Detecting Duplicate Questions

Alfonce Nzioka
Department of Computer Science
Stanford University
alfonce@stanford.edu

Abstract

We present a solution to the problem of identifying duplicate questions. We focus on a recent dataset of question pairs annotated with binary paraphrase labels and show that the decomposable attention model (1) outperforms most other more complicated neural architectures. Furthermore, when the model is pretrained on a noisy dataset of automatically collected question paraphrases, it results in a higher performance on the task.

1 Introduction

The task of identifying duplicated questions can be viewed as an instance of the paraphrase identification problem, which is a well-studied NLP task that uses natural language sentence matching (NLSM) to determine whether two sentences are paraphrase or not (2). This task has wide array of useful NLP application. For example, in question-and-answer (QA) forums, there are vast numbers of duplicate questions. Identifying these duplicates and consolidating their answers increases the efficiency of such QA forums. Moreover, identifying questions with the same semantic content could help web-scale question answering systems that are increasingly concentrating on retrieving focused answers to users' queries.

In this project, we focus on a dataset published by Quora.com containing over 400K annotated question pairs containing binary paraphrase labels.[1]. We believe that this dataset presents a great opportunity for the NLP practitioners tue to its scale and quality; it can result in systems that accurately identify duplicate questions, thus increasing the quality of many QA forums. We examine a simple model family, the *decomposable attention model* (1) that has shown promise in modeling natural language inference and has inspired recent work in similar tasks (Chen et al., 2016; Kim et al., 2017).

To significantly improve our model performance, we pretrain all our model parameters on the noisy, automatically collected question-paraphrase corpus Paralex (Fader et al., 2013) followed by fine-tuning the parameters on the Quora dataset. This two stage training procedure achieves comparable or better results with respect to several complex neural architectures, all using pretraining word embeddings.

2 Related Work

Paraphrase identification is a well-studied task in NLP (Das and Smith, 2009; Chang et al., 2010). Here, we focus on an instance, that of finding questions with identical meaning. With the renaissance of neural networks, several neural-based frameworks have been proposed for the task of paraphrase identification. The first framework is based on a siamese neural network consisting of two sub-networks joined at their outputs, where the sub-networks share the same weights at all levels and are responsible for extracting features from the input, and the output level computes the distance between the two feature vectors generated by the sub-networks (3). The shortcoming of this approach

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

is that there is no interaction between two sentences during the training process, which might cause information loss. The "compare-aggregate" approach is proposed (4), which captures the interation between two sentences by performing a word-level matching and aggregating the results into a vector the final classification. However, this approach fails to account for other types of matchings such as phrase-by-sentence and only performs matching in a single direction, thus neglecting information in the sentence pairs.

Wang et al. (2017) present the *bilateral multi-perspective matching* model (BiMPM) to tackle the limitations of neural-based frameworks. This approach uses a character-based LSTM at its input representation layer, a layer of bi-LSTMs for computing context information, four different types of multi-perspective matching layers, an additional bi-LSTM aggregation layer, followed by two-layer feedforward network for prediction. In contrast, the decomposable attention model uses four simple feedforward networks to attend, compare and predict, leading to a more efficient architecture.

3 Approach

3.1 Baseline

We implemented three baseline models all of which use 300-dimensional Glove embeddings and are based on the Siamese network framework. In the first model, each question is passed through two LSTM layers. The outputs from the final layer are concatenated and fed into a dense layer to produce the final classification result. The second approach is similar except that it uses Bi-LSTMs to encode each question. The third approach is based on the MaLSTM model (5). Two embedded matrices representing two questions are fed into a 50-dimensional LSTM. A similarity function is applied to the Manhattan distance between the final states of each LSTM to compute the relatedness label.

3.2 Decomposable Attention Model

First we present the problem formulation. Let DA model divides the prediction into three steps: Attend, Compare and Aggregate. Due to space limitations, we only provide a brief overview and refer to Parikh et al. (2016) for further details on each of these steps.

Attend Let $a = (a_1, ..., a_{l_a})$ and $b = (b_1, ..., b_{l_b})$ be two input texts containing l_a and l_b tokens. Let $\overline{a}$ and $\overline{b}$ be input vectors representing two input texts. The elements of $\overline{a}$ and $\overline{b}$ are aligned using a variant of neural attention to decompose the problem into the comparison of aligned phrases.

$$e_{ij} := F(\overline{a}_i)^T F(\overline{b}_i)$$

The function F is a feedforward network. The aligned phrases are computed as follows:

$$\beta_i = \sum_{j=1}^{l_b} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_a} \exp(e_{ik})} \overline{b}_j$$

$$\alpha_j = \sum_{i=1}^{l_a} \frac{\exp(e_{ij})}{\sum_{k=1}^{l_a} \exp(e_{kj})} \overline{a}_i$$

β_i is the subphrase in $\overline{b}$ that is aligned to $\overline{a}_i$ while α_j is aligned to $\overline{b}_j$.

Compare Second, we separately compare the aligned phrases $\{(\overline{a}_i, \beta_i)\}$ and $\{(\overline{b}_j, \alpha_j)\}$ using a feedforward network G

$$v_{1,i} = G([\overline{a}_i, \beta_i]) \forall i \in (1, ..., l_a) v_{2,i} = G([\overline{b}_j, \alpha_j]) \forall j \in (1, ..., l_b) \quad (1)$$

where the brackets [] denote concatenation.

Aggregate The sets $\{v_1, i\}_{i=1}^{l_a}$ and $\{v_2, j\}_{j=1}^{l_b}$ are aggregated by summation. The sum of the two sets is concatenated and passed through another feedforward network followed by a linear layer to predict the label y.

4 Experiments

4.1 Implementation Details

Datasets We evaluate our models on the Quora question paraphrase dataset which contains over 400K question pairs with binary labels. We split the data into 10K pairs each for development and test, and the rest for training. We duplicated the training set, which has approximately 36 % positive and 64 % negative pairs, by adding question pairs in reverse order. In pretraining the parameters for the decomposable attention model, we use the Paralex corpus (6) which consists of 36 million noisy paraphrase pairs including duplicate reversed paraphrases. We created 64 million artificial negative paraphrase pairs by combining the following three types of negatives in equal proportions: (1) random unrelated questions, (2) random questions that share a single word, and (3) random questions that share all but one word.

Hyperparameters We tuned the following hyperparameters by grid search on the development set. The leading settings for the decomposable attention model are presented in Table 1.

Table 1: Hyperparameters

Embedding dimension	300
Pretraining batch size	256
Tuning batch size	64
Learning rate	0.1
Dropout ratio	0.1

4.2 Results

Other than our baselines, we compare with h Wang et al. (2017) in Table **??**

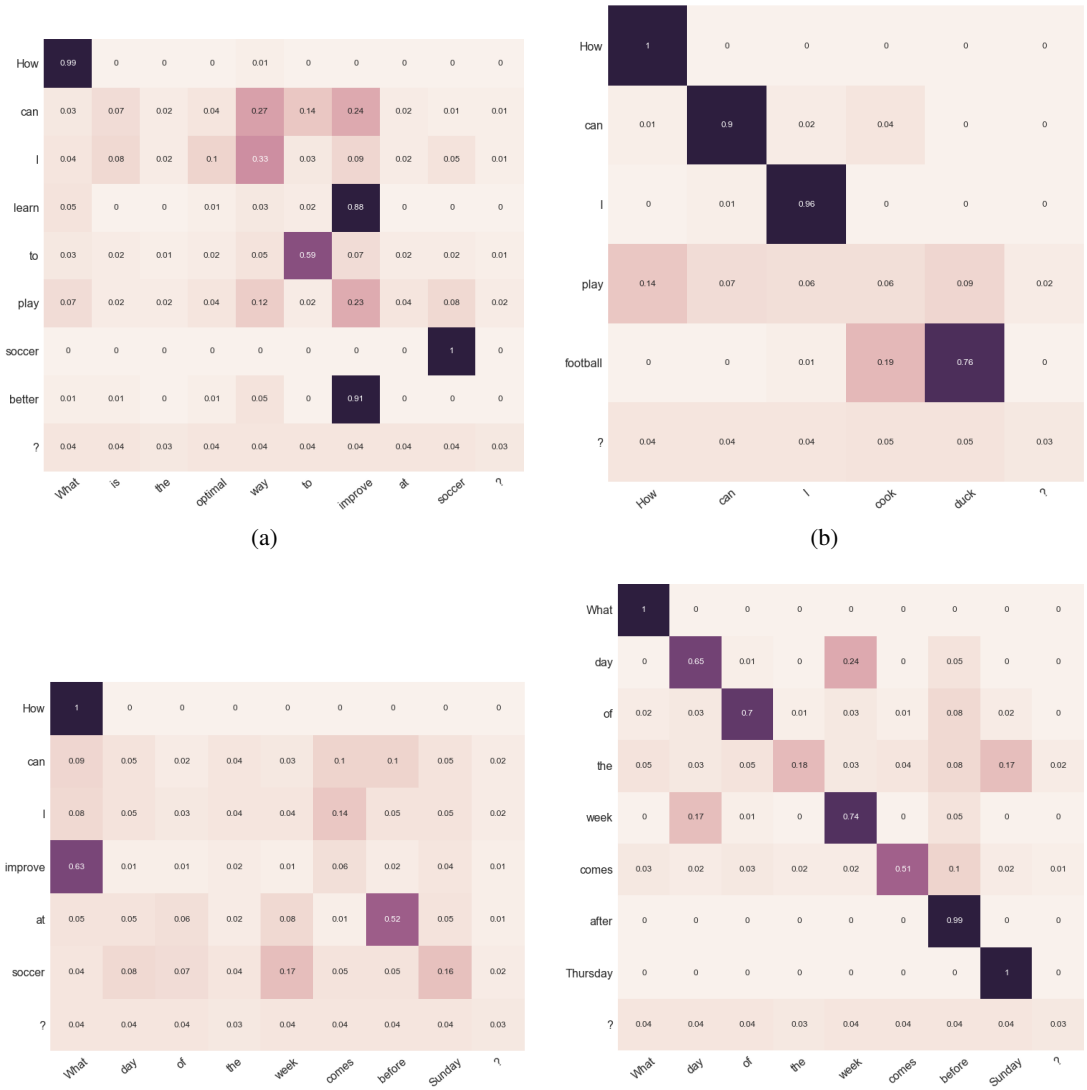
Method	Dev Acc	Test Acc
Siamese-CNN	-	300
Multi-Perspective CNN	-	81.38
Siamese-LSTM	-	82.58
Multi-Perspective-LSTM	-	83.21
L.D.C	-	85.55
BIMPM	88.69	88.17
LSTM concat	81.45	80.31
BiLSTM concat	84.15	83.19
MaLSTM	81.57	79.65
Attention	86.80	85.91
Attention-Pretrained	**87.45**	**86.82**

Table 2: Results on the Quora dev and test sets in terms of accuracy. The first six rows are taken from (Wang et al., 2017)

Our vanilla attention model (without pretraining) outperforms most of the other models. The final row shows our best performing model which leverages the full power of pretraining the model on Paralex. However it still falls short of the BiMPM model.

5 Analysis

The attention mechanisms for sample queries are presented in the visualizations below.



(a) (b)

(c) (d)

Figure 1: Visualization mechanisms

Figure 1(a) shows visualization for the question pair: “How can I learn to play soccer better?” and “What is the optimal way to improve at soccer?”. Our attention model correctly labels the pair as duplicate. In this example the network focuses on “learn”, “play”, “soccer”, “better”, “optimal”, “way”, and “improve”.

Figure 1(b) shows visualization for the pair: "How can I cook duck?" and "How can I play football?". In this case “How”, “can”, and “I” match up very well. However, "football" matches up with "duck". The network therefore correctly predicts that these questions are not duplicates.

Figure 1(c) shows attention mechanism for the pair: "How can I improve at soccer?" and "What day of the week comes before Sunday?". The visualization shows that these two questions have very little overlap, thus not duplicate.

Figure 1(d) is an example where the model fails. In this case the question pair "What day of the week comes before Sunday?" and "What day of the week comes after Thursday?" are marked as duplicate. "Thursday" and "Sunday" are very related by the attention mechanism. This example shows how the model lacks knowledge of the outside world: the relation of days of the week.

6 Conclusion and future work

We presented a contribution on question paraphrase identification on the recently published Quora corpus. We showed that pretraining the model on automatically labelled noisy but task specific-data results in better accuracy on this task.

References

[1] Ankur Parikh, Oscar Tackstr om, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of EMNLP*.

[2] Wang, Zhiguo, Wael Hamza, and Radu Florian. "Bilateral Multi-Perspective Matching for Natural Language Sentences."*arXiv preprint arXiv:1702.03814* (2017).

[3] Bromley, Jane, et al. "Signature Verification Using A "Siamese" Time Delay Neural Network." IJPRAI 7.4 (1993): 669-688.

[4] Wang, Shuohang, and Jing Jiang. "A Compare-Aggregate Model for Matching Text Sequences." *arXiv preprint arXiv:1611.01747* (2016).

[5] Mueller, Jonas, and Aditya Thyagarajan. "Siamese Recurrent Architectures for Learning Sentence Similarity." AAAI. 2016.

[6] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of ACL.*

[7] Ming-Wei Chang, Dan Goldwasser, Dan Roth, and Vivek Srikumar. 2010. Discriminative learning over constrained latent representations. In *Proceedings of HLT-NAACL.*

[8] Dipanjan Das and Noah A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of ACL-IJCNLP.*