Abstract

In this survey, we introduce methods to use knowledge features for natural language processing tasks. We first introduce current models for leveraging knowledge base features, then we discuss works of modeling data using graph neural networks or neural module networks.

1 Leveraging Knowledge Base features in NLP Tasks

There is a line of work leveraging knowledge base features for different natural language processing tasks. (Ahn et al., 2016) uses a simple idea to train a neural language model with the help of knowledge bases. For an entity word phrase in the sentence, the algorithm links the entity and finds all fact tuples \((s, r, o)\) in the knowledge graph. Then they sample words from object \(o\) to predict the following words. Since not all words in the sentence can be linked with an entity in the knowledge graph, there is also the original LSTM model for next work generation. To combine these two sources of prediction, a binary gate with learnable weight is used to control which word is used.

The following work (Yang & Mitchell, 2017) improved this one-hot search from knowledge bases. It learns the embedding of knowledge entities using model defined in (Yang et al., 2014), and find several candidate entities for each word. Then the model combines word embedding and entity embedding in a LSTM model for next work generation. To combine these two sources of prediction, a binary gate with learnable weight is used to control which word is used.

The following work (Yang & Mitchell, 2017) improved this one-hot search from knowledge bases. It learns the embedding of knowledge entities using model defined in (Yang et al., 2014), and find several candidate entities for each word. Then the model combines word embedding and entity embedding in a LSTM model for next work generation. To combine these two sources of prediction, a binary gate with learnable weight is used to control which word is used.

Graph Neural Networks

2.1 Graph Convolutional Networks

In this section we discuss popular methods to model graph structure using neural networks. (Kipf & Welling, 2016). The feature representation of GCN is as following:

\[
Z = f(X, A) = \text{Softmax}(\text{ReLU}(\hat{A}XW^{(0)})W^{(1)})
\]

Here \(Z \in \mathbb{R}^{N \times F}\) is the feature output, \(W^{(0)}\) and \(W^{(1)}\) are learnable weight of the feature transformation. This two-hop graph neural network is stacked with two layers of linearized graph filters, which was originally defined in (Defferrard et al., 2016). These filters are Chebyshev multinomials that approximate a linear filter in the graph Fourier domain.

Graph Attention networks (Veličković et al., 2017) is the improvement of GCN. It adds a learnable weight to multiply the adjacency matrix of the graph. This weight works as attention mechanism in other networks, to select the proper feature in the graph level. This work can be also extended to model relational data (Schlichtkrull et al., 2018).
2.2 Interpretablility of Graph Neural Networks

The work (Neil et al., 2018) advocate their model for link prediction by its interpretability. The work uses two methods to examine the weight of the feature. The first one is to examine the weight of the edge attention vector $C_{i,j,r}$, bigger weight means more important in classification. The other way to interpret the importance of feature is to augment the graph structure.

3 Dynamic Structure

The work (Andreas et al., 2016) uses dynamic network structure to deal with the visual question answering (VQA) problem. Given a question sentence such as “Where is the dog?”, the algorithm first uses semantic parsing algorithm to process the question as function modules: attend[dog] and classify[where], then call the neural module for each function to combine for the answer. The advantage of these models is that each type of the module share the same representation of features, and these sub-tasks are already solved by computer vision studies. However, this work needs manually labeled semantic parsing for the questions in the training tasks. When wrong model chosen for training, the noise become huge and the performance of model drastically decreases. There are a couple of works that uses the similar structure for NLP area (Wu et al., 2018; Li et al., 2018). (Li et al., 2018)

References


Jiang, He, Song, Yangqiu, Wang, Chenguang, Zhang, Ming, & Sun, Yizhou. 2017. Semi-supervised Learning over Heterogeneous Information Networks by Ensemble of Meta-graph Guided Random Walks.


