IIIT-H @ CLScisumm-18

Kritika Agrawal and Aakash Mittal

International Institute of Information Technology, Hyderabad
kritika.agrawal@research.iiit.ac.in
International Institute of Information Technology, Hyderabad
aakash.mittal@students.iiit.ac.in

Abstract. In this report, we present our system for the 4th Computational Linguistics Scientific Document Summarization Shared Task (CL-SciSumm 2018), which poses the challenge of identifying the spans of text in a reference paper (RP) that most accurately reflect a citation (i.e. citance) from another paper to the RP and identifying the discourse facet of the corresponding text. We modeled Task1A as Classification Prob-lem which classifies a pair of sentences into classes paraphrase and non-paraphrase. We used Convolutional Neural Network along-with transfer learning for this. Task1B is solved as multi label classification problem.

Keywords: Paraphrase Detection · Classification · Transfer Learning

1 Methodology

1.1 Task Description

CL-SciSumm explores summarization of scientific research, for the computational linguistics research domain. An ideal summary of computational linguistics research papers would be able to summarize previous research by drawing comparisons and contrasts between their goals, methods and results, as well as distil the overall trends in the state of the art and their place in the larger academic discourse. There are two tasks in CL-SciSumm 2018. The training dataset contains 40 topics of documents. A topic is consisted of a Reference Paper (RP) and Citing Papers (CPs) that all contain citations to the RP. In each CP, the text spans (citances) have been identified that pertain to a particular citation to the RP. In Task 1A, for each citance, we need to identify the spans of text (cited text spans) in the RP that most accurately reflect the citance. In Task 1B, for each cited text span, we need to identify what facet of the paper it belongs to, from five predefined facets, which are Aim, Method, Results, Implication and Hypothesis. In Task 2, we need to generate a structured summary of the RP from the cited text spans of the RP.

1.2 Task1A

In this task, for each citance, we need to identify the spans of text (cited text spans) in the RP that most accurately reflect the citance. We modeled this task as

classification problem where we are classifying pair of sentences into Paraphrase and non-paraphrase. We used the method described in "Text Matching as Image Recognition" paper of using CNN, ideally used for images, for text matching problem. Firstly, a matching matrix whose entries represent the similarities between words is constructed and viewed as an image. Then a convolutional neural network is utilized to capture rich matching patterns in a layer-by-layer way. By resembling the compositional hierarchies of patterns in image recognition, the model can successfully identify salient signals such as n-gram and n-term matchings.
We train the MatchPyramid model introduced in "" on Microsoft Research Paraphrase Detection Dataset. We then re-train the model using transfer learning on the given Training Dataset.

1.3 Task1B

In this task, for each cited text span, we need to identify what facet of the paper it belongs to. There are 5 predefined facets - Aim, Hypothesis, Method, Result, Implication. We have modeled this problem as multi label classification problem as for many sentences in annotation, a text belonged to more than one facet. We are using term frequency-inverse document frequency(tf-idf) features and trained NB-SVM on it. NBSVM is an SVM model trained on naive bayes features.

2 Experiments and Results

We created 3 datasets from the given Training Set for our approach:
a. N4 dataset - For each citance sentence we create a positive sample as the pair of the citance sentence and the corresponding reference sentence. We also create 4 negative samples as pair of citance sentence and 2 sentence before and 2 sentences after the reference sentence.
b. N2 dataset - For each citance sentence we create a positive sample as the pair of the citance sentence and the corresponding reference sentence. We also create 2 negative samples as pair of citance sentence and 1 sentence before and 1 sentences after the reference sentence.
c. N1 dataset - For each citance sentence we create a positive sample as the pair of the citance sentence and the corresponding reference sentence. We also create 1 negative sample as pair of citance sentence and randomly selecting a sentence from the first 2 sections of the paper which is not a reference sentence.

1. Preprocessing
All the words are converted to lower case and stemmed so as to decrease the vocabulary size.
2. Model
Our model consist of 2 embedding layers, one for the citance sentence and one

for the reference sentence. A cross is taken between the 2 layers to generate a matrix (the matrix will be given as an image to convolution network). A convolutional layer is applied on the matrix followed by an dynamic pooling layer. The activation function used is Relu. The output of the dynamic layer is flattened and then fed to a fully connected layer. The activation function used is softmax.

We first trained the model on Microsoft Research Paraphrase detection dataset. We followed the fine parameter tuning approach for transfer learning. We retrain all the layers in the model with weights initialized with those learnt on Microsoft paraphrase dataset.

We performed experiments on two models:
1. TL: Model initially trained on MSR Paraphrase Detection dataset and then trained on our dataset using Transfer learning as explained above. 2. Non-TL : Model same as mentioned in point 3 but trained directly on our dataset.

Parameters used :
Kernel Count : 32,
Kernel Size :, [3, 3],
Dynamic Pooling Size : [3, 10]

Table 1. Results.

Dataset	Unicode Encoding (No TL)	Unicode Encoding (TL)	ASCII Encoding (No TL)	ASCII Encoding (TL)
N4	0.767980	0.642020	0.771263	0.648687
N2	0.685290	0.552935	0.694420	0.583804
N1	0.671408	0.599155	0.682160	0.557793

2.1 Task1B

First we created a bag of words representation, as a term document matrix, for the citances and the reference text in the annotations. We have used ngrams, as suggested in the NBSVM paper. Then we find the log count ratio for each class. We fit a model for one class at a time. NBSVM is identical to the SVM, except we use x (k) = f (k) , where f (k) = r f (k) is the elementwise product. Here, x(k) is the feature set given to SVM and f (k) are the original features.
We had total of 748 samples in the training set. We split this into training and validation set in the ratio 4:1. Both Citance and Reference text are used as features. We obtained the highest accuracy of 90.303% on the validation dataset.

3 Conclusion

Submitted system runs for the following models :
1. Model trained on N4 dataset with no transfer learning.

2. Model trained on N4 dataset with transfer learning.
3. Model trained on N2 dataset with no transfer learning.
4. Model trained on N1 dataset with transfer learning.

References

Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S., & Cheng, X. (2016, February). Text Matching as Image Recognition. In *Proceedings of the AAAI Annual Meeting* (pp. 2793-2799).

Wang, S., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*: Short Papers-Volume 2 (pp. 90-94). Association for Computational Linguistics.