LanguageNet: Learning to Find Sense Relevant Example Sentences

Shang-Chien Cheng[1], **Jhih-Jie Chen**[2], **Ching-Yu Yang**[2], **Jason S. Chang**[2]
[1]Institute of Information Systems and Applications
National Tsing Hua University
[2]Department of Computer Science
National Tsing Hua University
{ashleycheng, jjc, chingyu, jason}@nlplab.cc

Abstract

In this paper, we present a system, *LanguageNet*, which can help second language learners to search for different meanings and usages of a word. We disambiguate word senses based on the pairs of an English word and its corresponding Chinese translations in a parallel corpus, UM-Corpus. The process involved performing word alignment, learning vector space representations of words and training a classifier to distinguish words into groups of senses. *LanguageNet* directly shows the definition of a sense, bilingual synonyms and sense relevant examples.

1 Introduction

The polysemy of words, namely words with more than one sense, is one of the major challenges for English as a Second or Other Language (ESOL) learners. The issue of disambiguating polysemous words has attracted considerable attention to the NLP community. In order to derive and provide information about word senses, large knowledge based semantic lexicons have been developed, such as WordNet (Miller, 1995) and *BabelNet* (Navigli and Ponzetto, 2012). These resources are useful as a sense inventory for many NLP tasks. However, these knowledge bases contain few or even no example sentences. Thus, it is important to obtain more sense relevant WordNet examples, which can be more useful for language learners, or for training sense-aware NLP systems. Previous work has pointed out that "*two languages are more informative*" and there is typically "*one sense per translation*", since different word senses typically translate differently into a foreign language. Intuitively, sense relevant examples could be obtained using parallel corpora to distinguish word sense based on counterpart translations.

For example, the word "*plant*" has at least two different meanings in English: one is ORGANISM and the other is BUILDING, with corresponding Chinese translations 植物 and 工廠. Consider the two senses of the word "*plant*". The good example for the sense ORGANISM not "The *plant* has 300 workers.", which is irrelevant to the sense, but rather "Rice is a model *plant*". The Chinese translation of the first sentence is "該工廠有300餘名工人。" , while the translation for the second sentence is "水稻是用於研究的模式植物。" . The two Chinese translations of plant for the sense of building and organism is respectively "工廠" and "植物". Intuitively, by learning the characteristic translations of the category (ORGANISM or BUILDING) of a word sense, we can identify the meaning of head word in the given sentence, and thus retrieving sense relevant examples.

We use Chinese translations of a English head word, HW in a parallel corpus to disambiguate and identify the word sense of HW in WordNet, expected to provide example sentences for different senses of a word. Our approach learns how to effectively classify a word into its intended senses by using a collection of word-translation pairs (e.g., e-HowNet) and a class system based on basic level concepts in WordNet.

The rest of the paper is organized as follows. We review the related work in the next section. Next, we present our method for automatically learning to classify Chinese translations to possible set of senses and expected to provide good examples. Then we introduce the system design and interface. Finally, we exploit the great potential of the system and envision the future works.

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 99–102
Santa Fe, New Mexico, USA, August 20-26, 2018.

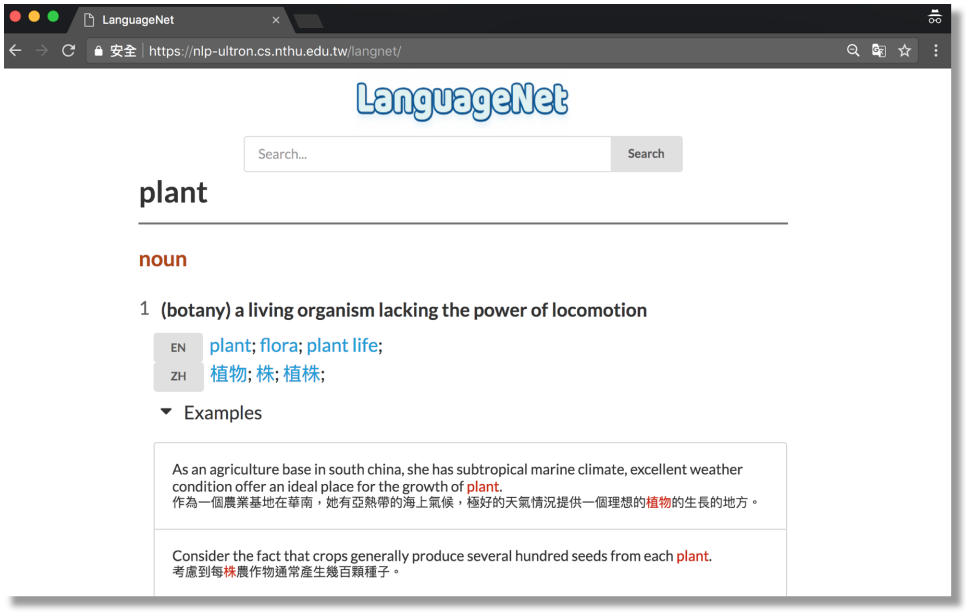

Figure 1: Example using *LanguageNet* typing "plant".

2 Related Work

Word sense disambiguation (WSD) has been an area of active research. WSD involves predicting the intended sense of a word in context based on a predefined set of senses. (Resnik and Yarowsky, 1999) use translate distinctions in the foreign language to identify sense distinctions in the source language for word sense disambiguation.

Recent work (e.g., Guo et al. (2014)), Upadhyay et al. (2017) also use parallel or mutilingual corpora to learn muti-sense vectors and capture different meanings of the same word in word sense disambiguation. In the area of word embedding, Chen et al. (2014) and Iacobacci et al. (2016) propose to use word embeddings instead of surface word as features to improve WSD performance. Yuan et al. (2016) propose to use similarity based on word embedding to identify the intended word sense.

3 Method

Problem Statement: We are given a polysemy W. We want to disambiguating the meaning of a word in sentences from a parallel corpus by providing bilingual examples for different senses of a word. Our goal is to find good examples for a given sense in *WordNet*. For this, we use a word-aligned parallel corpus, and then extract and classify translations for each W in question. Once the translations are classified and sense-tagged, we can then select example sentences for a given word sense.

3.1 Aligning and Extracting Word Translations

We use a large English-Chinese parallel corpus, UM-Corpus to obtain diverse kind of translations of a source-language word (Tian et al., 2014). For this, we use a word aligner, *fast align* (Dyer et al., 2013) to produce word alignments between the source words and translations in a parallel corpus. Then, for every pair of English word and translation, we computes pair similarity based on Dice coefficient and extract and classify translation words which frequently appear together with Dice similarity higher than a threshold. In the final step, we use these pairs of English and Chinese words to select example sentences and generate sense-tagged data.

3.2 Translation Similarity and Sense Labeling

In order to classify a given translation, we need a similarity measure between translations. First, we train a vector space representations of words using *word2vec* (Mikolov et al., 2013) on Chinese Wikipedia. Then, we define sense categories based on WordNet hierarchy (the top-level hypernym), and take a

Table 1: Example of word-translation pairs for training the classifier under two sense categories related to "plant"

Sense Category	Words	Translations
plant	grass, moss, fungus, ginkgo, crop	草類, 苔類, 菌類, 銀杏, 作物
building complex	factory, mill, sander, workshop	廠家, 製造廠, 研磨機, 製造場

Table 2: Result of classification under sense categories and corresponding gloss in WordNet related to the word "plant"

Translations	Sense category	Gloss in WordNet
植物, 植株	Plant	(botany) a living organism lacking the power of locomotion
工廠, 裝置, 廠	Building complex	buildings for carrying on industrial labor

collection of word-translation pairs (*W*, *T*) in c (e.g., *E-HowNet*) and assign each translation *T* to its all possible sense categories of the English word *W* based on WordNet hierarchy and a list of Basic Level Concepts. Since nouns have the explicit hierarchical structure of hypernym relationship in WordNet, we focus on noun senses. Although these (*T*, *CAT*) pairs may contain errors, we use them to train a classifier for tagging translations and apply the word embeddings trained with *word2vec* as features. We used 2,442 sense categories of noun based on WordNet concept and take word-translation pairs of 14,991 nouns from the bilingual dictionary for training. Table 1 displays example word-translation pairs for classifying translations related to the two sense categories related to "plant".

3.3 Classifying Chinese Translation to WordNet Senses

To identify the sense of a polysemy and its Chinese translation, we use the support vector machine (SVM) classifier trained on the sense category and translation, as described in the previous subsection. We use only unambiguous words with only one sense category (even when there are more than one WordNet sense). The feature is simply the word embedding of the translation. The output of the model are sense categories with probability and the probable category coinciding with the given English word will be returned as the output. We train a model to predict each sense category based on vector of translation T. If an ambiguous word has two senses, our SVM classifier use the feature vectors generated from training data of these two sense categories to learn a hyperplane which separates these two senses in high dimensional space. Given a translation of the word, the classifier then predict the sense category by its word vector and predict the sense according to the side of the hyperplane the vector lies in. After classification, we convert the category to the relevant sense to the English word belonging to this category. Table 2 displays the result of classification under sense categories and corresponding gloss in WordNet related to the word "plant".

3.4 Selecting example sentences

Finally, the Chinese translations of a English polysemous word which we extracted from corpus are sense-tagged. To help the user quickly and straightforwardly learn the usage for each sense of a word, good examples are really important. Therefore, we adopt the GDEX method (Adam Kilgarriff, 2008) to select representative sentences in candidates with translations of the same sense from the parallel corpus. The GDEX method score sentences by considering sentence length, word frequency, the presence of pronouns, location of the head word, and most importantly collocations.

3.5 *LanguageNet*

Our goal is to disambiguate word sense based on WordNet and provide sense relevant examples for the user using the additional information of Chinese translations of a English polysemy word. The

preliminary evaluation shows that the system can predict the relevant sense category with an accuracy rate of over 90% for a set of 12 words used in the WSD evaluation literature. We develop *LanguageNet* as a web application, which can assist second language learners to search for different meanings and usages of a word. We extracted 5,148 nouns in the UM-corpus to disambiguate senses and produce a bilingual word sense dataset for the system. An example *LanguageNet* search for the word "plant" is shown in Figure 1. *LanguageNet* has determined the intended senses of "plant" in sentences by predicting the sense class of the counterpart Chinese translations (e.g., 植物, 工廠, 廠, 植株). *LanguageNet* is accessible at (`http://nlp-ultron.cs.nthu.edu.tw/langnet/`).

4 Conclusion and Future Work

LanguageNet not only shows the definition and synonyms for each sense of a word, but also provides good sense relevant examples for the user. The preliminary assessment shows *LanguageNet* can provide reasonable accurate sense relevant translations and examples to support learning English with learner's native language (e.g., Chinese).

LanguageNet provide the best of the both worlds by combining a dictionary and a concordance to help English learners. Alternatively, the sense-labeled data can also be used by an NLP system to exploit semantic information. As future work, we plan to use the sense-labeled data to improve other WSD tasks and train the sense-specific word embeddings.

References

Katy McAdam Michael Rundell Pavel Rychlý Adam Kilgarriff, Miloš Husák. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In Janet DeCesaris Elisenda Bernal, editor, *Proceedings of the 13th EURALEX International Congress*, pages 425–432, Barcelona, Spain, jul. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 497–507.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 897–907.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. volume 5, pages 113–133. Cambridge University Press.

Liang Tian, Derek F Wong, Lidia S Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *LREC*, pages 1837–1842.

Shyam Upadhyay, Kai-Wei Chang, Matt Taddy, Adam Kalai, and James Zou. 2017. Beyond bilingual: Multi-sense word embeddings using multilingual context. *arXiv preprint arXiv:1706.08160*.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.