A Sememe-based Approach for Knowledge Base Question Answering

Peiyun Wu and Xiaowang Zhang

College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
{wupeiyun,xiaowangzhang}@tju.edu.cn**

Abstract. In this poster, we present a sememe-based approach to semantic parsing in question answering over knowledge base by leveraging a sememe-level semantics to improve the performance of semantic similarity between question and relations. Firstly, we propose a double-channel model to extract both sememe-level semantics and word-level semantics. Moreover, we present a context-based representation to encode the sememe of questions to refine sememe incorporation for reducing noise. Finally, we introduce a hierarchical representation to encode the sememe representation of relations to remove the ambiguity of words maximally. Experiments evaluated on benchmarks show that our model outperforms off-the-shelf models

1 Introduction

Knowledge base question answering (KBQA) is the task of accurately and concisely answering a natural language question over knowledge base (KB) by understanding the intention of the question. As a critical branch of KBQA, semantic parsing based approaches construct semantic parsing trees or equivalent query structures (also called *query graph*) to represent the given question, and then ranking them by calculating the semantic similarity with the question.

Most current works focus on selecting the semantic relations that most similar to a question to find the optimal query graph. Unfortunately, those existing approaches are limited in differentiating two relations with the similar word-level semantics due to the following issues: (1) Polysemous and low-frequency words often undermine the overall performance of semantic similarity measurement. (2) More minimal semantics of words between question and relations are ignored. Existing models heavily dependent on the embeddings closest to the question representation instead of extracting minimal semantic similarity between them.

To overtake the above limitations, we leverage sememe from external lexical-semantic resources. Sememes are minimum semantic units of word meanings [3], a word may have multiple senses, and a sense consists of several sememes. In this poster, we present a sememe-based approach to semantic parsing in KBQA by leveraging a sememe-level semantics.

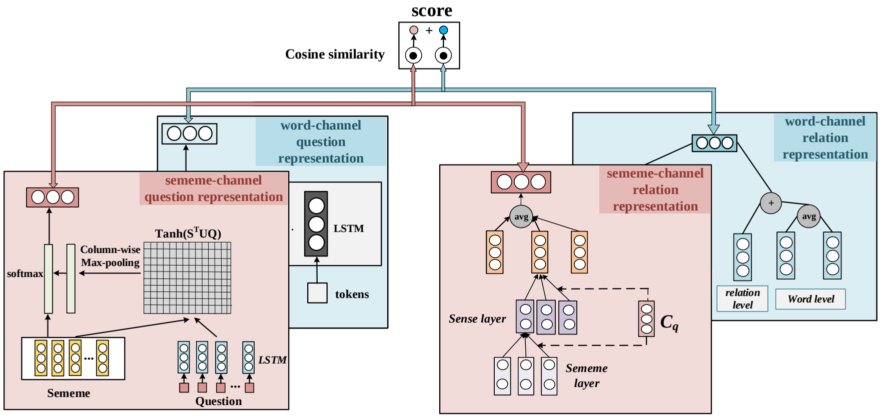
score
Cosine similarity
+
word-channel question representation
LSTM
tokens
word-channel relation representation
sememe-channel question representation
Tanh(S^TUQ)
Column-wise Max-pooling
softmax
Sememe
LSTM
Question
sememe-channel relation representation
avg
Sense layer
C_q
Sememe layer
relation level
Word level

Fig. 1: Diagram for our double-channel model.

2 Approach

Our model is shown in Fig. 1. In this paper, based on a heuristic algorithm in [2], we generate candidate query graphs with considering five kinds of semantic constraints: entity, type, temporal (explicit and inexplicit time), order, and compare.

2.1 Word-Channel Representation

In this part, we generate word-channel representation of question and relations. Given a question $\{w_1, w_2, ..., w_n\}$, we feed it into a bi-directional long short-term memory network (Bi-LSTM) to generate the hidden representation $Q = (h_1, \ldots, h_n)$ and obtain h_q after pooling operation. Then we transform h_q with a fully connected layer and a *ReLU* function to get the word-channel representation of the question:

$$q^w = \mathrm{ReLU}\left(W_q \cdot h_q + b_1\right) \tag{1}$$

where W_q denotes the linear transformation matrix.

To encode the word-channel relation representation, we take the relation-level (e.g. "*contained by*") and word-level(e.g. "*contained*","*by*") relation names into consideration. Given relations $\{r_1, r_2, ..., r_n\}$ in a query graph, for relation-level representations, we simply take each relation name as a whole unit, and translate it into vector representation as $\{r_1^{rl}, r_2^{rl}, ..., r_n^{rl}\}$. For word-level representations, we represent the word sequence of each relation using word averaging as $\{r_1^{wl}, r_2^{wl}, ..., r_n^{wl}\}$. We get the final vector of each word-channel relation representation as $r_i = r_i^{rl} + r_i^{wl}$. Finally, we apply pooling operation over all relations and obtain the word-channel representation of relations, denoted by r^w.

2.2 Sememe-Channel Representation

Sememe-Channel Question Representation We denote S_m as the set of all sememes occurring in the question. Then we map S_m into the vectors $S = (s_1, \ldots, s_n)$ and adopt a context attention mechanism to deemphasize irrelevant sememes and focusing on more correlative to context ones. The interactive context matrix is calculated as $S^q = \tanh(S^\top UQ)$. Then we obtain the vector s^q by the column-wise max-pooling operation over S^q and use the softmax function. Finally, we get the sememe-channel question representation as following:

$$q^s = W_{sq}(\text{softmax}(s^q) \cdot S) + b_2 \tag{2}$$

where W_{sq} is the parameter matrix.

Sememe-Channel Relation Representation In this part, we adopt a hierarchical attention method to obtain the sememe-level representation of relations and maximally remove the ambiguity. We denote $\mathcal{R}^{\text{sense}}_{w_{ij}}$ as a set of sense vectors of word w_{ij} as $\mathcal{R}^{\text{sense}}_{w_{ij}} := \{se_{ij1}, \ldots, se_{ijk}\}$. We denote $\mathcal{R}^{\text{sememe}}_{s_{ijk}}$ as a set of sememe vectors of sense s_{ijk} as $\mathcal{R}^{\text{sememe}}_{s_{ijk}} := \{sm_{ijk1}, \ldots, sm_{ijkm}\}$.

To obtain the context information of the given question, we denote its word embeddings average as q_{avg} and construct a context representation C_q as following:

$$C_q = \sum_{i=1}^{n} \text{softmax}(\tanh(w_i^\top \cdot q_{avg})) \cdot w_i \tag{3}$$

Through this, the vector of sense se_{ijk} in $\mathcal{R}^{\text{sense}}_{w_{ij}}$ is represented as below:

$$se_{ijk} = \sum_{c=1}^{m} \text{softmax}(W_{sm} \cdot \tanh(sm_{ijkc}^\top \cdot C_q) + b_3) \cdot sm_{ijkc} \tag{4}$$

where W_{sm} is a weight matrix. And the representation of the j-th word in the i-th relation is a weighted sum of its senses $\{se_{ij1}, \ldots, se_{ijk}\}$:

$$r_{ij} = \sum_{y=1}^{k} \text{softmax}(W_{se} \cdot \tanh(se_{ijy}^\top \cdot C_q) + b_4) \cdot se_{ijy} \tag{5}$$

Finally, we can apply average operation over all words in all relations and obtain the sememe-channel representation of relations, denoted by r^s. In this way, we can compute the semantic similarity score of two channel as following:

$$\mathit{Score} = \mathit{cos}\left(q^w, r^w\right) + \mathit{cos}\left(q^s, r^s\right). \tag{6}$$

3 Experiments and Evaluations

Due to Freebase no longer up-to-date, including the unavailability of APIs and new dumps, we use the full Wikidata dump as our KB. We conduct our experiments, namely, *WebQuestionSP* (WebQSP),*QALD-7*(Task 4, English). We use sememe annotations in HowNet for sememe-channel representation.

Table 1: Overall Average Results over Wikidata

Model	**WebQSP**			**QALD-7**		
	Precision	Recall	F1	Precision	Recall	F1
STAGG(2015) [2]	0.1911	0.2267	0.1828	0.1934	0.2463	0.1861
Yu et al.(2017) [5]	0.2094	0.2453	0.1987	0.2173	0.2084	0.1958
Sorokin et al.(2018) [1]	0.2686	0.3179	0.2588	0.2176	0.2751	0.2131
Maheshwari et al.(2019) [4]	0.2678	0.3182	0.2619	0.2493	0.2691	0.2436
word-channel	0.2686	0.3179	0.2588	0.1948	0.2535	0.2048
sememe-channel	0.2467	0.2991	0.2438	0.2309	0.2919	0.2382
Double-channel(Our)	**0.2721**	**0.3343**	**0.2776**	**0.2678**	**0.3182**	**0.2619**

By Table 1, we show that our model is superior to all datasets and metrics. Our model achieves 51.9%, 39.7%, 7.3%, 6.0% higher F1-score compared to STAGG, Yu et al. (2017), Sorokin et al. (2018), Maheshwari et al. (2019) on WebQSP. Analogously, we achieve 40.6%, 33.8%, 22.9%, 7.5% higher F1-score on QALD-7. We can conclude that our double-channel representation method performs better than all baselines. We observe that ignoring either word or sememe, perform worse than the double-channel settings. The comparison demonstrates that two-channel representation preserve the complementary.

4 Conclusion

In this poster, we present a sememe-based approach to differentiate relations with similar semantics in KBQA, where sememe can be leveraged as the minimal semantics of words as an extra natural knowledge to enrich semantics for parsing. In future work, we are interested in maximizing the sememe-level semantics in overtaking the weakness of the word-level semantics in KBQA.

5 Acknowledgments

This work is supported by the National Key Research and Development Program of China (2017YFC0908401) and the National Natural Science Foundation of China (61972455). Xiaowang Zhang is supported by the Peiyang Young Scholars in Tianjin University (2019XRX-0032).

References

1. Sorokin, D., Gurevych, I.: Modeling semantics with gated graph neural networks for knowledge base question answering. In: *COLING'2018*, pp.3306–3317.
2. Yih, W., Chang, M., He, X., Gao, J.: Semantic parsing via staged query graph generation: question answering with knowledge base. In: *ACL'2015*, pp.1321–1331.
3. Bloomfield, L.: A set of postulates for the science of language. Language, 1926, **2**(3): 153-164.
4. Maheshwari, G., Trivedi, P., Lukovnikov, D., Chakraborty, N., Fischer, A., Lehmann, J. (2019). Learning to rank query graphs for complex question answering over knowledge graphs. In: ISWC'19, pp.487-504.
5. Yu, M., Yin, W., Hasan,K.S., Santos, C.N., Xiang,B., Zhou,B.: Improved neural relation detection for knowledge base question answering. In: *ACL'2017*, pp.571–581.