememe in various NLP tasks including word similarity computation (Liu and Li, 2002), word sense disambiguation (Zhang et al., 2005), question classification (Sun et al., 2007) and sentiment analysis (Dang and Zhang, 2010; Fu et al., 2013). Among these researches, Liu and Li (2002) is one of the most influential works, in which similarities of given two words are computed by measuring the degree of resemblance of their sememe trees.

Recent years also witness some works incorporating sememes into neural network models. Niu et al. (2017) propose a novel word representation learning model named SST that reforms Skipgram (Mikolov et al., 2013) by adding contextual attention to senses of target word, which are represented with combinations of corresponding sememes' embeddings. Experimental results show that SST can not only improve the quality of word embeddings but also learn satisfactory sense embeddings to do word sense disambiguation.

Gu et al. (2018) incorporate sememes into the decoding phase of language modeling where sememes are predicted first, and then senses and words are predicted in succession. The proposed model shows enhancement in the perplexity of language modeling and performance of down-

stream task headline generation.

Besides, HowNet is also utilized in lexicon expansion (Zeng et al., 2018), semantic rationality evaluation (Liu et al., 2018), etc.

Considering that human annotation is time-consuming and labor-intensive, some works attempt to employ machine learning methods to predict sememes for new words automatically. Xie et al. (2017) propose the task firstly and present two simple but effective models SPWE, which is based on collaborative filtering and SPSE, which is based on matrix factorization. Jin et al. (2018) further take the internal information of words into account when predicting sememes and achieve considerable boost of performance. And Li et al. (2018) take advantage of definitions of words to predict sememes. As for Qi et al. (2018), they propose the task of cross-lingual lexical sememe prediction and present a bilingual word representation learning and alignment-based model which demonstrates effectiveness in predicting sememes for cross-lingual words.

4 OpenHowNet

With the support of founders of HowNet, Open-HowNet encompasses the core data of HowNet and provides free download. In addition, Open-HowNet comprises OpenHowNet Web and Open-HowNet API.

OpenHowNet Web[1] gives a concise introduction to OpenHowNet, provides statistics of Open-HowNet dataset, lists sememe-related researches as well as history and founders of HowNet. Moreover, it has the functions of sense retrieval and sememe tree demonstration. Figure 3 shows how OpenHowNet Web illustrates the information of a sense in HowNet. From the figure, we can see that the exhibited information of a sense includes sense ID, POS tag, sememe-based definition, sememe tree and semantically near senses, which are computed with the sememe-based word similarity algorithm proposed by Liu and Li (2002).

Also, on the website of OpenHowNet, you can apply to download the whole OpenHowNet dataset and semantic representation learning results, which are learned by SST model (Niu et al., 2017) and include word, sense and sememe embeddings. Downloading is totally free and you just need to fill in a simple form including email

Figure 3: An example of sense information illustration on OpenHowNet Web

address in order to inform you of update of Open-HowNet.

OpenHowNet API[2] provides some useful APIs including searching senses and sememes from OpenHowNet dataset, drawing sememe tree of a given sense, computing word similarity based on sememe tree comparison (Liu and Li, 2002), etc. There are documents describing how to uses these APIs and you can also try the Jupyter Notebook tutorial.

5 Conclusion and Future Work

In this paper we present OpenHowNet, an open sememe-based lexical knowledge base which is based on well-known HowNet. We give an introduction to sememes, philosophy and structure of HowNet, sememe-related researches as well as basic components of OpenHowNet and their functions.

In the future, we will try to utilize machine learning methods to check and improve the annotation consistency of OpenHowNet. Moreover, we will attempt to ensemble various automatic sememe prediction methods together with online interactive annotating to annotate more words with sememes and enlarge the scale of OpenHowNet. Also, we want to transfer the existing sememe knowledge to other languages and build a multi-lingual version of OpenHowNet.

---

[1] https://openhownet.thunlp.org/home

[2] https://github.com/thunlp/OpenHowNet-API

References

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

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

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

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.

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*.

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*.

Wei Li, Xuancheng Ren, Damai Dai, Yunfang Wu, Houfeng Wang, and Xu Sun. 2018. Sememe Prediction: Learning Semantic Knowledge from Unstructured Textual Wiki Descriptions. *arXiv preprint*.

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

Shu Liu, Jingjing Xu, Xuancheng Ren, and Xu Sun. 2018. Evaluating semantic rationality of a sentence: A sememe-word-matching neural network based on hownet. *arXiv preprint*.

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

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

Fanchao Qi, Yankai Lin, Maosong Sun, Hao Zhu, Ruobing Xie, and Zhiyuan Liu. 2018. Cross-lingual Lexical Sememe Prediction. In *Proceedings of EMNLP*.

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.

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*.

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