# SciSumm 2017: Employing Word Vectors for Identifying, Classifying and Summarizing Scientific Documents

Aniket Pramanick, Salma Mandi, Monalisa Dey, and Dipankar Das

Jadavpur University, Kolkata, West Bengal, India

**Abstract:** This paper describes our approach on "Recognizing Reference Spans,Classifying Their Discourse Facets and Summarizing from Reference Text" as an attempt in the shared task on relationship mining and scientific summarization of computational linguistics research papers at SIGIR 2017.

## 1 Introduction

The 3rd CL-SciSumm Shared task provides resources for scientific paper summarization.An overview of the shared task,including specific details on the dataset and subsequent analysis for each task.In this report we provide a short description of the methods we have used for the task 1A and 1B.

### 1.1 Summary

This system is a rule based implementation of Artificial Neural Network.

## 2 Dataset and Preprocessing

The original reference is divided into sentences. The similarity between each of the sentences to the cited sentence in the citance is measured in using three standard measuring rules: a. Jaccards Coefficient (Trigram Model)(J(a, b)) b. Clough and Stevenson Coeffecient (Trigram Model) (C(a,b)) c. Model Probability Measure (Bigram Model) (P(a))

Thus for each sentence in the reference text we get an 3-tuple (J, C, P). Now this vector is fed to an Artificial Neural Network to find whether the sentence in the reference text is cited in the citance.

Using the outputs of this Artificial Neural Network we get the *Candidate Sentences*, the subset of which is the set of *"Citation Sentences"*.

Actually this Neural Network is used as an Filtering method.

## 3 System Framework

### 3.1 Task 1A:Identification

A very simple method has been used to identify the possible *Citation Texts*.

For each *Candidate Sentence* in the *Reference Text* the cosine similarity to the *Cited Sentence* is measured using cosine similarity, and the Cosine Similarity score is incremented by unity. Thus, each Candidate Sentence of a *Reference Text* gets assigned to a Cosine Similarity Score.

If a *Reference Text* has no *Candidate Sentence* with score greater than 1.2, we say that the *Reference text* has not been cited at all.

Otherwise the sentence or the sentence segment with maximum score is declared to be the *Citation Text*.

The code to measure the Cosine Similarity Score is as follows:

```
from keras.preprocessing.text import Tokenizer, text_to_word_sequence
import math
from scipy import spatial

def cosine_similarity(t1, t2):
        texts=[t1, t2]
        tknzr=Tokenizer(lower=True, split=" ")
        tknzr.fit_on_texts(texts)
        x=tknzr.texts_to_matrix(texts)
        v1=x[0]
        v2=x[1]
        sumxx=0
        sumxy=0
        sumyy=0
        for i in range(len(v1)):
                a=v1[i]
                b=v2[i]
                sumxx=sumxx+a*a
                sumyy=sumyy+b*b
                sumxy=sumxy+a*b
        return (float(sumxy)/float(math.sqrt(sumxx*sumyy))+1)
```

## 3.2 Task 1B:Classification

Task 1B considered as text classification problem.The five discourse facets of a sentence in a reference paper are Aim,Methode,Implication,Result and Hypothesis.For this task we are following an unsupervised method.

- A bag of words are created for each class.
- For each bag compute its bag vector.
- Compute sentence vector for each cited text span.
- For each bag vector measure cosine similarity between sentence vector and bag vector.

The most similarity value corresponding to a bag vector will be assigned as class of cited text.

**Bag of Words:** Each bag is a list of relevent words to a particuler class.Each bag is made based on unigram.For each class we have seperated the reference text from training data set.Then made a list of words from reference sentences and calculated their tf-idf score.Bag is constructed by taking words with highest tf-idf score.

**Bag Vectors:** We used the pre-trained 200 dimensional GloVe(http://nlp.standford.edu/software/CRF-NER.html) on Twitter data 2billion tweets(http://nlp.standford.edu.projects/glove/) to create the vectors of the reference text and the word bags.

The word bag vectors are created by taking the normalized summation of the vector of words in word bags which were present in the vocabulary of the pre-trained Glove model.Out of vocabulary words are assigned to null vector.

$\vec{qi} = \frac{1}{Nv(qi)} \sum_{j=1}^{Nv(qi)} \vec{Wij}$ and $\vec{Wij} = \vec{0}$ where,$\vec{qi}$ = Topic vector of ith word bag,Nv(qi)= Number of words in qi present in vocabulary, $\vec{Wij}$ = Vector of jth word in ith word bag.

**sentence vectors** The sentence vectors are created by taking the normalized summation of the vectors of the words in the sentence,which were present in the vocabulary of the pre-trained GloVe model.In cases where the word was not a part of the model vicabulary,it was assigned to the null vector.

$\vec{ti} = \frac{1}{Nv(ti)} \sum_{j=1}^{Nv(ti)} \vec{uij}$ and $\vec{uij} = \vec{0}$ Where,$\vec{ti}$ = Sentence vector of ith sentence,ti. Nv(ti)= Number of words in ti present in vocabulary. $\vec{uij}$ = Vector of jth word in ith sentence.

**Cosine Similarity** We used cosine similarity measure to calculate the cosine similarity,S between the sentence vector and the topic vector.

$$S = cosine - sim(\vec{ti}; \vec{qj}) = \frac{\vec{ti}\vec{qj}}{||\vec{ti}||||\vec{qj}||}$$

A high value of S denotes higher similarity between the sentence vector,$\vec{ti}$ and the topic vector $\vec{qj}$ and vice-versa.

### 3.3 Task 2:Summarization

From the outputs obtained from Task 1A and 1B, a community summary had to be formed, which is a structured extractive summary of the Reference Paper (RP) generated from the cited text spans of the RP.

**Community summary** The output of Task 1b, contained the cited text spans, along with their facets for each RP. The five discourse facets of a sentence in a reference paper are Aim, Method, Implication, Result and Hypothesis. For each facet, duplicate entries and stop words were removed from the text spans. A similarity score was calculated between the text spans for each facet using the cosine similarity measure. If the cosine similarity score was high, then one out of the two sentence vectors were selected randomly as a probable candidate for summarization. It was analysed from the output that any sentence with word length less than three, contributed no meaning to the summary generated and hence were discraded.

Table 1: Performance of our System in Task 1

| | | Run1 | Run2 |
|---|---|---|---|
| Task 1a Micro Avg | Precision | 0.045 | 0.051 |
| | Recall | 0.031 | 0.035 |
| | F1 | 0.037 | 0.042 |
| Task 1a Micro Avg | Precision | 0.057 | 0.066 |
| | Recall | 0.037 | 0.046 |
| | F1 | 0.045 | 0.054 |
| Task 1a ROUGE2 | Precision | 0.058 | 0.058 |
| | Recall | 0.132 | 0.132 |
| | F1 | 0.065 | 0.065 |
| Task 1b Micro Avg | Precision | 0.045 | 0.051 |
| | Recall | 0.031 | 0.035 |
| | F1 | 0.037 | 0.042 |
| Task 1b Micro Avg | Precision | 0.000 | 0.400 |
| | Recall | 0.000 | 0.057 |
| | F1 | 0.000 | 0.100 |

Table 2: Performance of our System in Task 2

| | | Run1 |
|---|---|---|
| Vs. Abstract - ROUGE 2 | Precision | 0.149 |
| | Recall | 0.278 |
| | F1 | 0.191 |
| Vs. Abstract - ROUGE SU4 | Precision | 0.091 |
| | Recall | 0.289 |
| | F1 | 0.133 |
| Vs. Human - ROUGE 2 | Precision | 0.243 |
| | Recall | 0.152 |
| | F1 | 0.181 |
| Vs. HUman - ROUGE SU4 | Precision | 0.249 |
| | Recall | 0.099 |
| | F1 | 0.129 |
| Vs. Community - ROUGE 2 | Precision | 0.135 |
| | Recall | 0.138 |
| | F1 | 0.132 |
| Vs. Community - ROUGE SU4 | Precision | 0.133 |
| | Recall | 0.138 |
| | F1 | 0.119 |

# 4 Evaluation

We have submitted two different runs for Task 1a and 1b and a single run for Task 2. Task 1a and 1b is scored by the overlap of text spans measured by number of sentences in the system output vs gold standard. Task 2 is scored using the ROUGE family of metrics between i) the system output and the gold standard summary fromt the reference spans ii) the system output and the asbtract of the reference paper. The performance of our system in Task 1 and 2 is shown below in Table 1 and Table 2 respectively.

# 5 Conclusion and Future Work

In this paper,we presented a brief overview of our system to address automatic paper summarization in the Computational Linguistics domain.Recognizing the cited text spans and determining their discourse facets are very challenging task for the summarization of scientific papers.We have observed, task 1A involves much more than similarity problem.More features

that reflect the citation intentions should be explored. For task 1B,building word bags which contain all the topic words relevant to the facet showed better results than the rest.We could do that approach better using bigram or trigram. Task 2 evaluations show that more features have to be introduced in order to calculate the importance of the cited text spans.

# 6 References


- Kokil Jaidka,Muthu Kumar Chandrasekaran,Sajal Rustagi and Min-Yen Kan(2016).Overview of the 2nd Computational Linguistics Scientific Document Summarization Shared Task(CL-SciSumm 2016),To appear in the proceedings of the Joint Workshop on Bibliometric-enhanced Informatilon Retrival and Natural Language Processing for Digital Libraries(BIRNDL 2016),Newwork,New Jersey,USA.

- Surojeet Dasgupta,Abhash Kumar,Dipankar Das,Sudip Kumar Naskar,Shivaji Bandyopadhyay.Word Embeddings for Information Extraction from Tweets.Microblog Track at Forum for Information Retrival Evaluation(FIRE)2016.