# CLSciSumm Shared Task: On the Contribution of Similarity measure and Natural Language Processing Features for Citing Problem

Elnaz Davoodi*, Kanika Madan*, Jia Gu
(* Equal contribution)

Thomson Reuters, Center for Cognitive Computing,
120 Bremner Blvd, Toronto, ON, M5J 3A8, CA

{elnaz.davoodi, kanika.madan, jia.gu}@thomsonreuters.com

**Abstract.** This paper introduces our system submitted to the CLSciSumm 2018 Shared Task at the BIRNDL 2018 Workshop. Our model is trained on a corpus of 40 articles of training set and a corpus of 20 articles from CL-SciSumm 2018. For the purpose of model training, we use random sampling from the articles. We build an ensemble classifier to predict sentences in the reference articles. Also, a multilabel classifier is built to predict the discourse facet of each citation instance. We evaluate the performance of our models using 10-fold cross validation.

**Keywords:** Scientific reference prediction, Ensemble learning, Text Summarization, Discourse Structure in scholarly discourse, Natural Language Processing, Computational Linguistics



## 1 Introduction

Digital documents cite each other frequently, for example, academic papers, news articles, and legal documents usually have citations to each other. In the scientific domain, these citations are even more valuable as these help the researchers to collaborate, acknowledge and extend the research work. The CL-SciSumm Shared Task 2018 focuses on identifying these citation-relationships between the citing and the reference papers by using computational linguistics, natural language processing and text summarization. Text summarization helps to identify the different components of the papers to be able to better identify the cited text in the reference paper.

The dataset in CL-SciSumm 2018 [1] contains sets of Reference Papers (RP) and Citation Papers (CP). The Citation papers contain citations to the Reference papers, and in each such citation paper, the text spans (citances) to a specific citation in the reference paper have been provided in the dataset. The tasks are divided into the following components:

1. (a) For each citation in the citation paper to the reference paper, identify the text spans in the reference papers that contain the given citance. These text spans can be any number of consecutive sentences between 1 and 5, inclusive.
   (b) Given a pre-defined set of facets, identify which section does the cited text identified in (a) belongs to.

2. The bonus task consists of generating a summary of the cited text spans in the reference papers, with a word limit of 250 words.

In this paper, we discuss how we approached these tasks. For Task 1A, we trained a Gradient Boosting Tree classifier on a set of 50 features extracted from the citing and reference citance texts. For Task 1B, we trained a Random Forest classifier on these 50 features on the text from Task 1 to predict the respective facet.

## 2 Methodology

### 2.1 Task Description

This paper explains our approach and results for task 1 of CL-SciSumm 2018, which consists of two subtasks. The training set consists of 40 topics of documents. Each annotated document contains Reference Paper (RP) and Citing Papers (CPs). Each instance in the annotated document refers to a text span in the CP referring to a text span in the RP. In addition, each instance contains a discourse facet which shows the type of relation between the text span in the RP and the corresponding text span in the CP. The first subtask (task 1A) is focused on finding the text span in the RP given a text span in the CP and the goal of the second subtask (task 1B) is to predict the discourse facet.

### 2.2 Task 1A

The goal of this subtask is to find the most relevant sentences (text span) in the reference document, give a text span in the citance. We treat this problem as a binary classification problem by considering the instances give in the annotations as positive instances and sampled negative instances from the reference article. We use various classes of features, including sentence similarity measures, natural language processing features, semantic similarity of reference and citance text spans, etc. We categorize the features into classes of features and provide a brief explanation of each feature as shown in Table 1.

These features can be broadly classified into the following categories:

1. **Similarity based features**:
    a. n-gram based similarity: we converted each of the texts from the citance and reference paper into n-grams, and then applied a number of similarity based metrics on these n-grams. We used n-grams from 1 to 5.
    b. chunk-based similarity: we extracted the noun and verb phrases from the two texts of the citance and reference papers. and calculate the similarity between these extracted chunks.
    c. embedding-based similarity: we used the glove embeddings to generate a summary vector for each of the two texts of the citance and reference papers, and generate a similarity feature using the cosine similarity between the two vectors.
    d. Character and token match: we generated these features by finding jaccard similarity between the characters and tokens of the two texts.

2. **Positional features:** These features capture the token/character positional match.
    a. character match offset features consider the character level match between the two texts of the citation and reference texts.
    b. token match offset features are generated using the token level match between the two texts of the citation and reference texts.
    c. lemma match offset features are calculated using the lemma match between the two texts of the citation and reference texts.
    d. minimal spanning tree based features take into consideration the distance between the common nodes of the two texts of the citation and reference texts.

3. **Frequency based features:** we generated these using the common word and WordNet synonym frequency between the two texts from the citation and reference texts.

**Table 1. List of feature categories used in both Task 1A and Task 1B.**

| Feature Type | Feature category | Feature Definition |
|---|---|---|
| Similarity based features | Word_embedding_similarity | Cosine similarity of the average embedding vectors for citation and reference sentences (used glove pre-trained embedding) |
| | Jaccard_similarity | Jaccard similarity of all tokens from ref and citing sentences |
| | n-gram_Lemmatized | Jaccard similarity of lemmatized and tokenized ngrams for citation and reference sentences using NLP4J |

| | | |
|---|---|---|
| | unigram_Lemmatized | Jaccard similarity of lemmatized and tokenized unigrams for citation and reference sentences using NLP4J |
| | char_match_np | Avg/Min/Max Character match scores between nouns and noun phrases for citation and reference sentences using NLP4J for POS tagging |
| | char_match_vp | Avg/Min/Max Character match scores between verb phrases for citation and reference sentences using NLP4J for POS tagging |
| Frequency based features | Common_word_freq_pos | Relative frequency of common words between reference and citing sentences filtered by POS tags (V, N, Adj, Adv) |
| | Common_syn_freq_pos | Relative frequency of common WordNet synonyms between referencesq2 and citing sentences filtered by POS tags (V, N, Adj, Adv) |
| Positional Features | avg_lemma_word_offset | Create a match set of words with common lemmas in the citation and reference sentences. From each word from this match set, create an offset of the word indices and take an average of these offsets. |
| | avg_match_depth | Create a match set of words with common lemmas in the citation and reference sentences. Take an average of the min depths in the dependency trees for each word in the above match set. |
| | avg_min3_match_depth | From the depths of common nodes in the feature “avg_match_depth”, take three words with min match depth and take an average over these |
| | avg_min_tree_dist | Create a match set of words with common lemmas in the citation and reference sentences. For each pair of nodes in this set, create a minimal spanning tree from the dependency tree such that the distance of the nodes of the two words is minimized from the root. For each word pair, find the distance between them by taking sum of distances of each word from the root. Take an average of these distances. |

| | | |
|---|---|---|
| | avg_sym_diff | Avg of symmetric difference of tokens in sliding windows from texts of the ref and citing sentences |
| | tok_match_score | Max of jaccard similarity between set of all noun phrases and verb phrases from words in the two texts from the ref and citing sentences using NLP4J for POS tagging. |

### 2.3 **Task 1B**

The goal of this subtask is for a give text span in reference and citing paper, we predict the discourse facet. Discourse facets are pre-defined categories and each instance can have multiple discourse facets. So, we treat this problem as a multilabel classification problem. We use the same set of features as explained in Table 1. We also use the reference text span predicted in task 1A.

## 3 Experimental Results

For training our classifier, we generated a training set by sampling positive and negative instances from the dataset. For instances labeled positive, we generated reference and citance text pairs from the lists of reference and citance texts provided. For negative instances, we generated the negative labeled pairs by sampling sentences in the citance paper which have not been provided in the citance text.

We experimented with two values for the negative-to-positive sampling ratios: (a) sample one negative instance from the reference text for each reference and citance pair, and (b) sample two negative instances from reference text for each reference and citance pair. We trained a Gradient Boosting Tree classifier on 50 text-based features using 10-fold cross validation.

Table 2 & Table 3 show the performance of our 10-fold cross validation for task 1A.

**Table 2. Performance of model trained for Task 1A with negative to positive sample ratio = 1.**

| **Label** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|
| 0 | 0.97 | 0.97 | 0.97 |
| 1 | 0.96 | 0.97 | 0.97 |
| Avg | 0.97 | 0.97 | 0.97 |

**Table 3. Performance of model trained for Task 1A with negative to positive sample ratio = 2.**

| Label | Precision | Recall | F1-score |
|---|---|---|---|
| 0 | 0.98 | 0.99 | 0.98 |
| 1 | 0.97 | 0.95 | 0.96 |
| Avg | 0.98 | 0.98 | 0.98 |

As explained above, for Task 1B, we used the same features as in Task 1A, and converting this problem to a multi class classification problem. We used Random Forest classifier for this, and used 10-fold cross validation for evaluation.

Table 4. shows the performance of our model trained for Task 1B.

**Table 4. Performance of model trained for Task 1B.**

| Precision | Recall | F1-score |
|---|---|---|
| 0.90 | 0.95 | 0.92 |

## References


1. Jaidka, Kokil, et al. "Overview of the CL-SciSumm 2016 shared task." Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). 2016.
2. Jaidka, K., Chandrasekaran, M. K., Jain, D., & Kan, M. Y. (2017). The CL-SciSumm shared task 2017: results and key insights. In Proceedings of the Computational Linguistics Scientific Summarization Shared Task (CL-SciSumm 2017), organized as a part of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2017)
3. Jaidka, K., Chandrasekaran, M. K., Rustagi, S., & Kan, M. Y. (2017). Insights from CL-SciSumm 2016: the faceted scientific document summarization Shared Task. International Journal on Digital Libraries, 1-9.Jaidka, K., Chandrasekaran, M. K., Rustagi, S., & Kan, M. Y. (2017). Insights from CL-SciSumm 2016: the faceted scientific document summarization Shared Task. International Journal on Digital Libraries, 1-9