Semantic Question Matching

Prakruthi Prabhakar
Computer Science Department
Carnegie Mellon University
prakruthi@cmu.edu

Megha Arora
Computer Science Department
Carnegie Mellon University
marora@cmu.edu

1 Introduction

Semantic Question Matching measures the degree of equivalence in the underlying semantics of paired textual questions. While making such an assessment is trivial for humans, constructing algorithms and computational models that mimic human level performance represents a difficult and deep natural language understanding (NLU) problem.

We formulate the problem as follows:

> Given a pair of questions, determine the semantic similarity between the pair of questions to provide the probability that the questions are duplicate, with 0 indicating that the two questions are completely different and 1 indicating duplicate pairs.

Our work will aim to solve this problem for questions in English using a dataset provided by Quora.

2 Related Work

Semantic Textual Similarity has been an active area of research in the last few years. Initial approaches to the problem involved identifying word level alignment using POS Tags and semantic similarity between words. There have been some attempts to model the problem using random forests with tens of handcrafted features, including cosine similarity of the average of the word2vec embeddings of tokens, the number of common words, the number of common topics labeled on the questions, and the part-of-speech tags of the words. Inspired by recent advancements in deep learning, currently, many neural network architectures and approaches are being tried. Among them, the notable ones include using LSTMs to generate representations for both questions and perform a late fusion using different notions of similarity, and machine translation based approaches using attention to identify token-level alignment between question pairs.

3 Methodology

We plan to use an Long Short Term Memory(LSTM)-based model, because they are better at capturing long-term dependencies. We will generate question embeddings for the two questions, and then feed those question embeddings into a representation layer. To merge the two embeddings, we will try concatenating the two vector representation outputs from the representation layers, or use custom features like distance and angle between the two vectors, and feed that into a dense layer to produce the final classification result.

4 Datasets

We will use Quora question similarity pairs Dataset [5] for training and evaluation. Additionally, we plan to augment this dataset with CoNLL Semantic Textual similarity datasets [6].

References

[1] Juri Ganitkevitch, Benjamin Van Durme, Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In Proceedings of NAACL-HLT.

[2] Lushan Han, Abhay Kashyap, Tim Finin, James May- field, and Jonathan Weese. 2013. UMBC EBIQUITY-CORE: Semantic textual similarity systems. In Proceedings of the Second Joint Conference on Lexical and Computational Semantics.

[3] Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.

[4] Ankur P Parikh, Oscar Täckström, Dipanjan Das, Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference, arXiv 2016

[5] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

[6] http://ixa2.si.ehu.es/stswiki/index.php/Main_Page