A Multi-Task Learning Approach for Answer Selection: A Study and a Chinese Law Dataset*

Wenyu Du,[1,2] **Baocheng Li,**[2,3] **Min Yang,**[1] **Qiang Qu,**[1] **Ying Shen**[4]
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]Westlake University, [3]Northeast Normal University, [4] Peking University Shenzhen Graduate School

Abstract

In this paper, we propose a Multi-Task learning approach for Answer Selection (MTAS), motivated by the fact that humans have no difficulty performing such task because they possess capabilities of multiple domains (tasks). Specifically, MTAS consists of two key components: (i) A category classification model that learns rich category-aware document representation; (ii) An answer selection model that provides the matching scores of question-answer pairs. These two tasks work on a shared document encoding layer, and they cooperate to learn a high-quality answer selection system. In addition, a multi-head attention mechanism is proposed to learn important information from different representation subspaces at different positions. We manually annotate the first Chinese question answering dataset in law domain (denoted as LawQA) to evaluate the effectiveness of our model. The experimental results show that our model MTAS consistently outperforms the compared methods.[1]

Introduction

Law Community Question Answering (CQA) forums are gaining popularity online since it offers a new opportunity for individuals to get free legal advice directly from experienced lawyers and users. A question often has hundreds of answers, which makes it time consuming for users to inspect the high-quality answers. Thus, it is essential that we have automatic answer selection techniques to select good answers to new questions in a community-created discussion forum. In the literature, answer selection have been extensively studied in the last decade using both non-neural approaches (Wang and Manning 2010) and neural ones (Yu et al. 2014).

Despite the effectiveness of previous studies, answer selection remains a challenge since conventional methods still have several drawbacks. (1) Prior answer selection approaches basically apply a uniform model for the questions from different text categories. However, according to what we observe, the answer styles in different categories can vary to a large degree in law domain. (2) Existing studies often rely on a single attention function to capture important parts of the input. However, in different representation subspaces, the important information may appear at different positions (Vaswani et al. 2017). (3) There is no publicly available benchmark for CQA in the law domain.

In this study, all the aforementioned limitations are considered and alleviated to some extent. MTAS simultaneously optimizes two coupled objectives: text categorization and answer selection, in which a document modeling module is share across tasks. The main purpose of our multi-task model is to strengthen the representation learning of questions, and safeguard the performance of answer selection in the scale corpus. To capture the comprehensive semantics of the whole input sequence, we employ a multi-head attention mechanism to focus on important information that may appear at different positions according to different representation subspaces. To empirically demonstrate the effectiveness of our approach, we created a Chinese data set (LawQA) in law domain by collecting question and answer pairs from a Chinese law forum.

The main contribution of our approach is three-fold: (i) we design a novel multi-task framework with multi-head attention mechanism for answer selection. (ii) we create the first Chinese law dataset (LawQA). The release of it would push forward the research in this field. (iii) the experimental results demonstrate the effectiveness of MTAS in answer selection.

Our Model

Given a question q, our model aims to rank a set of candidate answers $A = \{a_1, \ldots, a_n\}$. MTAS jointly trains two related tasks: answer selection (primary task) and text categorization (auxiliary task). Next, we will elaborate these two tasks in details.

BiLSTM First, we employ a word embedding layer to convert each word w into a low-dimensional vector e^w. Then, we use a BiLSTM to learn the hidden states of words in the question and answer. Formally, given the input word embedding e_t at index t in the document, the forward and backward hidden states $\overrightarrow{h}_t \in \mathbb{R}^u$ and $\overleftarrow{h}_t \in \mathbb{R}^u$ can be updated as:

$$\overrightarrow{h}_t = \overrightarrow{LSTM}(\overrightarrow{h}_{t-1}, e_t), \quad \overleftarrow{h}_t = \overleftarrow{LSTM}(\overleftarrow{h}_{t-1}, e_t) \quad (1)$$

*M. Yang is corresponding author (min.yang@siat.ac.cn). This work is supported by CAS Pioneer Hundred Talents Program.

[1]Source code: https://github.com/Ange1o/MTAS-LawQA

We concatenate the forward and backward vectors to form the final hidden state $h_t^i = [\overrightarrow{h}_t^i || \overleftarrow{h}_t^i]$ at time step t. Thus, we can use the Bi-LSTM network to obtain the hidden states $H^q = [h_1^q, \ldots, h_m^q]$ and $H^a = [h_1^a, \ldots, h_n^a]$ for the question and answer, respectively.

Multi-head attention We use multi-head attention mechanism to model the semantics of answers over questions. Specifically, given the output representation of the question (h_t^q) and answer (h_t^a) at time step t, we have:

$$A_t = \exp(W_m m_t), \;\; \hat{h}_t^a = flatten(A_t h_t^a) \quad (2)$$

$$m_t = tanh(W_a h_t^a + W_q h_t^q) \quad (3)$$

Where $\hat{h}^a$ is the answer representation after multihead attention, W_a, W_q, and W_m are weight parameters to be learned. $A_t \in \mathbb{R}^{b \times m}$ is the attention matrix, where b is the number of hops of attention. $flatten$ is an operation that flattens matrix into vector form. Here, we set $b = 4$.

Answer Selection Task The cosine similarities between the final representations of the question and the answer will then be calculated. Following the ranking loss in (dos Santos et al. 2016), we define the training objective as a hinge loss:

$$L_1 = max\{0, M - cosine(q, a_+) + cosine(q, a_-)\} \quad (4)$$

where a_+ is a ground truth answer, a_- is a randomly chosen incorrect answer, and M is a constant margin.

Text Categorization Task Text categorization is an auxiliary task that helps to learn better category-aware text representations. Text categorization and answer selection tasks share the same BiLSTM and Multi-head attention networks.

We feed the representations of question (i.e., H^q) into a two-layer fully-connected network and a softmax layer to obtain the predicted text category.

$$f = \tanh(V_1 H^q), \quad \hat{y} = \text{softmax}(V_2 f) \quad (5)$$

where V_1 and V_2 are projection parameters. We minimize directly the cross-entropy between the predicted label distribution $\hat{y}$ and the ground truth distribution y as the objective function:

$$L_2 = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \mathbf{I}(y = j) \log(\hat{y}) \quad (6)$$

where $\mathbf{I}(\cdot)$ is an indicator such that $\mathbf{I}(\text{true}) = 1$ and $\mathbf{I}(\text{flase}) = 0$. C is the category number and N is the number of questions in the corpus.

Joint training Overall, our model consists of two subtasks, each has a training objective. For the purpose of strengthening the learning of the share document-query representations, we train these two related task simultaneously. The joint multi-task objective function is minimized by:

$$L = (1 - \alpha) * L_1 + \alpha * L_2 \quad (7)$$

where α is the hyper-parameter that determine the weights of L_1 and L_2. Here, we set $\alpha = 0.1$ via cross validation.

Experiments

Experimental Data We collect a large LawQA dataset that contains 40,000 questions from 10 balanced categories and 72,416 positive QA pairs in law domain. We also manually collect negative samples by randomly selecting one answer from the other categories. Totally, there are 144,832 QA pairs for training. We randomly select 2000 QA pairs for testing and 1000 QA pairs for validation.

Implementation Details In our experiments, all word embeddings are initialized by a 150 dimension word2vec model. All the weights are given their initial values by sampling form a truncated normal distribution $N(0, 0.1)$. The hidden size of BiLSTM and attention size are set to 1000 and 300 respectively. We perform a 4-head attention on the answer representations.

Experimental Results We evaluate our model with several strong competitors: CNN (Yu et al. 2014), BiLSTM (Tan et al. 2016),Bi-LSTM-attention (Tan et al. 2016), IARNN-word (Wang, Liu, and Zhao 2016), AP-LSTM (dos Santos et al. 2016). Top-1 accuracy, MAP and MRR are used as the evaluation metrics for answer selection. The experimental results are summarized in Table 1. From Table 1, we observe that our model performs better than the compared methods. We also report the ablation test of MTAS in terms of discarding text categorization task. Text categorization contributes great improvement to MTAS. This is within our expectation since missing category information will lead the answer selection unspecific.

	Top1 Acc	MAP	MRR
CNN	0.521	0.569	0.640
Bi-LSTM	0.561	0.601	0.674
Bi-LSTM-attention	0.573	0.619	0.688
IARNN-word	0.534	0.584	0.657
AP-LSTM	0.556	0.591	0.669
MTAS w/o multitask	0.577	0.622	0.691
MTAS (Ours)	**0.588**	**0.636**	**0.700**

Table 1: Experiment result on answer selection task

References

dos Santos, C. N.; Tan, M.; Xiang, B.; and Zhou, B. 2016. Attentive pooling networks. *CoRR, abs/1602.03609* 2(3):4.

Tan, M.; Santos, C. d.; Xiang, B.; and Zhou, B. 2016. Lstm-based deep learning models for non-factoid answer selection. *ICLR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Wang, M., and Manning, C. D. 2010. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *COLING*, 1164–1172.

Wang, B.; Liu, K.; and Zhao, J. 2016. Inner attention based recurrent neural networks for answer selection. In *ACL*.

Yu, L.; Hermann, K. M.; Blunsom, P.; and Pulman, S. 2014. deep learning for answer sentence selection. In *NIPS*.