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

Supplimental Material

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].

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. It is not unusual for a question to have hundreds of answers, which makes it time consuming for users to inspect and winnow the high-quality answers. Thus, it is essential that we have automatic techniques to select good answers to new questions in a community-created discussion forum.

Answer selection, which is a key component of question answering (QA), has attracted increasing attention recently due to its broad applications in natural language processing and information retrieval, such as factoid question answering [Voorhees and Tice, 2000] and community-based question answering [Bian et al., 2008]. Given a question, answer selection aims to pick out the most relevant answer from a set of candidates. In the literature, answer selection have been extensively studied in the last decade using both non-neural approaches [Wang et al.,

[1] Data and codes are available at: https://github.com/Ange1o/MTAS-lawqa

2007, Wang and Manning, 2010, Yih et al., 2013] and neural ones [Yu et al., 2014, Dos Santos et al., 2015, Yin et al., 2016, Dong et al., 2017, Tymoshenko et al., 2017].

Despite the effectiveness of previous studies, answer selection remains a challenge since conventional methodologies 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. The document representations should focus on different aspects of the topics which belong to the corresponding categories. When the question category is given, a category-aware text representation can largely promote the performance of answer selection. (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]. All the information forms the comprehensive semantics of the whole input sequence, especially for long documents, which usually happen in the law domain. (3) There is no publicly available benchmark for community question answering in the law domain.

In this study, all the aforementioned limitations are considered and alleviated to some extent. Our method, Multi-Task learning approach for Answer Selection (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. We find that, though our final goal is answer selection, a good text categorization model could help the semantics analysis and comprehension of documents by learning robust category-aware text representations. 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. LawQA is partitioned into 10 different categories, each of which corresponds to a specific category. To the best of our knowledge, it is the first CQA dataset in the law domain. The release of it would push forward the research in this field. We conduct experiments on this corpus, and the results shows that our model outperforms the compared methods. In addition, LawQA is a challenging dataset since the Top 1 accuracy and MAP values are less than 70% for the best benchmark.

2 Data Collection

2.1 Process

We elaborate the process of generating our LawQA dataset in this section. Firstly, we collect a large pool of law related QA pairs with categorical information from the forum[2]. All the questions asked by netizens will be answered by the licensed lawyers. The questions are divided into ten categories, including Violence, Traffic accident, Medical accident, Consumer Rights, Property disputes, Bank, Criminal defense, Divorce, Inheritance and Labor Contract.

Then, we remove the redundant QA pairs, and set the minimum length of question and answer to be 14 characters, to avoid the vagueness in the text. Our resized QA dataset contains 10 balanced categories with 40,000 questions. Since one question may have multiple answers, we have a clean QA dataset with overall 72,416 positive QA pairs. Table 1 shows one pair of LawQA.

继承 Inheritance	我和丈夫的财产，丈夫前妻的儿子有权继承我和丈夫的财产吗? Does the son of my husband's ex-wife have the right to inherit property from me and my husband?
	你好，你丈夫前妻的儿子若是你丈夫与前妻的共同子女，有权继承你丈夫的遗产。 If the son of your husband's ex-wife is also your husband's son, he will have the right to inherit legacy from your husband.

Table 1: An example of LawQA pair

To build the training set for answer selection, we manually collect negative samples by randomly selecting one answer form another category to form the negative sample for each QA pair (positive sample). Finally, we have a 144,832 QA pairs for training. In terms of testing set, for each distinct question, we set the candidate pool to be 100.

2.2 Statistics and analysis

Table 2 shows the statistics of LawQA dataset and makes a comparison to three other answer selection datasets, TrecQA [Robertson et al., 1996], WikiQA [Yang et al., 2015] and InsuranceQA [Feng et al., 2015].

In legal domain, one key challenge is the domain gap between question and answer. The askers are netizens without much legal knowledge and questions are quite ambiguous and often in informal or oral language. On the contrary, these questions are answered by well-trained professionals and they answer purely in formal and written language. This phenomenon is extremely noticeable in the legal

[2] http://china.findlaw.cn/

Dataset	Train	Dev	Test	Avg len of Q	Avg len of A
LawQA	144,832	1000	2000	45.39	30.92
InsuranceQA	18,540	1000	50	7.16	49.5
TrecQA	1162	65	38	11.39	30.39
WikiQA	873	126	9	7.18	25.15

Table 2: Statistics of the LawQA dataset and other answer selection datasets

消费者权益 Consumer Rights	被美容院忽悠办卡，不给退怎么办 (I was) fooled by a beauty salon to buy its membership, what should (I) do if they don’t give my money back (?)
	可以协商处理，协商不成可以到 工商局投诉，或者到法院起诉。 You can try negotiation first, then you could complain to Administration for Industry and Commerce or prosecute via court.

刑事辩护 Criminal Defense	打电话威胁要20万，随即报警，应该构成什么罪 A phone call threatened (me) for 200 thousand (Yuan), (and I) called police immediately. Which crime did it be (?)
	需要结合实际情节，涉嫌构成敲诈勒索罪或绑架罪。 Suspicion of Extortion or Kidnapping. Detailed information is required.

Table 3: Two examples of LawQA pairs - Different language styles between QA

scenario. Table 3 shows two question-answer pairs that questions are in informal language while answers are in formal language from legal experts.

3 Our Model

Our model MTAS, whose architecture is illustrated in Figure 1, jointly trains two related tasks: answer selection (primary task) and text categorization (auxiliary task). Next, we will elaborate these two tasks in details.

3.1 Answer selection

Given a question q, our model aims to rank a set of candidate answers $A = \{a_1, \ldots, a_n\}$.

3.1.1 Word embedding

Firstly, we employ a word embedding model to compact each word into a distributed embedding. Each word w in the corpus is mapped to a low-dimensional word embedding e^w through a word embedding layer.

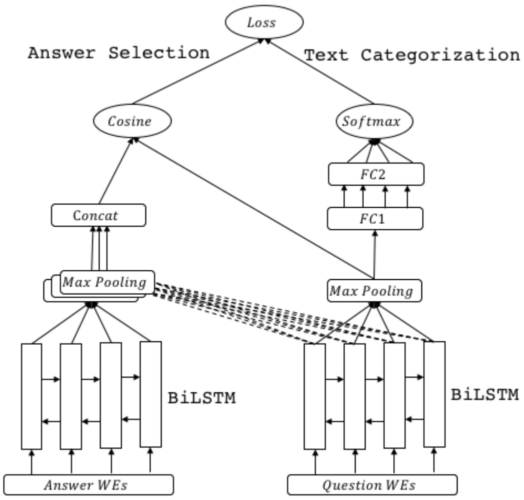

Figure 1: The architecture of MTAS

3.1.2 BiLSTM

Then, we use a BiLSTM [Hochreiter and Schmidhuber, 1997] 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 bu update as:

$$\overrightarrow{h}_t^i = \overrightarrow{LSTM}(\overrightarrow{h}_{t-1}^i, e_t^i) \tag{1}$$

$$\overleftarrow{h}_t^i = \overleftarrow{LSTM}(\overleftarrow{h}_{t-1}^i, e_t^i) \tag{2}$$

The concatenation of forward and backward vectors 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.

3.1.3 Multi-head attention

We use multi-head attention mechanism to model the semantics of answers over questions, which helps to capture the important information from different representation subspaces at different positions. Specifically, given the output representation of the question (h_t^q) and answer (h_t^a) at time step t, we have:

$$m_t = tanh(W_a h_t^a + W_q h_t^q) \tag{3}$$

$$A_t = \exp(W_m m_t) \tag{4}$$

$$\hat{h}^a = flatten(A_t h^a) \tag{5}$$

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 $b = 4$.

3.1.4 Answer Selection Objective

Finally, the question representations h^q and the attended answer representations $\hat{h}^a$ will be fed through a max-pooling layer. The cosine similarities between the question and the answer will then be calculated. Following the same ranking loss in [Weston et al., 2014, Hu et al., 2014, Feng et al., 2015], we define the training objective as a hinge loss:

$$L_1 = max\{0, M - cosine(q, a_+) + cosine(q, a_-)\} \tag{6}$$

where a_+ is a ground truth answer, a_- is an incorrect answer randomly chosen from the entire answer space, and M is a constant margin.

3.2 Text Categorization

Text categorization is an auxiliary task to help learn better category-aware text representations. This task can be seen as a multi-class classification problem. Text categorization and answer selection model share the same BiLSTM and Multi-head attention networks with the answer selection task.

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.

$$f = \text{tanh}(V_1 H^q), \quad \hat{y} = \text{softmax}(V_2 f) \tag{7}$$

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:

$$L_2 = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \mathbf{I}(y = j) \log(\hat{y}) \tag{8}$$

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.

3.3 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:

$$L = (1 - \alpha) * L_1 + \alpha * L_2 \tag{9}$$

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.

4 Experiments

4.1 Evaluation Metrics

To quantitatively evaluate the proposed model on the LawQA dataset, we adopt three most widely used evaluation metrics for answer selection, including Top-1 accuracy, MAP (Mean Average Precision) and MRR (Mean Reciprocal Rank).

4.2 Implementation Details

In our experiments, all word embeddings are initialized by a 150 dimension word2vec [Mikolov et al., 2013] 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. To train our model, we perform a mini batch gradient descent with batch size as 512 and clip the gradient d when $d > 5$.

4.3 Baselines

We evaluate and compare our model with several strong competitors:

CNN [Yu et al., 2014]: A bi-gram convolutional neural network to model the text.

Bi-LSTM [Tan et al., 2015]: A BiLSTM network to learn the representations.

Bi-LSTM-attention [Tan et al., 2015]: A BiLSTM network followed by an attention mechanism.

IARNN-word [Wang et al., 2016]: An attention mechanism before a BiLSTM network.

AP-LSTM [dos Santos et al., 2016]: A combined attention and pooling mechanism after a BiLSTM network.

4.4 Experimental Results

The experimental results are summarized in Table 4. From Table 4, we observe that our model performs better than the compared methods. For example, MTAS improves the Top1 accuracy from 0.573 to 0.588. The similar trends are observed on MAP and MRR metrics. To investigate the effectiveness of our multi-task learning, we also report the ablation test of MTAS in terms of discarding text categorization task (denoted as MTAS w/o multitask). Text categorization contributes great improvement to MTAS. This is within our expectation since missing category information will lead the answer selection unspecific. Category information is vital when selecting the best answer for the given question.

	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 4: Experiment result on answer selection task. Numbers with * mean that improvement is statistically significant over the baseline methods(t-test, p-value¡0.05).

4.5 Text classification results

We report the text classification results in Table 5. From the results, we can observe that the results co-trained with answer selection task is better than the results of only performing text classification.

	Precision	Recall	F1
Only text classification	0.789	0.787	0.777
MTAS	0.801	0.800	0.794

Table 5: Text classification results on LawQA.

4.6 Experiments on α

Table 6 shows the experimental results on α. With a larger α, the train phase will pay more attention on the loss in text classification task and a smaller α causes answer selection task plays a more important role in the whole train phase. Because the main goal is answer selection task, we found 0.1 is the optimum that focuses more on the answer selection task and still contains enough text categorical information.

α	0.01	0.02	0.05	0.1	0.2	0.5
Top1 acc	0.581	0.581	0.588	0.588	0.567	0.540
MAP	0.626	0.625	0.630	0.636	0.617	0.594
MRR	0.691	0.692	0.697	0.700	0.683	0.660

Table 6: Experiment on hyperparameter α

4.7 Error analysis

We randomly select 100 samples that are incorrectly predicted by our model from the test set. We observe that these samples are imbalanced across categorical domains. As shown in Table 7, some question types like Consumer Rights and Bank are more difficult than others like Inheritance and Medical Accident. We argue the reason may be because that the questions in Bank and Consumer Rights are more easily confused with other categories. We showed two examples in the case study section.

Consumer Rights	14	Traffic Accident	9
Bank	14	Property Distributes	9
Divorce	12	Labor Contract	8
Criminal Defense	11	Inheritance	7
Violence	10	Medical Accident	6

Table 7: Incorrect samples statistics.

4.8 Case study

Table 8 gives one example for each Consumer Rights and Bank categories, where the samples are easily confused with other categories. Although the first case is annotated in Consumer Rights, it is also plausible to be classified into Property Disputes. Similarly, the second case can also be easily mismatched to Criminal Defense despite its annotation as Bank.

5 Conclusion

In this paper, we proposed a multi-task framework for answer selection, which treated text categorization and answer selection as two subtasks, strengthening the representation learning in document modeling. We also created a new QA task in the law domain to evaluate the effectiveness of our model and will release this new corpus to push forward the research in this field. The experiments demonstrate the superiority of MTAS.

References

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. Finding the right facts in the crowd: Factoid question answering over social media. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 467–476, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2.

消费者权益 Consumer Rights	武汉**购房后交付定金4万，后因资金周转原因欲退订，但房产公司答复是定金没收不退，请问如何才能挽回损失 (I) booked an estate in Wuhan and paid 40 thousand as deposit, but later I wanted to cancel the deal due to capital turnover. But the real-estate company replied deposit was nonrefundable, so how can I retrieve the loss (?)
	如果是定金，交付一方不履行合同的，无权请求返还。 If one party does not fulfill the contract, this party has no right to ask for refunding deposit.

(a) Easily mismatched to Property Disputes

银行 Bank	信用卡恶意透支300万元无力偿还会判多少年? My credit card was maliciously overdrafted 3 million Yuan, (and I) was unable to pay off. How many years will I be sentenced to prison?
	你这个情节属于特别严重，可能会判刑10年以上的，建议立即委托律师处理看是不是有减轻情节。 Your situation is extremely severe and could be sentenced to over 10 years in jail. Suggest to consult an attorney immediately to see if there are any mitigating circumstances.

(b) Easily mismatched to Criminal Defense

Table 8: Two examples of LawQA pairs - Consumer Rights and Bank categories

doi: 10.1145/1367497.1367561. URL http://doi.acm.org/10.1145/1367497.1367561.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, 2017.

Cicero Dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. Learning hybrid representations to retrieve semantically equivalent questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 694–699, 2015.

Cıcero Nogueira dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. Attentive pooling networks. *CoRR, abs/1602.03609*, 2(3):4, 2016.

Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. Applying deep learning to answer selection: A study and an open task. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 813–820. IEEE, 2015.

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

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. pages 109–126, 1996.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*, 2015.

Kateryna Tymoshenko, Daniele Bonadiman, and Alessandro Moschitti. Ranking kernels for structures and embeddings: A hybrid preference and classification model. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 897–902, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 200–207, New York, NY, USA, 2000. ACM. ISBN 1-58113-226-3. doi: 10.1145/345508.345577. URL http://doi.acm.org/10.1145/345508.345577.

Bingning Wang, Kang Liu, and Jun Zhao. Inner attention based recurrent neural networks for answer selection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1288–1297, 2016.

Mengqiu Wang and Christopher D Manning. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1164–1172. Association for Computational Linguistics, 2010.

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

Jason Weston, Sumit Chopra, and Keith Adams. # tagspace: Semantic embeddings from hashtags. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1822–1827, 2014.

Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, 2015.

Wen-tau Yih, Ming-Wei Chang, Christopher Meek, and Andrzej Pastusiak. Question answering using enhanced lexical semantic models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1744–1753, 2013.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *TACL*, 4:259–272, 2016.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. deep learning for answer sentence selection. In *Proceedings of Deep Learning and Representation Learning Workshop*. NIPS, 2014.