Title	語義の箱埋め込み学習とその応用
Author(s)	小田, 康平
Citation	
Issue Date	2024-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/18879
Rights	
Description	Supervisor: 白井 清昭, 先端科学技術研究科, 修士(情報科学)

Abstract

In general, a word is polysemous, i.e., has two or more senses. It is important to handle polysemous words appropriately to understand human language by a machine. Word Sense Disambiguation (WSD) is a well-known problem on polysemy of words. It is a task to select one appropriate sense of a word in a context from a predefined set of senses. WSD has been studied for many years in the field of natural language processing. Recent trends in WSD apply supervised learning of classification models based on neural networks using word sense tagged sentences as training data. However, most of the conventional WSD researches suppose that word senses in test data are always known, that is, appear in the training data. However, senses of words are changed and new senses are generated day by day. Therefore, it is required to judge whether the target word has an unknown sense that does not appear in training data (new sense), rather than simply selecting the appropriate sense from a predefined set of senses. It is also important to provide some useful information about a new sense to help human to understand it. For example, a superordinate sense of a new sense can be helpful to grasp a general meaning of the new sense.

The goal of this study is to learn box embeddings of word senses. The box embedding of a sense is an abstract meaning of a sense represented by a region in a vector space. Unlike previous WSD researches that represent a sense as a single vector, a box embedding can represent how broad or narrow a concept of a sense is. Box embeddings enable us to determine whether a word in a given context has a new sense and to predict a superordinate sense of a new sense. In this thesis, we extend an existing WSD method, MetricWSD, to learn box embeddings of senses. In addition, we propose two policies to create a small dataset used for computation of loss in training of the model of box embeddings. Our proposed method is applied to three tasks: WSD, new sense classification, and superordinate sense prediction of a new sense, then the performance of the proposed method is empirically compared with conventional methods that represent a sense as a single vector.

Box embedding is formulated as a pair of vectors: one represents the center of the box and the other represents the size of the box. Our proposed model produces a box embedding for each word in a sentence that represents the meaning of the word in that context. We call it "contextual box embedding" of an instance of a sense (a sense appeared in a particular sentence). The box embedding of a sense is obtained by averaging the contextual box embeddings of instances of the sense.

In MetricWSD, Bidirectional Encoder Representations from Transformers (BERT) was used as a model to obtain the contextual embedding of the target word. In this research, a fully connected layer (FCL) is attached to the final layer of BERT

to obtain the contextual box embedding of the target word. The input of the FCL is an embedding of the target word in the final layer of BERT, and the output is a vector representing the contextual box embedding. The half dimensions of the output vector represent the center of the box embedding, and another half represent the size of the box embedding.

The loss function for training the model is computed by considering the overlap of box embeddings of two senses. Intuitively, the loss function is designed so that the overlap of box embeddings of a sense and its superordinate sense becomes greater, and less otherwise. A dataset to be used for calculation of the loss, called episode, is prepared for each sense. In each episode, the model is trained by the following procedures. (1) Select N_C senses from all senses appearing in the training data. (2) For each sense, randomly select N_S sense instances (sentences including the sense) as the support set. (3) For each sense, randomly select N_Q sense instances, which are not chosen in the support set, as the query set. (4) Obtain a prototype representation of each sense by averaging contextual box embeddings of the instances in the support set. (5) Calculate the loss based on the output of the model (contextual box embeddings) for the query set and the prototype representations. (6) Update the model parameters to minimize the loss.

In the above step (1), we propose two policies to select N_C word senses. The first policy is S_r. The target sense and its superordinate senses are selected first, then the rest is selected randomly. The second policy is S_n. The target sense, its superordinate senses, its subordinate senses, and sibling senses are chosen first, then the rest is selected randomly. The policy S_n emphasizes that the box embedding of a sense does not overlap with that of its subordinate and sibling senses.

In the experiment to evaluate the proposed methods, our methods are applied to three tasks: WSD, new sense classification, and superordinate sense prediction of a new sense. To evaluate the ability of the model to learn box embeddings of different number of senses, three datasets with different sizes are prepared: $\mathcal{D}_{\text{living\_thing.n.01}}$, $\mathcal{D}_{\text{artifact.n.01}}$, and $\mathcal{D}_{\text{entity.n.01}}$. living_thing.n.01, artifact.n.01, and entity.n.01 are senses in WordNet (called synset), and each dataset contains sentences of subordinate senses of one of the three root senses. $\mathcal{D}_{\text{entity.n.01}}$ consists of senses of all nouns, while $\mathcal{D}_{\text{living\_thing.n.01}}$ and $\mathcal{D}_{\text{artifact.n.01}}$ consist of subsets of nouns. Two baselines are prepared: BERT-NN , which based on a pre-trained BERT, and MetricWSD.

As for the WSD task, the proposed methods outperformed the baselines for small datasets ($\mathcal{D}_{\text{living\_thing.n.01}}$ and $\mathcal{D}_{\text{artifact.n.01}}$). However, for the dataset consisting of many senses of all nouns ($\mathcal{D}_{\text{entity.n.01}}$), the proposed methods performed worse than the baselines. It indicates that the box embeddings of senses trained by the proposed model are not always good when the box embeddings of too many senses are trained simultaneously. As for the new sense classification task, the proposed methods outperformed the baselines on the dataset consisting of many senses of all nouns, which is inconsistent with the results of WSD. As for the task of superordinate sense prediction of a new sense, the proposed method outperformed the baseline in all three datasets. The parameter used for choosing candidates of superordinate senses was set to 0.5, 0.7, or 0.9, and it was found that the best thresholds were different for three datasets and two policies (S_r and S_n). To sum, the box embeddings of the senses learned by the proposed method was better representation than an ordinary single vectors for several sense-related tasks in some conditions. Especially, the box embeddings has ability to represent hypernym-hyponym relation of senses appropriately.