Arcada Working Papers 3/2018
ISSN 2342-3064
ISBN 978-952-5260-92-2

ARCADA

Deep Learning Models for Cross-Lingual Short-Text Matching

Patrick Jansson, Shuhua Liu

www.arcada.fi

Arcada Working Papers 3/2018
ISSN 2342-3064
ISBN 978-952-5260-92-2

Deep Learning Models for Cross-Lingual Short-Text Matching*

Patrick Jansson[i] and Shuhua Liu[ii]

Abstract

Short-text matching is a fundamental task in many important NLP applications such as question answering, machine translation and conversational assistants. The CIKM2018 AnalytiCup takes on the challenge of language adaptation issue in short-text matching and aims to promote development of advanced models for cross-lingual matching of question pairs. This article reports our participation in the challenge. We explore a number of machine learning models for question pair matching and compare their performance. Our best performing model achieved a log loss of 0.39294. The AnalytiCup winner model achieved a log loss of 0.31731.

Keywords: Short text matching; question pairs; cross-lingual; neural embeddings; deep learning models

1 INTRODUCTION

The CIKM2018 AnalytiCup takes on the challenge of cross-lingual short-text matching to advance the development of chatbots or online service assistants. Short-text matching is a fundamental task in many important NLP applications such as paraphrase identification, question answering and machine translation systems. While the focus of the AnalytiCup is on the automatic matching of question pairs in different languages, the problem is defined as to determine whether two questions in the same or different languages have similar meaning or express the same semantics. Participants of the challenge are provided with customer service datasets, containing ecommerce user questions in English and Spanish (https://tianchi.aliyun.com/markets/tianchi/CIKM2018).

* Funding from Helsinki Region Urban Research Program (http://www.helsinki.fi/kaupunkitutkimus/) and Arcada TUF Foundation (http://tuf.arcada.fi) are gratefully acknowledged.
[i] Arcada University of Applied Sciences, Dept. of Business Management and Analytics, [patrick.jansson@arcada.fi]
[ii] Arcada University of Applied Sciences, Dept. of Business Management and Analytics, [shuhua.liu@arcada.fi]

The training datasets include (1) English training data, which contains a pair of questions in English and their respective translations to Spanish, and label of whether the pair of questions are a match; (2) Spanish training data, which contains a pair of Spanish questions and their respective translations to English, and label of whether the pair are a match; (3) An unlabeled dataset that has only Spanish question and its translation in English; and (4) Test data contains only Spanish question pairs, as models' performance are evaluated on the target language, i.e. Spanish language. In addition to the data, word vectors are provided, and participants are given some constraints in developing the solutions[1].

Question pair matching could mean question type matching, entity attribute matching, term matching and term sequence matching, which sometimes matters, sometimes not. The challenge lies in that many semantically similar questions may have relatively little word overlap, with less context to help the assessment, or they are in different languages and the automatic translation may bring errors and more complexity. In this study, we approached the problem with combined use of supervised and semi-supervised methods for building deep neural network models. We developed two types of deep learning models: LSTM models with attention and the Transformer models. In addition we tested an embedding based baseline model. Our best result achieved a log loss 0.39294 while the winner's best result achieved a log loss 0.31731. In the following sections we describe the data, methods, models and our analysis.

2 DATA AND PREPROCESSING

2.1 Datasets overview

An overview of the datasets is shown in Table 1.

Table 1: Training, test and unlabeled datasets

Datasets	Columns	Size
English training	en1, es1, en2, es2, label	20000
Spanish training	es1, en1, es2, en2, label	1400
Test A	es1, es2	5000
Test B	es1, es2	10000
Unlabeled	es, en (direct translation)	55669

[1] Participants can only use the data provided by the organizer, including the labeled data, unlabeled data, translations, word vectors. Only fastText pretrained word vectors are allowed to use. No other data or pretrained models are allowed. Training a translation model is not recommended.

The samples of training sets are in both languages. The test sets are Spanish only. We considered different extractions and combinations of the datasets to make most out of the labelled and unlabelled data for building the models. To "create" more data, we give label 1 to every translation of the original questions. In other cases the questions get their original label. The datasets are utilized in a number of different ways when developing the different models, which we will describe in detail in section 3.

2.2 Preprocessing

Preprocessing include POS tagging and dependency tree parsing using spacy (https://spacy.io/usage/linguistic-features). As the provided FastText vectors are not the binary version so we can't use them to get OOV (out of vocabulary) embeddings. We tried to use a simple method to get around the problem. The idea is to find the most similarly spelled words for each word which doesnt have a fasttext vector, and using their average as the vector for the OOV word. For an OOV word, find the 3 most similar words within the FastText vocabulary (language specific) in terms of their characters using pythons difflib; then get the embedding for each of the 3 words and use their average as the embedding for the OOV word. This helps to certain extent but can't get embeddings for all OOV words.

3 DEEP LEARNING FOR QUESTION PAIR MATCHING

In the world of deep learning for natural language processing, recurrent neural networks (RNN) in general, and long short-term memory (LSTM) and gated recurrent neural networks (GRU) in particular[2], have been established as state of the art approaches in sequence modeling problems such as language modeling and machine translation (Vaswani et al, 2017; Conneau et al, 2018). Attention mechanisms have also become an integral part of sequence models in various tasks, allowing modeling of dependencies without paying attention to their distance in the input or output sequences (Vaswani et al, 2017).

There has also been interesting development in machine learning approaches to sequence modelling tasks. While supervised learning has been the dominant approach for long, semi-supervised learning is gaining more and more attention, promoting better use of unlabeled data in model development (Dai and Le, 2015).

Inspired by the above works and developments in this area (Vaswani et al, 2017; Conneau et al, 2018; Dai and Le, 2015), we explore the potential of building effective question pair matching models using LSTM model with attention, and transformer model with semi-supervised learning.

[2] There is no fundamental different between LSTM and GRU models, the choice is more of technical nature.

3.1 LSTM model with Attention

LSTM networks as a special kind of RNN are capable of learning long-term dependencies in text (Hochreiter and Schmidhuber, 1997). They work well on a large variety of problems. In sequence modeling the LSTM models or CNN models often include an encoder and a decoder, and the best performing models also connect the encoder and decoder through an attention mechanism (Vaswani et al, 2017). This is the first approach we applied for building our question pair matching models.

In this set of experiments, question pairs in the same languages from both english and spanish training sets are used for training. For spanish - english question pairs from unlabeled dataset, label is set as 1 due to it being a translation (making the assumption that the translation is good). For spanish - english pairs from unlabeled data where columns are randomly shuffeled, label is set to 0 (making the assumption that the shuffle didnt randomly match any two texts).

The architectures of our models using the LSTM network and the Transformer network are shown in Figure 1. They share a very similar four components structure: the embedding representation, the sentence encoder, the Max pooling component and the classification component. Their main difference is the use of LSTM-encoder vs Transformer-encoder.

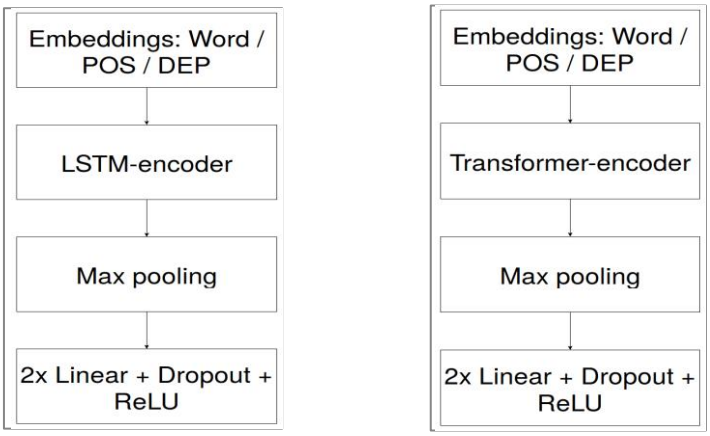

Figure 1. Model Architectures for both LSTM with Attention and Semi-supervised Transformer Models

The LSTM with attentions models take a similar form as in the work of Conneau et al (2018) on supervised learning of universal sentence representations from natural language inference data[3]. The Embeddings component has three layers, concatenating three

[3] Their work showed that universal sentence representations can be trained with supervised learning from the Stanford Natural Language Inference datasets (SNLI, Bowman et al, 2015). The evaluation indicated that the learned representations can consistently outperform unsupervised methods like SkipThought vectors (Kiros et al., 2015) on a wide range of transfer tasks. Their test with using the Multi Genre NLI (MultiNLI, Williams et al., 2018), which contains ten distinct genres of written and spoken English, covering most of the complexity of the language, observed a significant boost in performance overall compared to the model trained only on SLNI, suggesting that having a larger coverage for training helps learn even better general representations.

types of word embeddings to into one embedding to represent each word: (1) pretrained fastText word embedding, they are not retrained during training; (2) POS embedding (Part of Speech embeddings), 20-dimensional embedding, language specific; (3) DEP embedding (Dependency embeddings), 30-dimensional embedding, language specific. The POS and DEP embeddings are updated during model training process[4]. After the embeddings component come the sentence encoder and Max pooling components. The Sentence Encoder is a single layer Bidirectional LSTM with 512 units. A single vector representation for the text sequence (i.e. the question sentences) is then obtained by max pooling similar as in the InferSent (Conneau et al 2017). Finally, at the Classification step, the vector representations of the question pairs, their absolute difference and their hadamard product are concatenated as input to a neural network with 2 linear layers followed by dropout and ReLU as activation function. The two different languages each have their own embedding representations and LSTM encoders.

The idea and form of attention mechanism between the two different questions comes from the work of Vaswani et al (2017, Attention Is All You Need) on transformer model, which relyies entirely on self-attention to compute representations of its input and output without using RNNs or convolution. Self-attention (also called intra-attention) is an attention mechanism ”relating different positions of a single sequence in order to compute a representation of the sequence”. In a self-attention network, each token is connected to any other token in the same sentence directly via self-attention. Self-attention has been used in a variety of tasks including reading comprehension, abstractive summarization, textual entailment, etc (Vaswani et al, 2017).

We trained LSTM based question pair matching models both with and without self-attention. Our results do not support the effectiveness of self-attention mechanism in short text matching problem. Our final/best performing LSTM model is a regular LSTM with max-pooling, no attention, achieved a log loss of 0.39294.

3.2 Transformer based question pair matching model

Comparing to LSTM models, the transformer model proposed by Vaswani et al (2017) has a simpler neural network model architecture based solely on attention mechanisms to draw global dependencies between input and output, without the use of recurrent or convolutional structure at all. The attention networks of the transfomer can have multiple attention heads. Their experiments showed such models to be superior in quality on two machine translation tasks (English-to-German, English-to-French) while being more parallelizable and requiring significantly less time to train (Vaswani et al, 2017).

Similar to the LSTM based question pair matching models, our transformer based models also have four components. The embeddings represnetation, max pooling and classification components are basically the same as the LSTM models. The only difference is with the sentence encoder, which now consist of 4 hidden encoder layers with a size of 512; 4 attention heads; **key and value depth** as well as filter size 256. The hidden state of the encoder is passed onto the classifier which determines if two sentences are a match or not.

[4] POS and DEP parsers are from spaCy (https://spacy.io).

For this set of experiments, the traning set contains pairs in the same language from english and spanish training sets. In addition, all translations of original questions are used with a label 1; en_1 - es_2 and es_2 - en_1 pairs take the true labels. For spanish - english question pairs from unlabeled data, labels are set as 1 due to it being a translation; for spanish - english pairs from unlabeled data where columns are randomly shuffeled, labels are set to 0. We found that this technique didn't work for transformer models, thought it worked better for the LSTM based model.

The best performance of the transformer based models stands at a log loss of 0.4628.

3.3 Baseline model

In addition to the above two models, we also tested a much simpler baseline model, in which the questions are simply models as bags of word embeddings, their representation as the average embedding of all the words in each question. Their similarity is measured by calculating the cosine similarity between the two embeddings.

Fasttext word embeddings are used, but not sentence embeddings. The word embeddings in English and Spanish wasn't aligned as no training is required. The similarity was just measures on the Spanish-Spanish test set with the provided Spanish embeddings.

The Test A log loss was very high at 1.28 (the organizer reported 1.27718 using this method). This primitive test indicated that simple model of short text as bag of embeddings is not enough. The two deep learning models delivered much better results.

4 SUMMARY AND CONCLUSION

In this paper, we reported our participation in the CIKM 2018 AnalytiCup challenge on cross-lingual question pair matching. We described our models that leverages the power of LSTM models, attention mechanisms, transformer models and unlabeled data. The performance of the different types of models are summarized in Table 2. Our best results achieved a log loss of 0.39294, while the winning team achieved a log loss 0.31731.

Table 2: Comparing cross-lingual questions pair matching modeling

Models	Baseline	LSTM with attention	Transformer model
Performance/Logloss	1.28	0.39294	0.46283

It should be noted that we did not try many other ways of incorporating more features in the models or using model ensembles to get extra performance gains. Some recent experiences from other related tasks seem to indicate that the best results are obtained by the combination of the word embedding based method and a unigram language model or by an ensemble of deep and shallow learning. This will be a very interesting next step work for us.

We did try to incorporate the Word Mover's Distance (WDM)[5] between texts as additional feature, but it didn't show any impact so it is not included in our best models.

We also experimented with regular GRU Rnn instead of the transformer. A very large single layer GRU seems to work a bit better even though it overfits. The transformer type models initially gave better results with less overfitting but we were able to train LSTM models with much better performance. Overall, our results do not support the effectiveness of self-attention mechanism in short text matching problem.

We also compared the effect of having a shared model for both languages or having separate encoders for each language. Our experiments indicates that it does seem like having a shared model for English and Spanish works a bit better than having two separate encoders.

REFERENCES

Conneau Alexis, Douwe Kiela, Holger Schwenk, Loic Barrault and Antoine Bordes, Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, EMNLP 2017, September 9-11, 2017, Copenhagen, Denmark

Conneau Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer and Herve Jegou, Word Translation Without Parallel Data, ICLR 2018, April 30 - May 3, 2018, Vancouver Convention Center, Vancouver, Canada

Dai Andrew M. and Quoc V. Le, Semi-supervised Sequence Learning, the 29th Conference on Neural Information Processing Systems (NIPS), December 7-12, 2015, Montreal, Canada.

Kusner Matt J., Yu Sun, Nicholas I. Kolkin and Kilian Q. Weinberger, "From Word Embeddings To Document Distances", Proceedings of the 32nd International Conference on Machine Learning, 2015, Lille, France.

[5] Kusner et al, From Word Embeddings To Document Distances. WMD tried to measure the dissimilarity between text documents. It formulates the distance as a minimum cost that one document (based on word embedding representations) need to take to exactly match the other, which is a standard linear transportation problem. They showed its effectiveness in document classification.

Bowman Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning, A large annotated corpus for learning natural language inference. In Proceedings of EMNLP 2015, September 17-21, 2015, Lisbon, Portugal.

Kiros Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, Sanja Fidler, Skip-Thought Vectors, the 29th Conference on Neural Information Processing Systems (NIPS 2015), December 7-12, 2015, Montreal, Canada.

Vaswani Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin, Attention Is All You Need, the 31st Conference on Neural Information Processing Systems (NIPS 2017), December 4-9, 2017, Long Beach, CA, USA.

Williams Adina, Nikita Nangia, and Samuel R Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in Proceedings of NAACL-HLT 2018, June 1-6, 2018, New Orleans, Louisiana, Association for Computational Linguistics.