Analyzing Contextualized Knowledge Base Aware Representation Learning models for WSD

AAAI 2023, Mozhgan Saeidi

The Problem

Natural Language Processing (NLP) includes many tasks which, most of which are engaged with representation learning (RL). Text representation learning has shown its essential impact on the final results of NLP tasks, including Word Sense Disambiguation (WSD), Information Retrieval (IR), and Question Answering. After the development of deep neural networks, different approaches have been widely used to solve NLP tasks. Some of these deep neural networks are convolutional neural networks (CNNs), recurrent neural networks (RNNs), graph-based neural networks (GNNs), and attention mechanisms. Representation learning is also one of the tasks that use the power of deep learning to alleviate feature engineering difficulties. RL models usually use low-dimensional and dense vectors to represent the syntactic or semantic features of the language implicitly.

On the large corpus, the pre-trained models can learn language representations and then be used to solve downstream tasks. Between different RL approaches, the Skip-gram and GloVe are such models that are very shallow for computational efficiencies. While by emerging the deep models, including transformers, the RL architecture is transferred from shallow to deep. The pre-trained embeddings capture the semantics of the words they represent, but they suffer from the context in their representations. The importance of context in word representation is vital in some NLP tasks, like WSD.

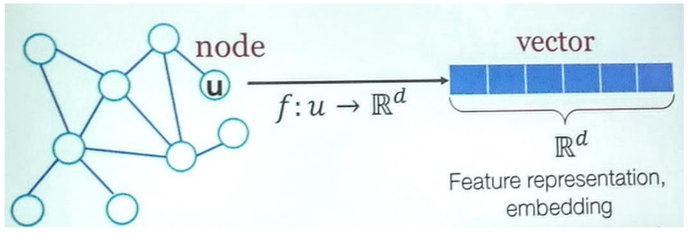

Fig. 1: A depict of vector representation learning

The biomedical domain is one of these domains that needs specific pre-trained language models; this need is because of the volume of the biomedical text, which is growing at a good speed and needs analysis for different problems. On average, more than 3000 new articles are published every day in peer-reviewed journals, excluding pre-prints and technical reports such as clinical trial reports in various archives. Consequently, there is increasingly more demand for accurate biomedical text mining tools to extract text information.
Word representations may improve the effectiveness of the disambiguation models if they carry helpful information from the context and the knowledge base. In this work, first, we provide an in-depth quantitative and qualitative analysis of existing transformer-based language models to understand their capabilities and potential limitations in encoding and recovering word senses. Second, we present a novel contextual-knowledge base aware sense representation learning method. The name of this new embedding approach is C-KASE stands for Contextualized-Knowledge base Aware Sense Embedding. The novelty in our representation is the integration of the knowledge base and the context. This representation lies in a space comparable to contextualized word vectors, thus allowing a word occurrence to be easily linked to its meaning by applying a simple nearest-neighbor approach. Finally, we compare our approach with state-of-the-art embedding methods for WSD.

The Architecture of C-KASE Embedding

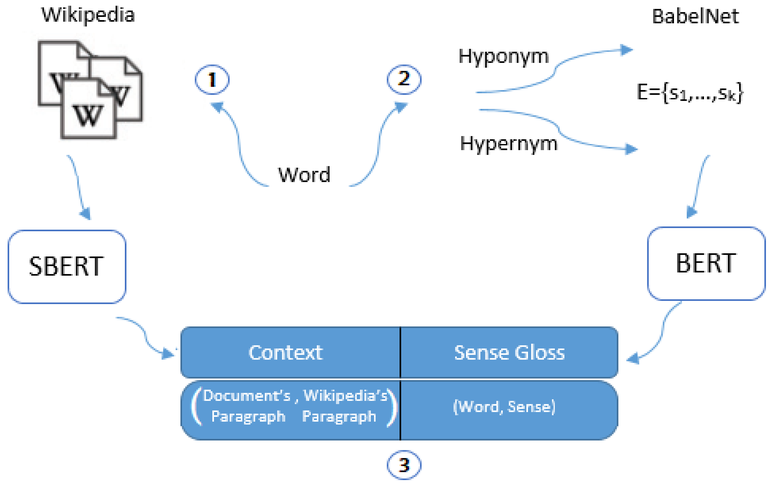

Fig. 2: Demonstration of the C-KASE representation and its three components. Component 1) Collecting all Wikipedia pages for ambiguous words, Component 2) Using hypernymy and hyponymy relations to extract all synsets for ambiguous words from Babelnet in set E, Component 3) Concatenating (word, sense) representation for all senses in E from the second component with (Document's paragraph, Wikipedia's paragraph) representation from the first component as context

C-KASE Representation Learning Model

Our proposed algorithm, C-BERT, is a pre-trained language representation model. C-BERT is created by combining semantic and textual information from the first paragraph of each sense's Wikipedia page and the paragraph of the input document text, which includes the senses. C-KASE is based on three components; Context Retrieval, Word Embedding, and Sense Embedding.

1-Context Retrieval:
For each synset s, we collect all the connected concepts to s from Wikipedia and BabelNet. We show this set of related synsets to s by R_s, which is

$$R_s = \{s' | (s, s') \in E\} \tag{1}$$

2-Word Embedding:
In the second component, we use BERT to extract the given ambiguous word from the input text.

$$WO(p_1, p_2) = (\sum_{w \in O} \frac{1}{r_w^{p_1} + r_w^{p_2}})(\sum_{i=1}^{|O|} \frac{1}{2i})^{-1} \tag{2}$$

3-Sense Embedding:
In this last component, we build the final representation of each mention. From the previous step, we took the representation of mention, R(m), and the representation of each one of its senses. Our unique representations combine the mention representation with sense representation, concatenating the two vector representations of R(m) and $R(s_i)$.

$$\text{Sim}(m, s_i) = Cosine(R(m, s_i), R(PD, PW)), \text{for } i = 1, \ldots, k \tag{3}$$

Experiments

We use the English WSD test set framework, which is constructed by five standard evaluation benchmark datasets. For WSD modeling, we employed a 1-nearest neighbor approach as previous methods in the literature to test our representations on the WSD task. For each target word m in the test set, we computed its contextual embedding using BERT and compared it against the embeddings of C-KASE associated with the senses of m. Hence, we took as a prediction for the target word the sense corresponding to its nearest neighbor.

Model	Senseval-2	Senseval-3	Semeval-7	Semeval-13	Semeval-15	All
BERT	77.1	73.2	66.1	71.5	74.4	73.8
LMMS	76.1	75.5	68.2	75.2	77.1	75.3
SensEmBERT	72.4	69.8	60.1	78.8	75.1	72.6
ARES	78.2	77.2	71.1	77.2	83.1	77.8
C-KASE	79.6	78.5	74.6	79.3	82.9	78.9

Fig. 3: The accuracy performance of WSD evaluation framework on the test sets nominal instances of the unified dataset.

Model	Nouns	Verbs	Adjectives	Adverbs
BERT	76.2	62.9	79.7	85.5
LMMS	78.2	64.1	81.3	82.9
ARES	78.7	67.3	82.6	87.1
C-KASE	79.6	69.6	85.2	89.3

Fig. 4: The performance of the 1-NN WSD of each embedding on All dataset split by parts of speech

	ARES		BERT		C-KASE	
Type	#Mis-D	ER	#Mis-D	ER	#Mis-D	ER
Noun	916	0.21	1023	0.24	877	0.20
Verb	540	0.33	613	0.37	502	0.30
Adj.	166	0.17	194	0.20	141	0.14
Adv.	45	0.13	50	0.14	37	0.10

Fig. 5: Error rate analysis of the 1-NN WSD evaluation framework with ARES, BERT, and C-KASE representations on the All dataset, separated by type.

Model	Noun	Verb	Adj.	Adv.
BERT	0.75	0.11	0.1	0.04
LMMS	0.80	0.1	0.07	0.03
SensEmBERT	0.79	0.1	0.09	0.02
ARES	0.81	0.09	0.07	0.03
C-KASE	0.84	0.085	0.06	0.01

Fig. 6: Confusion-Error table for Noun type by each model. This table shows how models are confused by the type of word at the time of disambiguation