Improved Similar Question Retrieval through LSTM Semantic Embeddings and Siamese Network

Heng Lin
University of Amsterdam
henry.lin@student.uva.nl

Joop Pascha
University of Amsterdam
joop.pascha@student.uva.nl

Luca Simonetto
University of Amsterdam
luca.simonetto@student.uva.nl

Murïel Hol
University of Amsterdam
muriel.hol@student.uva.nl

Abstract

introduction to the problem. contributions in terms of improvements on previous implementation. comparison of results to start-of-the-art.

Keywords— *todo keywords*

1 Introduction

The discipline of Question Answering (QA) is concerned with systems that can automatically answer questions posed by humans in natural language. Today, a large set of questions that regular users have posed on websites such as stackoverflow and yahoo answers have already been answered that could potentially be used to answer similar questions directly. This poses the advantage of avoiding *question-starvation* in which some questions are never answered or take a significant time. However, finding similar questions requires a thorough understanding of the semantic meaning thereof. The *lexicon-syntactic gap* poses a significant problem as questions can be significantly different lexically and syntactically while having the same semantic meaning. Here, an attempt is made to uplift this limitation with the use of Siamese networks in a similar fashion to Das et al. [1] with a change of input representation and network structure.

Das et al. [1] propose the use of Siamese networks to resolve both the problem of finding an appropriate embedding that projects similar questions close to each other in semantic space while at the same time solving the issue of having no question-to-question dataset that is sufficiently large. The latter is important as semantic embeddings generally require an abundance of data. They propose an unsupervised training method in which question-answer pairs (Q,A) are used to train a model of which only the semantic embeddings are used to compare (Q,Q) during testing. The way this is achieved is by using the weight-sharing property of the Siamese network with a contrastive loss function that tries to bring the two inputs (Q,A) closely together in similar space. As a heuristic, negative examples are created by taking answers that do not belong to the question. We believe that an other benefit of this approach compared to using question-to-question pairs only, is that answers tend to be longer and encapsulate more of the semantic meaning within the question. Therefore this approach could indirectly learn the semantic meaning of the questions.

Our first attempt to improve upon the approach of Das et al. [1] is to change the input representation from having a *bag-of-words* assumption to one that is more *context-dependent*. In the original implementation a binary vector is used in which each element denotes whether a trigram is present or not. By using a technique that transforms the question or answers sentence into a representation that takes context into account, a Long Short Term Memmory (LSTM) network, we hypothesize that the semantic embedding thereafter will be greatly improved. Secondly, we hypothesize that their network can be improved upon by replacing the convolutional layer with Relu by either adding multiple non-linear layers or adding more non-linear layers before it. It is expected that the used sparse binary input vector with indications that are uncorrelated (hence *bag-of-words*) are unsuited for convolutional network. Lastly, we give a comparative overview of how different similarity measures within the contractive loss function affect the model performance.

The main contributions of our paper are the followings:

- one
- two
- three

The remainder of this paper is broken up in the following sections. In Related Work (2) the approach of Das et al. [1] is explained in more detail and the relevant topics which we use in our approach such as Siamese networks and LSTM are explained more thoroughly. Thereafter, in Methods (3) our adaptations and enhancements are explained. A comparison of the performance the models are then shown in Results (5). For this we use qualitative methods by using a (Q,Q) test set that is available at todo. An overview of our most important findings and remarks are then discussed in the conclusion (6) after which we give our recommendation for future work in the following section (7).

2 Related Work

The current state of the art in modeling question-answer or question-question pairs appear to go into the direction of deep learning approaches and topic models. For an overview

of the history of question answering we defer to Das et al. [1]. Instead, here we focus first on given a formal definition of the Siamese twin network which is consequently used in Methods 3. Second, an explanation of the use of LSTMs is being given and how this could potentially be used as a feature extractor for the Siamese network. Third, a short explanation is given on the deep learning methods within Natural Language Processing and how this knowledge can be used to enhance the approach of Das et al. [1].

3 Methods

Data

For training the following dataset was used:

- The dataset was obtained from the Yahoo Labs Webscope: *L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part)* [1].

  For testing the following dataset was used:

-

Input Representation

Data Pre-processing

Each question and answer was pre-processed by lowercasing, stemming, stopword and special character removal.

Network Architectures

Loss Functions

Contractive loss function.

Model Training

4 Evaluation

Quantitative Analysis

Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Precision at K (P@K)

Qualitative Analysis

5 Results

6 Conclusion

7 Future Work

References

[1] Arpita Das, Harish Yenala, Manoj Chinnakotla, and Manish Shrivastava. Together we stand: Siamese networks for similar question retrieval. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 378–387, 2016.

[1] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l