A Key-Value Memory Network of Knowledge Base using Attentive Query

Wenting Qi wq244@nyu.edu

Zhiyuan Wang zw1946@nyu.edu

Yiyi Zhang yz2092@nyu.edu

Weicheng Zhu wz727@nyu.edu

Abstract

Question classification (QC) can be recognized as an effective method of improving Question Answering (QA). Unfortunately, for certain domain-specific tasks, for instance medical diagnosis, even QC suffers from its inherent conundrums like data sparsity for rare diseases. Leveraging external knowledge base (KB) to finesse such issues has proven effective, though traditional KB structures impose strong restrictions on both the source and the schema of KB. Our paper introduces **AKB-MN**, a novel BiLSTM-based model that leverages external knowledge, which effectively incorporates continuous representation from KB using attention mechanism as a query on our key (symptom) and value (disease) pairs to flexibly integrate with the representation of input text. To compare its performance, we evaluate our model on a stratified test set, with a focus on the per-label accuracy, and it achieves remarkable improvements on most minority labels.

1 Introduction

As a direction in improving existing question answering (QA) models, question classification (QC) is a studied task in NLP. Recent Neural-Network (NN) models achieve near state-of-the-art performance on a variety of tasks including QC. However, these models fail to deliver satisfactory performance when it comes to certain domain-specific problems. Particularly, in a genre such as medical diagnosis, traditional models might suffer from data sparsity issues where, even for large dataset, rare diseases have only a handful of samples. Thus, it would be difficult for the model to accurately capture characteristics of those diseases. In light of these observations, introducing external knowledge base (KB) has been proven effective to improve existing NN QC models for domain-specific tasks. Yet traditional KB usually imposes strong restrictions on the source and the schema of databases, causing its implementation to require tedious task-specific engineering.

We propose a new model, **A**ttentive-**K**nowledge-**B**ased **M**emory **N**etwork (**AKB-MN**), for effectively calibrating QC by incorporating attention and key-value memory within KB. We first establish a dynamic interaction between inputs and symptoms via attention mechanism. Furthermore, using the interaction as a query, we design a key-value (symptom-disease) network with deterministic memory to better incorporate KB guidance in our task. Our model structure imitates the way how humans seek help from authorized external resources and how context is stored in memory.

In our work, we investigate QC architectures using medical QA logs crawled from HealthTap with more than 1.6M records, where each record includes details like question (our input), answer, category (225 distinct target labels), related topics, etc.

2 Background

Question classification (QC) (Harabagiu et al., 2000; Hermjakob, 2001; Li and Roth, 2002) has become an approach in improving open-domain question answering (QA) (Harabagiu et al., 2000; Light et al., 2001) system. As per Li and Roth (2002), “locating an answer accurately hinges on first filtering out a wide range of candidates based on some categorization of answer types.” Such rational applies to certain domain-specific QA where a significant amount of sub-categories span the entire domain. For instance, Medical QA often involves knowledge of diseases that can be classified into at least over 100 categories according to NCBI BioSystems Database (Marchler-Bauer et al., 2009) and thus can be improved via QC as an interim step.

Meanwhile, as neural network develops rapidly these years, neural-network (NN) based models has achieved groundbreaking performance on various tasks (Glorot et al., 2011; Kim, 2014; Chen et al., 2014). One of its superiority is "revealing the true explanatory factors of the observed data" via latent variables (Glorot et al., 2011). In particular, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), a special kind of recurrent network (RNN) architecture, addresses the long-term dependency problem, that is essential in many language understanding tasks (Mesnil et al., 2013; Ghosh et al., 2016) but traditional RNN fails to deal with (Hochreiter and Schmidhuber, 1997; Bengio et al., 1994). In this work, we employ a slight variant of the classic LSTM, named bidirectional LSTM (BiLSTM) (Graves and Schmidhuber, 2005).

However, NN classification models still suffer from inherent conundrums rooted in extreme imbalanced data source (Choi et al., 2017). In order to tackle this issue, involving domain-specific knowledge into the model design has long been recognized (Elman, 1990) as a viable treatment. Primarily, previous works leverage knowledge bases (KBs) in two ways: encoding external knowledge as either discrete features (Ratinov and Roth, 2009; Rahman and Ng, 2011; Miwa and Bansal, 2016) or continuous representations (Yang and Mitchell, 2017). However, according to Yang, "not only do these [discrete] features generalize poorly, but they require task-specific feature engineering to achieve good performance." As a result, we incorporate KBs via continuous representations in our work.

Our proposed model amalgamates Key-Value memory network (Miller et al., 2016) with attention mechanism (Bahdanau et al., 2014; Luong et al., 2015) to form a novel KB structure. Attention has been proven efficient in many deep learning tasks such as machine translation (Bahdanau et al., 2014; Luong et al., 2015), speech recognition (Chorowski et al., 2015) and image captioning (Xu et al., 2015). With attention mechanism, NN-based models can better concentrate on crucial information and retain them dynamically (Kim et al., 2017). Building Key-Value memory network upon attention gives both "greater flexibility to prior knowledge about the task" and "more effective power in the model via nontrivial transforms between key and value" (Miller et al., 2016).

Technically, our attention mechanism acts as a distributional query vector that attempts to align appropriate keys (symptoms) within KB, where these symptoms are then used to infer the marginal probabilities of associated values (diseases) through the pre-determined memory matrix in our KB.

3 Model

In this section, we first introduce our baseline model, BiLSTM, which is prevalent in classification tasks. Then we explain how we construct knowledge base (KB) for our task. Finally, we demonstrate our **AKB-MN** model that extends existing BiLSTM by incorporating KB.

3.1 Bidirectional LSTM

LSTM is a variant of recurrent neural network (RNN) that additionally introduces three gates, forget gate f_t, input gate i_t, and output gate o_t, at each time stamp t to solve potential exploration or vanishing of gradient and lacking long-time dependency of traditional RNN (Hochreiter and Schmidhuber, 1997). At each t, the update rules for LSTM reads:

$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1}) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1}) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1}) \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

where x_t is embedded input; h_t is the state vector; σ is a logistic function; $f_t, i_t, o_t \in (0,1)^d$; W's, U's are different linear layers to be learned and $\odot$ is Hadamard product that acts as a gate switch.

Bidirectional LSTM (Graves and Schmidhuber, 2005) stacks two LSTMs which take input x_t in two directions: forward direction and backward direction respectively. BiLSTM returns vertically stacked outputs of two layers $\left[\overrightarrow{h_t} : \overleftarrow{h_{T-t}}\right]$. Our model encodes questions with $\tilde{h}_T = \text{concat}(\overrightarrow{h_T} \| \overleftarrow{h_1})$ as a summary of the original context.

3.2 Knowledge Base

In order to construct an instructive Knowledge Base (KB), we need a database that provides empirical evidence of how pathological symptoms are linked to various diseases. There exists many standardized medical knowledge bases in the format of Resource Description Framework (RDF) graph, e.g. ICD-9-CM, MIMIC, etc. However, databases like these reveal a classification of diseases and a classification system for diagnosis respectively, with no connections between symptoms and diseases whatsoever.

Zhou et al. (2014) designed a symptom-based network of human diseases (Human Symptoms Disease Network, HSDN) to construct "a comprehensive, high-quality map of disease-symptom relations" by using 7,109,429 medical bibliographic records and related Medical Subject Headings (MeSH) metadata. MeSH is the National Library of Medicine's controlled vocabulary thesaurus, an authorized, comprehensive and well-defined database. HSDN contains the interrelation of how more than 200 symptoms are empirically related to various, broad categories of diseases, and thus is an ideal candidate of our memory matrix in KB.

3.3 Attentive-KB Memory Network

Our model **AKB-MN** extends BiLSTM model with a query to medical KB that provides relational match between diseases and symptoms. Instead of word level matching, our KB introduces embeddings as representation of n symptoms $E_{sym} \in \mathbf{R}^{n \times d}$ and m diseases $E_{dis} \in \mathbf{R}^{m \times d}$ respectively, where d is the size of embedding. The model queries possible symptoms mentioned in question, compressed within representation vector h_T of the sentence, via attention mechanism and induces the representation of potential diseases V. Figure 1 depicts the mechanism how our KB works as a complement to BiLSTM model.

In KB block, each question sentence is encoded by word embedding and BiLSTM as state vectors h_T. Each symptom could be associated with several potential diseases based on the key-value memory we obtained from HSDN. The KB takes h_T as a query on keys E^i_{sym}'s in continuous space $\mathbf{R}^d$ and obtains attention weights α_i for each symptom:

$$e_i(h_T, E^i_{sym}) = v^T \tanh\left(W h_T \| U E^i_{sym}\right)$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{n} \exp(e_i)}$$

where v, W, U are linear weights to be learned and $\|$ denotes concatenation.

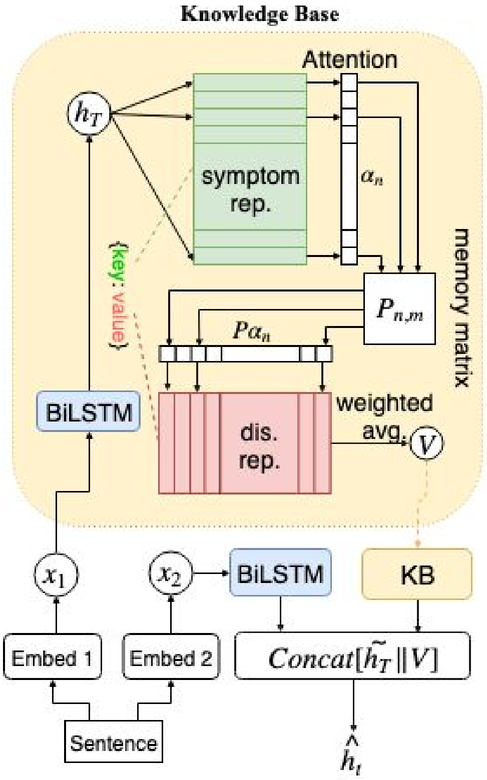

Figure 1: The architecture of **AKB-MN** model. Sentences are encoded with two embedding layes and BiLSTMs. One of state vectors h_t queries medical knowledge and generates V from KB. The other state vector $\tilde{h}_t$, together with V, works as the model output $\hat{h}_T$.

Next, using transitional probabilities $p_{ij} = p(\text{dis.} = j \mid \text{sym.} = i)$ from the key-value memory, we compute the marginal probability on each symptom:

$$p(\text{dis.} = j \mid h_T) = \sum_{i=1}^{n} p_{ij} \cdot \alpha_i.$$

To aggregate the information of diseases, marginal distribution of symptoms is used as weights to average-pool embeddings of diseases into $V \in \mathbf{R}^d$:

$$V = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \alpha_i E^j_{dis}.$$

We then concatenate the resulting V with the state vector $\tilde{h}_T$ of original BiLSTM to form a representation that includes both original textual meanings and medical knowledge from KB:

$$\hat{h}_T = \text{concat}(\tilde{h}_T \| V).$$

Finally we use two fully-connected layers to transform the output representation $\hat{h}_T$ into corresponding scores for classification.

4 Experiments

4.1 Implementation Detail

Stratified Sampling: Our QC task has 225 labels, and the top 10% minority labels consist of only 0.4% of our overall data. To prevent an unrepresentative population, we use the stratified sampling to split our training, validation and test data.

Training & Configuration: We perform grid search to tune our baseline model, and achieve best performance following hyper-parameters: 256 embedding size, 512 hidden dimension, and 4 layers. While training our **AKB-MN** model, in order to make training process more balanced and effective for every components within KB, we put two dropout layers with $p = 0.5$ on the BiLSTM outside KB. The BiLSTM outside has embedding size, hidden dimension, and number of layers of 256, 512, and 2 respectively. While for the BiLSTM inside, these hyper-parameters are 256, 256, and 1. As for our memory matrix within KB, there are 316 symptoms (keys) and 119 diseases (values), each of them is represented by a vector of 256 dimension.

4.2 Results

We take prediction accuracy as our evaluation metric. We compare the **AKB-MN** model with the baseline model basing on the performance of the overall accuracy and per-class accuracy for minority labels to examine the effectiveness of KB module in solving the data imbalance problem.

From Table 1, we observe a 0.9% improvement on the overall test accuracy and a significant improvement of 2.62% on unweighted average accuracy[1] among every class. More improvement on the accuracy calculated by unweighted mean compared to weighted mean indicates a remarkable step-forward on classifying minority labels. These observations demonstrate the effectiveness of leveraging medical knowledge in the **AKB-MN** model.

Overall Label Acc.	**Baseline**	**AKB-MN**
Unweighted Acc.	64.15%	**66.77%**
Weighted Acc.	66.13%	**67.03%**

Table 1: Overall Accuracy on Test Set

We then investigate the performance on rare labels. Most samples of label with low frequency in Table 2 are not well identified by vanilla BiLSTM model. The **AKB-MN** model significantly surpasses BiLSTM on most of these labels. The class accuracy on ”Dupuytren contracture” and ”gas bloating” do not improve by adding KB as they are typical examples of unusual diseases or common physiological phenomena that are not included in our medical KB. Despite some specific categories beyond KB memory, the architecture of KB module functions well in supplementing insufficient training data of minority categories with additional information.

Label Name	**Freq.**	**Baseline**	**AKB-MN**
Hypogonadism	0.025‰	58.33%	66.67%
Dupuytren Contracture	0.041‰	75.00%	75.00%
Nausea	0.052‰	40.00%	56.00%
Charley Horse	0.083‰	67.50%	77.50%
Gas Bloating	0.183‰	55.68%	55.68%
Stomach	0.208‰	39.00%	64.00%
Carbidopa Levodopa	0.392‰	45.74%	58.51%

Table 2: Per-class Accuracy on Selected Labels

5 Conclusions

5.1 Summary

The traditional Neural Network approach of Question Classification Task suffers from data sparsity issues. Our proposed **AKB-MN** model, that has a knowledge base module, is proved effective in improving the accuracy of minority labels. Furthermore, **AKB-MN** is a structure that can be generalized to other tasks with little restrictions on the degree of direct relevance between task and external knowledge, as long as the source provides relationship between informative keys and values.

5.2 Limitation

Our limitations intrinsically come from the qualified design of KB. The memory network we used in our model is a definitive structure with fixed numbers of keys and values. However, the keys and/or values in the network can suffer from incompatibility with the original dataset. Future work can modify this structure of incorporating external knowledge by using a more interactive form so that our keys and values can be adaptively updated.

Moreover, our memory matrix is obtained from analyzing past medical bibliographic records before 2016 and it may need to be regularly maintained up-to-date.

[1] Unweighted Accuracy is calculated as the arithmetic mean of per-class accuracy

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. 2014. Deep learning-based classification of hyperspectral data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7:2094–2107.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795. ACM.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585.

Jeffrey L. Elman. 1990. Finding structure in time. *COGNITIVE SCIENCE*, 14(2):179–211.

Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry P. Heck. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *CoRR*, abs/1602.06291.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 513–520, USA. Omnipress.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Sanda Harabagiu, Dan Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. 2000. Falcon: Boosting knowledge for answer engines. pages 479–488.

Ulf Hermjakob. 2001. Parsing and question classification for question answering. In *Proceedings of the Workshop on Open-domain Question Answering - Volume 12*, ODQA '01, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *arXiv preprint arXiv:1702.00887*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marc Light, Gideon S. Mann, Ellen Riloff, and Eric Breck. 2001. Analyses for elucidating current question answering technology. *Nat. Lang. Eng.*, 7(4):325–342.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Aron Marchler-Bauer, Chunlei Liu, Jane He, Lewis Y. Geer, Lianyi Han, Renata C. Geer, Siqian He, Stephen H. Bryant, and Wenyao Shi. 2009. The NCBI BioSystems database. *Nucleic Acids Research*, 38:D492–D496.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*.

Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. *CoRR*, abs/1606.03126.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 814–824, Portland, Oregon, USA. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

Bishan Yang and Tom Mitchell. 2017. Leveraging knowledge bases in lstms for improving machine reading. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).*

XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. 2014. Human symptoms–disease network. *Nature communications*, 5:4212.

A Appendix

Github Repositories
Source code:
https://github.com/jackzhu727/NLU-final-project
HSDN Dataset:
https://github.com/jackzhu727/hsdn

Contribution Guide
Wenqing Qi:
Research on KB, model implementation
Zhiyuan Wang:
Related work research, model architecture
Yiyi Zhang:
Related work research, research on KB
Weicheng Zhu:
Model architecture, model implementation