NLP project1 report

**Natural Language Inference for Fake News Detection**

SPRING 2019 資工四 B04902051 林承豫

Project 1 資工四 B04902099 黃嵩仁

Due Date: 2019.5.10 網媒博二 D06944009 蘇弘庭

Division of Work

資工四 B04902051 林承豫

SVM Model training and testing

Preprocessing

資工四 B04902099 黃嵩仁

SVM Model training and testing

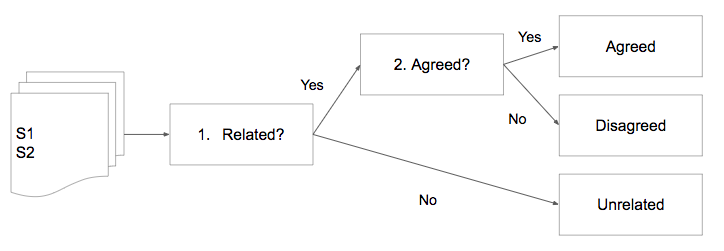
Preprocessing

網媒博二 D06944009 蘇弘庭

NN Model training and testing

**Methods**

Generally, we formulate the problem as a *hierarchical binary classification* as shown below for the following reasons. (1) Intuitively, both of agreed and disagreed indicate two sentences are related. (2) We argue that the model relies on different parts of the sentences to recognize the agreed/disagreed and related/unrelated . (3) Both of the classification tasks have similar tasks with other large-scale datasets (e.g. Quora Question Pairs for related/unrelated), we can benefit from pretraining the model with corresponding dataset.



1.naive based similarity

(1)Jaccard similarity: calculated based on the intersection part of two sentences.

(2)Cosine similarity: calculated based on the frequency of words appeared in two sentences.

(3)Human knowledge: We find out that most sentences labeled disagreed had a common word “rumors” in them. So we add a rule to classify disagreed sentence pairs by this findings. And the performance has improve above 10 percent.

2. Support vector machine

1. Use the jaccard and cosine similarity as two input for support vector machine. Construct two vector machine, one for classify related /unrelated, another one for agreed/disagreed. The performance is almost the same as similarity method with human knowledge.

3. Information Retrieval method

Step 1：calculate tf-idf value of each title (We convert all characters to lowercase before tokenizing, then we choose top 5000 words ordered by term frequency)

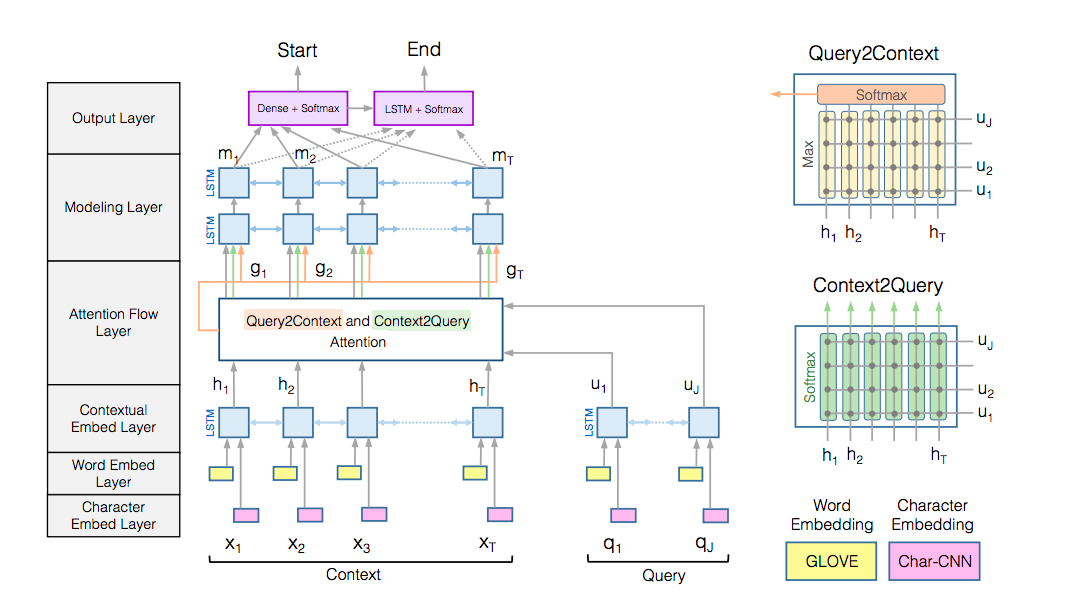
Step 2：Dimension Reduction (LSA：We use svd toolkit to factorized the matrix and keep first 256 dimension)

Step 3-1：Cosine similarity (We calculate the cosine similarity of title\_1 and title\_2, respectively. We get the Q1, Q2, Q3 of agreed/disagreed/unrelated set. Thus, we get the threshold to classify test data.)

Step 3-2：SVM model (We train two svm model, one for related/unrelated data, and one for agreed/disagreed data.)

4. Neural Networks: BiDAFake

Intuitively, to determine whether a sentence S2 agrees another sentence S1 or not, we pay attention to some specific words according to each other. More precisely, instead of simply represent a sentence by widely used bag of words or recurrent neural networks, we apply the attention mechanism to each ot the sentence based on another one (i.e. attend S1 with S2, and attend S2 with S1).



We utilize a classic model, BiDirectional Attention Flow (BiDAF, Seo et’al, ICLR 2017), which is originally used for reading comprehension task and widely adapted to reasoning-required tasks, as our backbone model, we modify the BiDAF reading comprehension model from RCZoo package. The original model takes a context C and a query Q and fuse the information of them by C2Q and Q2C attention, we replace the output layer with a classifier which inputs mean of modeling layer output. We respectively train the related and agreed model. Observing the imbalance of the data, we set the weight of positive/negative, 1:9 for the agreed model and 2:1 for the related model.

To satisfy our curiosity of the difference between the related and agreed on the point of view of the model, we train the model in three ways. (1) Jointly train two binary classification model with two separate binary classifier. (2) Train two models. (3) Train two models with layers shared, in this setting, we separately train the attention flow layer and binary classifier and share the remaining layers.

**Experiments** (ex. acuracy on validation set)

We evaluate our proposed method by (1) Weighted binary classification (WBC)accuracy. As we cannot access the origin testing set, we split 10% of training set for testing and train with the rest of dataset and set the weights of related to 2:1 and agreed to 1:9. (2) News Classification Weighted Accuracy (NCWA) with testing set on kaggle.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | WBC-related | WBC-agreed | NCWA |
| Similarity | Jaccard | 0.7164 |  | 0.58837 |
| Cosine | 0.6449 |  | 0.57204 |
| Similarity with human knowledge | Jaccard | 0.7281 |  | 0.72526 |
| Cosine | 0.6316 |  | 0.72365 |
| SVM | with naive similarity |  |  | 0.70253 |
| tf-idf | Cosine | 0.5121 | 0.5494 | 0.63804 |
| SVM | 0.5140 | 0.5643 | 0.60074 |
| BiDAFake | Single | 0.7425 | 0.1000 |  |
| Double | 0.8109 | 0.8004 |  |
| Shared | 0.8204 | 0.8084 |  |

**Discussion**

1. Does the human knowledge help?

The second block demonstrates the improvement in large margin by applying prior knowledge. It demonstrates the two sides of the coin: Utilizing the handcraft feature by carefully dig into the data enhances the feature and boosts the classification results. However, on the other hand, it indicates that the training data strictly follows certain patterns and consequently is very sensitive to adversarial attacks.

1. Does IR principles help?

The fourth block shows that applying TF-IDF weights improves 6% of NCWA score without requiring the human knowledge. Practically, the human prior knowledge requires human efforts which are usually expensive. Additionally, human knowledge is data dependent while TF-IDF heuristics are more general and can be automatically calculated.

1. Does the NN based model agree that, agreed and related are related?

The last block demonstrate the WBC results for three varieties of BiDAFake model. Single: Training two binary classification model jointly with two classifiers. Double: Training each of the classifier with two different networks. Nattn Shared: Training two model with separate attention flow layer and classifiers.

The results shows that the model is dominated by the related scores when training with a single model regardless the weights of loss are set according to the label distribution and results in extremely low agreed scores. Training with two models solves this problem and achieves much better results on both tasks. Sharing the layers further improves both scores by about 1% by observing more data and generalize the layers to solve both tasks, while the attention flow layers focus on different parts corresponding to the task.

**Conclusion**

We formulate the problem as hierarchical classification by observing the nature of the task and the possibility of transferring from similar tasks. We implement a similarity baseline and attempt three ways. (1) Leveraging human knowledge by digging into the data. (2) Applying the TF-IDF weight from information retrieval principle. (3) Utilizing the neural network based reading comprehension model. In our experiments, injecting the observed prior knowledge significantly boosts the news classification accuracy. Also, TF-IDF weighted similarity improves the baseline without requiring human efforts. Finally, the neural network based model is able to learn to classify whether two sentences are related or agreed without additional handcraft features or heuristics and achieves better results on both of binary classification tasks comparing to two baselines, and further improved by sharing the layers and separately learning to attend.

References:

Seo et’al, BI-DIRECTIONAL ATTENTION FLOW FOR MACHINE COMPREHENSION, ICLR 2017, <https://arxiv.org/pdf/1611.01603.pdf>

<https://github.com/lixinsu/RCZoo>