Helium @ CL-SciSumm-19 : Transfer learning for effective scientific research comprehension

Bakhtiyar Syed[1], Vijayasaradhi Indurthi[1], Balaji Vasan Srinivasan[2], and Vasudeva Varma[1]

[1] IIIT Hyderabad
{syed.b, vijaya.saradhi}@research.iiit.ac.in, vv@iiit.ac.in
[2] Adobe Research
balsrini@adobe.com

Abstract. Automatic research paper summarization is a fairly interesting topic that has garnered significant interest in the research community in recent years. In this paper, we introduce team Helium's system description for the CL-SciSumm shared task colocated with SIGIR 2019. We specifically attempt the first task, targeting in building an improved recall system of reference text spans from a given citing research paper (Task 1A) and constructing better models for comprehension of scientific facets (Task 1B). Our architecture incorporates transfer learning by utilising a combination of pretrained embeddings which are subsequently used for building models for the given tasks. In particular - for task 1A, we locate the related text spans referred to by the citation text by creating paired text representations and employ pre-trained embedding mechanisms in conjunction with XGBoost, a gradient boosted decision tree algorithm to identify textual entailment. For task 1B, we make use of the same pretrained embeddings and use the RAKEL algorithm for multi-label classification. Our goal is to enable better scientific research comprehension and we believe that a new approach involving transfer learning will certainly add value to the research community working on these tasks.

1 Introduction

Traditionally, summarization has been a key requirement for facilitating easier comprehension of documents. Automatic summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document. The CL-SciSumm 2019 Shared Task [6] helps in facilitating advances in scientific communication summarization. It encourages the comprehension of information in automatic scientific paper summarization in the form of facet identification, , and the use of new resources, such as the mini-summaries written in other papers by other scholars, and concept taxonomies developed for computational linguistics.

As the number of scientific publications increases, researchers have to spend more time reading them. Reading complete scientific publications to understand

and fully comprehend the content is time-consuming for researchers and enthusiasts alike. In many cases, it even drives away the newcomers from a particular field of study just because there are not enough background articles to aid the process of easier comprehension of highly technical research papers. While abstracts present in the paper help the researchers get a big picture of the paper, they may not cover all the important aspects of the paper. Moreover, the abstracts may be biased, overstated or understated, or the abstracts may not be considered as good summaries by the research community. Automatic summarization of the paper can help capture the details and contributions of a paper more accurately in an unbiased manner as compared to the abstract of the paper. Generating automatic summaries of scientific papers is hence, an important and challenging task.

A citation is a reference to a published or unpublished source in the body of the document. Many digital documents cite other relevant document(s) in their body. News documents, legal documents, wikipedia articles and scientific papers have citations to each other. Citations allow the reader to determine independently whether the referenced material supports the author's argument in the claimed way, and to help the reader gauge the strength and validity of the material the author has used. Citations of a scientific paper help understand the relevant ideas and their evolution. The sentences of the citing articles containing the citation to the publication, also known as the citances are useful in analysing the reference publications and thereby contribute to better summarization of scientific publications.

The rest of the paper is organized as follows. We briefly describe related work in Section 2. We describe and formulate the problem of the shared task in Section 3. Next, our focus shifts to the methodology and the experiments in Section 4. We conclude the paper with mentioning the inferred conclusions and possible directions for future work in section 5.

2 Related Work

The CL-SciSumm series of shared tasks [11, 17, 18] has garnered much attention and attracted many contributions from the research community.

A large number of related works exist as this shared task has been running since 2016. For the subtask 1A, which comprises of identification of text spans according to citations, the solution techniques can be categorized broadly into two types - solutions based on retrieval task and solutions based on classification task. The former methods formulate the problem as an information retrieval problem - learning to rank and selecting the top item from the ranked list.

For subtask 1A, Felber et al. [10] created an index of the reference papers and treating each citance as a query and the results were ranked by VSM and BM25 model. Prasad et al. [24] used tf-idf and LCS for the syntactic score and pairwise neural network ranking model to calculate semantic relatedness score. Cao et al. [4] simplify the problem as a ranking problem and select the first item from the ranked list retrieved.

The classification methods include works from Ma et al. [16], Cao et al. [4], Zhang et al. [29], Li et al. [14]. Ma et al. [16] used four classifiers with different features and a majority voting mechanism to vote for the final result. Cao et al. [4] use SVM with features like tf-idf, named entity features and position information of the reference sentence. Zhang et al. [29] computed features based on sentence-level and character-level tf-idf scores and word2vec similarity and then used a logistic regression classifier to classify sentences which will be selected or not. Li et al. [14] aggregated the results from several basic methods and used majority voting for the final result. Wang et al. [28] use Information Retrieval Model by incorporating word embeddings and domain ontology. L. Moraes et al. [9] use a sentence similarity method using Siamese Deep Learning Networks [21] and a Positional Language Model approach [15]

For subtask 1B, which comprises of identification of the facet, almost all the teams have formulated this as a text classification problem Classification methods for subtask 1B can be divided into rule based methods and supervised learning methods using a gamut of features ranging from TF-IDF together with a variety of supervised machine learning algorithms. Moraes et al. [9] use SVMs, Random Forests, Decision Trees and Multi layer Perceptron and an ensemble method AdaBoost using TF-IDF features. Some teams like Lauscher et al. [13]. have used deep learning text classification techniques like Convolutional Neural Networks (CNNs).

It can be observed that most of the researchers have formulated subtask 1A as a retrieval task than a classification task.

3 Problem Definition and Formulation

In this section, we define the problem formally. Our team, Helium participated in Task 1 - comprising of subtasks 1-A and 1-B specifically. We will introduce the problem definition and then proceed to describe our formulation of the problem.

3.1 Task Description

Given: A topic consisting of a Reference Paper (RP) and Citing Papers (CPs) that all contain citations to the RP. In each CP, the text spans (i.e., citances) have been identified that pertain to a particular citation to the RP.

Task 1-A: For each citance, identify the spans of text (cited text spans) in the RP that most accurately reflect the citance. These are of the granularity of a sentence fragment, a full sentence, or several consecutive sentences (no more than 5).

Task 1B: For each cited text span, identify what facet of the paper it belongs to, from a predefined set of facets. These pre-defined list of facets as defined by the organizers are *Method, Aim, Hypothesis, Implication, Results.*

We pose the problem of finding reference text spans from the citing sentences (Task 1A) as a sentence-pair classification problem. The closest problem in literature is the Paraphrase Detection problem [26] and the Entailment detection problem [8].

For Task 1B, since there are multiple instances in the dataset of two or more labels for each citance text within the CP, we formulate the problem as one belonging to the multilabel category of problems.

3.2 Dataset Creation

The task organizers provided us with 1018 Reference Papers(RPs) and their corresponding Citing Papers(CPs) along with their citations to the particular RP. Each citation consists of a *Reference Offset*. There may be one or more *Reference Offsets* for a particular citation. These offsets point to the sentence IDs in the RP wherein the citation is being referred to.

For our task, we simplify the dataset to create pairs for each citance text in the CP. Each citance text in the CP is paired with the reference text (corresponding to the reference offset). Each pair was labeled with a 1/0 binary variable to show if there is a possibility of entailment of the reference text with respect to the citance text. From the training set of 2018, we collect 180,685 such pairs. Similarly, we were able to collect 3,398,218 such pairs from the newly released 2019 dataset. As we pose task 1B of scientific facet comprehension as a multilabel classification task, we build the dataset from the existing labels available to us. Since the existing labels are only available for the 2018 dataset, we restrict ourselves to the 752 cleaned citance texts from the CPs for our prediction task.

4 Methodology and Experiments

In this section, we go over the critical aspects of our team's submitted system to the shared task. We first provide an overview of pre-trained text representations along with how we set the stage for the transfer learning process. Next, we look at each of the sub-tasks in detail and the algorithms used for final predictions for the dataset given.

4.1 Pre-trained text representations and Transfer Learning

Pre-trained text representations have been used fairly well for a number of natural language processing (NLP) tasks [19]. Unsurprisingly, performance of various NLP tasks have improved as a result of using pre-trained representations, most often as the embedding layer as the first step of a deep neural network architecture [27] . Word embeddings have been widely used in modern NLP applications as they provide a ready-to-use simple vector representation of words. They capture the semantic properties of words and the linguistic relationship between them. These word embeddings have improved the performance of many downstream tasks across many domains like text classification, machine comprehension etc. [3]. Multiple ways of generating word embeddings exist, such as Neural Probabilistic Language Model [2], Word2Vec [20], GloVe [22], and more recently ELMo [23]. These word embeddings rely on the distributional linguistic hypothesis. They differ in the way they capture the meaning of the words or the

way they are trained. Each word embedding captures a different set of semantic attributes which may or may not be captured by other word embeddings. In general, it is difficult to predict the relative performance of these word embeddings on downstream tasks. The choice of which word embeddings should be used for a given downstream task depends on experimentation and evaluation.

While word embeddings can produce representations for words which can capture the linguistic properties and the semantics of the words, the idea of representing sentences as vectors is an important and open research problem [7]. Finding a universal representation of a sentence which works with a variety of downstream tasks is the major goal of many sentence embedding techniques. A common approach of obtaining a sentence representation using word embeddings is by the simple and naïve way of using the simple arithmetic mean of all the embeddings of the words present in the sentence. Smooth inverse frequency, which uses weighted averages and modifies it using Singular Value Decomposition (SVD), has been a strong contender as a baseline over traditional averaging technique [1]. Other sentence embedding techniques include p-means [25], InferSent [7], SkipThought [12], Universal Encoder [5].

In particular, for our task, we use these pretrained sentence encoders to get a dense vector representations for each of the citance/reference texts and we then use these embeddings as features for building models for the downstream classification tasks. Our team's system focuses primarily on the use of Universal Sentence Encoder [5] to get the text representations which are then used in the transfer learning mechanism to other machine learning algorithms as we show below.

4.2 Finding Reference Text Correspondence (Task 1A)

As reported in section 3, we treat task 1A as a sentence-pair classification algorithm problem where a (reference text - citance text) pair will be classified as a positive class only if the reference text of the RP accurately reflects the citance text of the CP, negative class otherwise. We were able to construct 180,685 such pairs from the 2018 version of the dataset. Similarly, we were able to construct 3,398,218 pairs of citance-reference text from the 2019 dataset. As our pipeline suggests, we go through the following steps:

1. Obtain pre-trained dense sentence representations for the citance text and reference text separately for each pair using the Universal Sentence Encoder.
2. Once the pre-trained representations are available, construct the features for the transfer learning phase in the pipeline. These features are constructed by *element-wise subtraction* of the 512-dimensional dense vector representations. We hypothesise that this step enables carrying forward better features for the next step in the pipeline.
3. Post obtaining the 512-dimensional output from the above step, we use a variant of gradient boosted decision tree algorithm for the binary classification task. Specifically, we employ eXtreme Gradient Boosting (XGBoost). Gradient Boosting utilises the principles of Gradient Descent and Boosting

to form the core of its algorithmic prowess. Boosting, essentially is an ensemble of weak learners where the misclassified records are given greater weight (boosted) to correctly predict them in later models. These weak learners are later combined through a linear combination to produce a single strong learner. With eXtreme Gradient Boosting (XGBoost), we take advantage of the following features[3] of the tree-based boosting algorithm XGBoost:
- Approximation for split-finding.
- Column block for parallel learning.
- Sparsity-awareness.
- Cache-aware access.
- Out-of-core computation.
- Regularized Learning Objective.
- Fast!

Since XGBoost is also used in a lot of machine learning contests which are held competitively[4], we specifically chose this as a successor to the pre-trained representation in the transfer learning pipeline.

4. During the prediction phase: For each of the CPs - for each citance text, we then select the corresponding RP's reference texts and rank the top 5 candidates based on the probabilities of being an entailment. These probabilities are readily available when XGBoost is being used to classify the instances. We finally select the top 5 ranked reference texts from the RP for the corresponding citance text and reflect it in our system's submission.

Due to limitations on the compute memory requirements, we do not run cross-validation over the entire dataset, rather we select a random subset of 500 labels from the entailment category exhibiting the label 1 and similarly a random subset of 500 for the label 0. Our results for subtask 1A are detailed in Table 1.

Class	Precision	Recall	F-1
Reference text accurately reflects citance? (Label 1)	80.75	64.60	71.78
Reference text accurately reflects citance? (Label 0)	70.50	84.60	76.91

Table 1. *Precision, Recall* and *F1* scores (in percentage) for **Sub-Task A**

4.3 Scientific Facet Comprehension (Task 1B)

We have 752 instances of citance text - many of them for which for which more than one labels are given - corresponding to the 2018 dataset. There are 5 possible scientific facets as annotated - *Method, Aim, Hypothesis, Implication, Results.* We seek to build a system which can detect one or more than one labels for the given citance text at hand. Formally, multi-label classification is the problem of

[3] https://www.kdnuggets.com/2017/10/xgboost-concise-technical-overview.html
[4] www.kaggle.com

finding a model that maps inputs x to binary vectors y (assigning a value of 0 or 1 for each element (label) in y). In our case - y - is a 5-vector array for each of the given scientific facets.

Since the number of samples in the dataset is small, deep learning techniques do not perform well. We aim to thus use advantage of ensemble-based methods for the multi-label classification problem. Traditionally, multi-label classification problems have been tackled by problem transformation, i.e., either treating them as a *binary* classification problem - by building a separate classifier for each label, or as *multi-class* classification problem - by creating one binary classifier for every label combination present in the training set(also known as the label powerset transformation). Instead, we use an ensemble of classifiers which help us in the task. Specifically, we use the RAKEL algorithm - which employs random k-label subsets approach wherein there is extensive use of multiple label-powerset classifiers, each classifier trained on a random subset of the original labels for the instances (in our case, 5). Finally, a voting mechanism is employed to find the correct labels. The RAndom k-labELsets (RAKEL) algorithm constructs each member of the ensemble by considering a small random subset of labels and learning a single-label classifier for the prediction of each element in the powerset of this subset. In this way, the proposed algorithm aims to take into account label correlations using single-label classifiers that are applied on subtasks with manageable number of labels and adequate number of examples per label. The authors of the RAKEL algorithm show in their paper that t their experimental results on common multilabel domains involving protein, document and scene classification show that better performance can be achieved compared to popular multilabel classification approaches using the random k-labelsets approach.

For scientific facet comprehension, we first construct dense vector representations from the pre-trained Universal Sentence Encoder for the given citance texts from the citance papers (CPs) and do a 5-fold cross validation on the 752 instances in the dataset. We report precision, recall, F-1, accuracy scores for the same. To combat the harsh metrics (since we are dealing with a multi-label problem and accuracies can be unsurprisingly low), we also report the Hamming loss for the same.

The average *Hamming loss* is 0.222, whereas the average *accuracy* is 43.29%. The statistics for precision, recall and F-1 for each of the scientific facets are reported in Table 2.

Scientific Facet	**Precision**	**Recall**	**F-1**
Method	82.38	61.68	70.45
Aim	14.64	38.02	20.79
Results	28.19	42.89	33.94
Hypothesis	11.65	41.33	17.33
Implication	15.46	17.37	15.26

Table 2. *Averaged* Precision, Recall and F-1 scores (in percentage) over 5-fold cross-validation for each of the scientific facets in **Sub-Task B**

5 Conclusions and Future Work

The results of our experiments lead us to believe that transfer learning can pave way for better scientific comprehension and indeed bettering the cause as the first step towards building automated scientific research summarization systems. At the same time, techniques like utilizing pretrained word and sentence embeddings can help build systems for better understanding of different scientific facets and can aid in effective segmentation of research papers for further processing. Indeed, some of the shortcomings are that the precision and recall scores for a lot of the scientific facets other than the *Method* facet are low. This is because of a high imbalance in the data with the majority class being that of the *Method* facet, and cross-validation effectively tends to delete some of the labels for which we have low data, and tends to give lower performance metrics across some of the other classes. This is also a possible exploration for some of the future directions which can be improved upon. We did indeed notice even lower performances for some of the facets with a few of the other learning algorithms, which we omit in the interest of space.

A few of the future directions can be led by the possibilities of exploring training sentence embedding mechanisms from scratch – on all scientific research papers data for particular domains and see if it helps in increased performance across scientific comprehension tasks. A strong emphasis can also be laid on learning domain-specific features for the cause. This is a promising area and we believe such explorations will indeed benefit the cause of the scientific community in aiding automated scientific summarization.

References

1. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. A simple but tough-to-beat baseline for sentence embeddings (2016)
2. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research **3**(Feb), 1137–1155 (2003)
3. Camacho-Collados, J., Pilehvar, M.T.: From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research **63**, 743–788 (2018)
4. Cao, Z., Li, W., Wu, D.: Polyu at cl-scisumm 2016. In: Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). pp. 132–138 (2016)
5. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al.: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)
6. Chandrasekaran, M., Yasunaga, M., Radev, D., Freitag, D., Kan, M.Y.: Overview and results: Cl-scisumm sharedtask 2019 (2019)
7. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
8. Dagan, I., Dolan, B., Magnini, B., Roth, D.: Recognizing textual entailment: Rational, evaluation and approaches–erratum. Natural Language Engineering **16**(1), 105–105 (2010)

9. De Moraes, L.F., Das, A., Karimi, S., Verma, R.M.: University of houston@ cl-scisumm 2018. In: BIRNDL@ SIGIR. pp. 142–149 (2018)
10. Felber, T., Kern, R.: Query generation strategies for cl-scisumm 2017 shared task. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017) (2017)
11. Jaidka, K., Chandrasekaran, M.K., Rustagi, S., Kan, M.Y.: Overview of the cl-scisumm 2016 shared task. In: Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). pp. 93–102 (2016)
12. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in neural information processing systems. pp. 3294–3302 (2015)
13. Lauscher, A., Glavas, G., Eckert, K.: Citation-based summarization of scientific articles using semantic textual similarity. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017) (2017)
14. Li, L., Zhang, Y., Mao, L., Chi, J., Chen, M., Huang, Z.: Cist@ clscisumm-17: Multiple features based citation linkage, classification and summarization. In: BIRNDL@ SIGIR (2). pp. 43–54 (2017)
15. Lv, Y., Zhai, C.: Positional language models for information retrieval. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 299–306. ACM (2009)
16. Ma, S., Zhang, H., Xu, J., Zhang, C.: Njust@ clscisumm-18. In: BIRNDL@ SIGIR. pp. 114–129 (2018)
17. Mayr, P., Chandrasekaran, M.K., Jaidka, K.: Report on the 2nd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (birndl 2017). In: ACM SIGIR Forum. vol. 51, pp. 107–113. ACM (2018)
18. Mayr, P., Chandrasekaran, M.K., Jaidka, K.: Report on the 3rd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (birndl 2018). In: ACM SIGIR Forum. vol. 52, pp. 105–110. ACM (2019)
19. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. arXiv preprint arXiv:1712.09405 (2017)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
21. Mueller, J., Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
22. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
23. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
24. Prasad, A.: Wing-nus at cl-scisumm 2017: Learning from syntactic and semantic similarity for citation contextualization. In: BIRNDL@ SIGIR (2). pp. 26–32 (2017)
25. Rücklé, A., Eger, S., Peyrard, M., Gurevych, I.: Concatenated p-mean word embeddings as universal cross-lingual sentence representations. arXiv preprint arXiv:1803.01400 (2018)

26. Socher, R., Huang, E.H., Pennin, J., Manning, C.D., Ng, A.Y.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In: Advances in neural information processing systems. pp. 801–809 (2011)
27. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 384–394. Association for Computational Linguistics (2010)
28. Wang, P., Li, S., Wang, T., Zhou, H., Tang, J.: Nudt@ clscisumm-18. In: BIRNDL@ SIGIR. pp. 102–113 (2018)
29. Zhang, D., Li, S.: Pku@ clscisumm-17: Citation contextualization. In: BIRNDL@ SIGIR (2). pp. 86–93 (2017)