A Review on Deep Learning Techniques Applied to Answer Selection

Tuan Manh Lai [1], Trung Bui [2], Sheng Li [3]

[1] Purdue University, [2] Adobe Research, [3] University of Georgia

Introduction

- Answer selection is an important problem in NLP, with applications in many areas: Given a question and a set of candidate answers, the task is to identify which of the candidates contains the correct answer to the question.
- Many deep learning methods have been proposed for the task. They produce impressive performance without relying on any feature engineering or expensive external resources.
- In this work, we aim to provide a comprehensive review on deep learning methods applied to answer selection. In addition, we examine the most popular datasets and the evaluation metrics for the task

Overview

- Existing deep learning methods for answer selection can be examined along two dimensions: (i) *learning approaches* (pointwise, pairwise, and listwise) (ii) *neural network architectures* (Siamese architecture, Attentive architecture, or Compare-Aggregate architecture).

Method	Learning Approach	Model Architecture	MAP (Raw TrecQA)	MAP (Clean TrecQA)
TRAIN-ALL unigram+count (Yu et al., 2014)	Pointwise	Siamese	0.693	-
TRAIN-ALL bigram+count (Yu et al., 2014)	Pointwise	Siamese	0.711	-
QA-LSTM (Tan et al., 2015)	Pairwise	Siamese	-	0.682
QA-LSTM with attention (Tan et al., 2015)	Pairwise	Attentive	-	0.690
QA-LSTM/CNN (Tan et al., 2015)	Pairwise	Siamese	-	0.706
Attentive Pooling CNN (dos Santos et al., 2016)	Pairwise	Attentive	-	0.753
(Severyn and Moschitti, 2015)	Pointwise	Siamese	0.746	-
L.D.C Model (Wang et al., 2016b)	Pointwise	Compare-Aggregate	-	0.771
Pairwise Word Interaction Modelling (He and Lin, 2016)	Pointwise	Compare-Aggregate	0.758	-
Multi-Perspective CNN (He et al., 2015)	Pointwise	Siamese	0.762	0.777
HyperQA (Hyperbolic Embeddings) (Tay et al., 2018a)	Pairwise	Siamese	0.770	0.784
PairwiseRank+Multi-Perspective CNN (Rao et al., 2016)	Pairwise	Siamese	0.780	0.801
BiMPM (Shen et al., 2017)	Pointwise	Compare-Aggregate	-	0.802
Dynamic-Clip Attention (Bian et al., 2017)	Listwise	Compare-Aggregate	-	0.821
IWAN (Shen et al., 2017)	Pointwise	Compare-Aggregate	-	0.822
IWAN+CARNN (Tran et al., 2018)	Pointwise	Compare-Aggregate	-	0.829
MCAN (Tay et al., 2018b)	Pointwise	Compare-Aggregate	-	0.838

Table 1: An overview of many of the existing deep learning methods for answer selection. TrecQA is a widely used dataset for benchmarking different answer selection systems. There are two versions of TrecQA: Raw TrecQA and Clean TrecQA. Mean Average Precision (MAP) is a standard evaluation metric in Information Retrieval and Question Answering.

Learning Approaches

- The answer selection problem can be formulated as a ranking problem, where the goal is to give better rank to the candidate sentences that are relevant to the question.
- There are three most common approaches to learn the ranking function h_θ, namely, pointwise, pairwise, and listwise (Liu, 2011).
- In the **pointwise approach**, the ranking problem is transformed to a binary classification problem. More specifically, the training instances are triples (q_i, c_{ij}, y_{ij}), where q_i is a question in the dataset, c_{ij} is a candidate answer for q_i, and y_{ij} is a binary value indicating whether c_{ij} is correct. It is enough to train a binary classifier: $h_\theta(q_i, c_{ij}) \rightarrow \hat{y}_{ij}$, where $0 \leq \hat{y}_{ij} \leq 1$. During inference, given a question, the trained classifier h_θ is used to rank every candidate sentence, and the top-ranked candidate is selected (i.e., $argmax_{cij}\, h_\theta(q_i, c_{ij})$ should be selected as the answer to q_i).
- The second approach to ranking is the **pairwise approach**, where the ranking function h_θ is explicitly trained to score correct candidate sentences higher than incorrect sentences. For example, in (Feng et al., 2015), the training instances are triples (q_i, c_i^+, c_i^-) where q_i is a question, c_i^+ is a correct sentence for q_i, and c_i^- is an incorrect sentence sampled from the whole candidate sentence space. And the hinge loss function is defined as follows:

$$L = \max\{0, m - h_\theta(q_i, c_i^+) + h_\theta(q_i, c_i^-)\}$$

where m is the margin. Basically, the loss function is designed to encourage the correct answer to have a higher score than the incorrect answer by a certain margin.

- The pointwise approach and the pairwise approach ignore the fact that answer selection is a prediction task on list of candidate sentences. In the **listwise approach**, a single training instance consists of a question and its list of candidates.
- Even though many work adopted the pointwise approach, this approach is not close to the nature of ranking. The pairwise approach and the listwise approach exploit more information about the ground truth ordering of candidate sentences.
- (Rao et al., 2016) proposed a pairwise ranking approach that can directly exploit existing pointwise neural network models as base components. The approach outperforms many previous competitive pointwise baselines. (Bian et al., 2017) showed that the listwise approach performs better than the pointwise approach on public datasets such as TrecQA (Wang et al., 2007) and WikiQA (Yang et al.,2015).

Neural Network Architectures

- There are three main types of general architectures for measuring the relevance of a candidate sentence to a question, namely, Siamese architecture, Attentive architecture, Compare-Aggregate architecture.
- **Siamese Architecture**. In a Siamese architecture (Bromley et al., 1993), the same encoder (e.g., a CNN or a RNN) is used to build the vector representations for the input sentences (i.e., the candidate answer and the question) individually. After that, the relevance score is determined solely based on the encoded vectors. There is no explicit interaction between the input sentences during the encoding process.

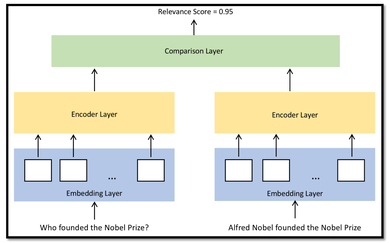

Figure 1: The general architecture of a Siamese model. Despite its conceptual simplicity, a disadvantage is the absence of explicit interaction between the input sentences during the encoding process. A question is always mapped to the same vector regardless of the candidate answer in consideration, and vice versa.

- **Attentive Architecture**. Instead of generating representations for the candidate answer and the question independently, attention mechanisms can be used to allow the information from an input sentence to influence the computation of the other's representation. For example, Tan et al. (2015) proposed a basic model with attention (Figure 2). The model employs a biLSTM network and a pooling layer to generate the question representation $\boldsymbol{o_q}$. The candidate representation $\boldsymbol{o_c}$ is calculated similarly, except that prior to the pooling layer, each biLSTM output vector will be multiplied by a weight, which is determined by the question representation $\boldsymbol{o_q}$. Conceptually, the attention mechanism gives more weights on certain words in the candidate answer, and the weights are computed according to the question information.

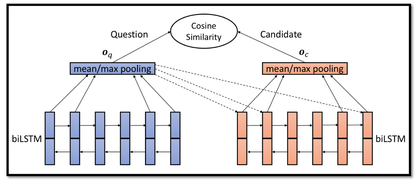

Figure 2: QA-LSTM with attention. The figure was adapted from (Tan et al., 2015).

- **Compare-Aggregate Architecture**. The Compare-Aggregate architectures can capture more interactive features between input sentences than the Siamese architectures and the Attentive architectures, therefore typically have better performance when evaluated on public datasets such as TrecQA (Wang et al., 2007) and WikiQA (Yang et al., 2015). In a Compare-Aggregate architecture, vector representations of small units such as words of the sentences are first compared. After that, these comparison results are aggregated to calculate the final relevance score. Figure 3 shows the architecture of BiMPM, a Compare-Aggregate model proposed in (Wang et al., 2017).

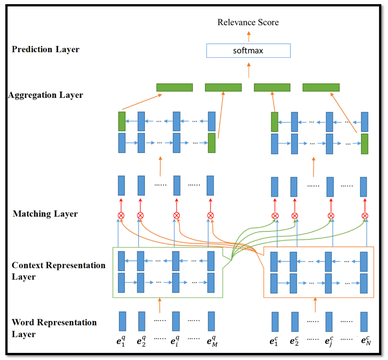

Figure 3: The architecture of the BiMPM model (figure adapted from (Wang et al., 2017))

- **The boundaries between the architecture types are not always crystal clear.**
  - For example, while many Siamese architectures typically capture less interactive features between the input sentences than the Attentive architectures, few recently proposed Siamese architectures have sophisticated comparison layer after the encoding layer (He et al., 2015; Rao et al., 2016). As a result, they even outperform some Attentive architectures.
  - In addition, most of the state-of-the-art compare-aggregate models make use of attention mechanisms.
- Even though the boundaries are not crystal clear, separating the existing different neural architectures into the three categories can provide the big picture more easily.

Datasets and Evaluation Metrics

- TrecQA, WikiQA, InsuranceQA, and SemEval-2016 cQA datasets have been widely used for benchmarking answer selection systems:
  - TrecQA (Wang et al., 2007) was created from the TREC Question Answering tracks. In the literature, we observe two versions of TrecQA: both have the same train set but their dev and test sets differ due to different pre-processing.
  - WikiQA (Yang et al., 2015) is an open-domain question answering dataset that was constructed from real queries of Bing and Wikipedia. All questions with no correct answers are usually removed when the dataset is used to train and evaluate answer selection systems (Bian et al., 2017; Shen et al., 2017).
  - InsuranceQA (Feng et al., 2015) is a domain specific answer selection dataset in which all question and candidate pairs are in the insurance domain.
  - SemEval-2016 cQA (Nakov et al., 2016) was created from the data extracted from the Qatar Living Forums. The dataset was used for the SemEval–2016 Task 3 on Community Question Answering.

Dataset	Example
TrecQA	**Question**: Who established the Nobel prize awards? **Positive Answer**: The Nobel Prize was established in the will of Alfred Nobel, a Swede who invented dynamite and died in 1896. **Negative Answer**: The awards aren't given in specific categories.
WikiQA	**Question**: How many albums has Eminem sold in his career? **Positive Answer**: He has sold more than 100 million records worldwide, including 42 million tracks and 49.1 million albums in the United States. **Negative Answer**: Eminem is one of the best-selling artists in the world and is the best-selling artist of the 2000s.
InsuranceQA	**Question**: Does Medicare cover my spouse? **Positive Answer**: If your spouse has worked and paid Medicare taxes for the entire required 40 quarters, or is eligible for Medicare by virtue of being disabled or some other reason, your spouse can receive his/her own medicare benefits. If your spouse has not met those qualifications, if you have met them, and if your spouse is age 65, he/she can receive Medicare based on your eligibility. **Negative Answer**: If you were married to a Medicare eligible spouse for at least 10 years, you may qualify for Medicare. If you are widowed, and have not remarried, and you were married to your spouse at least 9 months before your spouses death, you may be eligible for Medicare benefits under a spouse provision.
SemEval-2016 cQA	**Question**:Hi;Can any one tell me a place where i can have a good massage drom philipinies????? yesterday i had a massage in Bio-Bil they charged me 300qr for 01 hour bt it is totally waste... pls advice me if theres any philipinos.... **Positive Answer**: Try Magic Touch in Abu Hamour (beside Abu Hamour Petrol Stn)it will just cost you 60QR per hour and I've seen a lot of Qataris as their customers. **Negative Answer**: I dont know the name; you can call them. Do it fast; they have sooooo many reservations ;)

Table 2: Examples from the datasets presented in the paper

- The performance of an answer selection system is typically measured in Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), which are standard metrics in Information Retrieval and Question Answering.
- MRR only examines the rank of the highest-ranked correct candidate. MAP examines the ranks of all the correct candidate answers.

Potential Future Research Directions

- Extending existing answer selection methods to achieve state-of-the-art results on other sentence pair modeling tasks (e.g., paraphrase identification, natural language inference, …), and vice versa.
- Developing novel transfer learning techniques for boosting the performance of answer selection systems (Min et al., 2017).
- Applying answer selection techniques to real-world problems and applications (Lai et al., 2018).