Macquarie University at TAC 2014 SciSumm

Diego Mollá **Christopher jones**
Macquarie University
Sydney, Australia
diego.molla-aliod@mq.edu.au
christopher.jones4@students.mq.edu.au

Abstract

For the SciSumm track, we used the same settings as in our participation in the TAC 2014 BiomedSumm track. In this report we explain our experiments for the SciSumm track. We tried the same runs as for our participation in the TAC 2014 BiomedSumm track, except for runs that use UMLS since the domain of the documents in the SciSumm track was not biomedical. We also tried a variant of task 2 of the BiomedSumm track, where we aim to recreate the document abstracts by incorporating information from the citing papers. In general, the impact of the different methods used in the runs was the same as for the BiomedSumm track, but the absolute values of the results was lower.

1 Introduction

We wanted to test whether the results obtained in our runs in the TAC 2014 BiomedSumm track were also applicable to the domain of SciSumm. In particular, we replicated the runs of task1a, except for a run that used UMLS. In addition, we tried the same runs of task2 for a variant of task 2 where the goal is to recreate the abstract section of the reference paper. We observed that the impact of information from the citances is positive for biomedical publications (Mollá et al., 2014), and we wanted to know whether this is also true for the Computational Linguistics domain.

2 Finding the Best Fit to a Citance

Given the text of a citance, our system ranks the sentences of the reference paper according to its similarity to the citance. In all cases we modelled each sentence and citance as a vector, and used cosine similarity. We experimented with different forms of representing the information in the vectors, and different forms of using the similarity scores to perform the final sentence ranking.

2.1 Using *tf.idf*

In our simplest approach, we computed the *tf.idf* of all lowercased words, without removing stop words. We computed separate *tf.idf* statistics for each reference paper. In particular, for each reference paper we computed the *idf* component using the set of sentences in the paper and the citance text of all citing papers.

2.2 Adding texts of the same topic

Since the amount of text used to compute the *tf.idf* in Section 2.1 was relatively little, we extended the information used to compute the *idf* component by adding the complete text of all citing papers. Given that the citing papers are presumably of the same topic as the reference paper, by adding this text we hope to include complementary information that can be useful for computing the *idf* information.

2.3 Adding context

Further to the approach in Section 2.2, we extended the information of each sentence in the reference paper by including the neighouring sentences within a context window of 20 sentences centered in the target sentence.

2.4 Re-ranking using MMR

The last experiment used Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) to rank the sentences. All sentences were represented as *tf.idf* vectors of extended information as described in Section 2.3. Then, the final score of a sentence was the combination of the similarity with the citance and similarity of the other sentences of the summary according to the formula shown in Figure 1. We chose a value of $\lambda = 0.97$.

2.5 Evaluation and Results

Our evaluation is based on ROUGE (Lin, 2004). ROUGE is a popular evaluation method for summarisation systems that compares the text output of the system against a set of target summaries. In our case, the output is the set of selected sentences, and the target summaries are the sentences given by the annotators. Since ROUGE uses the actual contents words, and not the offset information, we expect that this metric will give non-zero results for cases when the system chooses a sentence that is similar to, but not exactly, the one chosen by the annotator.

Since all of the approaches were unsupervised, we used all the data without having to perform cross-validation experiments.

Table 1 summarises the results of our experiments with both the SciSumm data and the BiomedSumm data. In all results the systems were designed to return 3 sentences, as specified in the shared task. We also ignored all short sentences (under 50 characters) to avoid including headings or mistakes made by the sentence segmentation algorithm.

We observe a similar improvement of results in both domains, with the exception that MMR does not improve over the run that uses *tf.idf* over context in SciSumm, whereas there is an improvement in BiomedSumm. The absolute values are better in the BiomedSumm data, and looking at the confidence intervals we can presume that the difference between the best and the worst run is statistically significant in the BiomedSumm data. The results in the SciSumm data are poorer in general and we cannot find any values that are statistically significant. But given that the amount of data in SciSumm is smaller, and given that the improvement mirrors that of the BiomedSumm data, it seems that, in general, as we add more information to the models that compute *tf.idf*, the results improve. So we can presume that approaches that gather related information to be added for computing the vector models will produce even better results. The results with MMR appears to be contradictory across the two domains but the difference is so small that it might not be statistically significant even when we add more evaluation data.

Table 2 shows the ROUGE-L F1 scores of each individual reference document from the SciSumm dataset, for comparison with runs by other teams.

3 Building the Final Summary

We wanted to test whether information from the citances are useful for building an extractive summary, as is the case with the BiomedSumm data (Mollá et al., 2014). We therefore implemented extractive summarisation systems with and without information from the citances.

The summarisers without information from the citances scored each sentence as the sum of the *tf.idf* values of the sentence elements. We tried the *tf.idf* approach described in Section refsec:tfidf.

The summarisers with information from the citances scored each candidate sentence i on the basis of rank(i,c) obtained in task 1a, which has values between 0 (first sentence) and n (last sentence) and represents the rank of sentence i in citance c:

$$\text{score}(i) = \sum_{c \in \text{citances}} 1 - \frac{\text{rank}(i, c)}{n}$$

The summaries were evaluated using ROUGE-L, where the model summaries are the abstract section of the corresponding papers. Since paper X96-1048 of the SciSumm data did not have an abstract section, it was removed for this experiment. One of the 20 documents from the BiomedSumm was also removed for the same reason. Table 3 shows the evaluation results.

Again, we observe a similar pattern in both domains, and again the absolute values of the results are higher in the BiomedSumm task. The versions that use the data from task1a are better than the versions that do not use information from task1a. The difference is statistically significant with the

$$\text{MMR} = \arg \max_{D_i \in R \setminus S} \left[\lambda(\text{sim}(D_i, Q)) - (1 - \lambda) \max_{D_j \in S} \text{sim}(D_i, D_j) \right]$$

Where:

- Q is the citance text.
- R is the set of sentences in the document.
- S is the set of sentences that haven been chosen in the summary so far.

Figure 1: Maximal Marginal Relevance (MMR)

	SciSumm				BiomedSumm			
Run	R	P	F1	CI	R	P	F1	CI
tf.idf	0.316	0.198	0.211	0.185–0.240	0.273	0.326	0.279	0.265–0.293
topics	0.324	0.201	0.217	0.191–0.245	0.288	0.357	0.300	0.285–0.316
context	0.339	0.214	0.225	0.197–0.255	0.291	0.372	0.308	0.293–0.323
MMR	0.335	0.212	0.223	0.195–0.251	0.290	0.375	0.308	0.293–0.323

Table 1: ROUGE-L results of our runs for task 1a

Paper ID	*tf.idf*	topics	context	MMR
C90-2039_TRAIN	0.235	0.229	0.235	0.228
C94-2154_TRAIN	0.298	0.307	0.288	0.288
E03-1020_TRAIN	0.274	0.240	0.239	0.243
H05-1115_TRAIN	0.209	0.271	0.350	0.311
H89-2014_TRAIN	0.302	0.315	0.332	0.333
J00-3003_TRAIN	0.209	0.206	0.196	0.199
J98-2005_TRAIN	0.105	0.113	0.101	0.108
N01-1011_TRAIN	0.214	0.228	0.221	0.219
P98-1081_TRAIN	0.180	0.192	0.200	0.201
X96-1048_TRAIN	0.238	0.235	0.248	0.247
Micro-average	0.211	0.217	0.225	0.223
Macro-average	0.226	0.234	0.241	0.238

Table 2: ROUGE-L F1 results for individual SciSumm reference papers for task 1a

	SciSumm				BiomedSumm			
Run	R	P	F1	CI	R	P	F1	CI
tf.idf	0.379	0.157	0.214	0.168–0.264	0.293	0.192	0.227	0.190–0.261
task1a *tfidf*	0.396	0.203	0.259	0.207–0.307	0.425	0.266	0.322	0.294–0.355
task1a MMR	0.406	0.199	0.260	0.209–0.306	0.480	0.302	0.364	0.332–0.397

Table 3: ROUGE-L results of our runs for task 2

BiomedSumm data, and it is near the edge of statistically significance with the SciSumm data.

Table 4 shows the breakout of ROUGE-L F1 scores per document, for comparison with the runs by other teams.[1]

4 Conclusions

The results of the experiments reported in this paper suggest that information from related papers may be useful to find the sentences of the reference paper that best match the citances.

Our experiments also suggest that information from the citances may be useful for building an extractive summary. This conclusion is compatible with prior research that suggest that, in general, information from citing papers may be useful for building summaries, as was stated in the original goals of the BiomedSumm and SciSumm shared tasks.

Acknowledgments

This research was made possible thanks to a summer internship granted to Christopher Jones by the Department of Computing, Macquarie University.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, pages 335–336, New York, New York, USA. ACM Press.

Chin-Yew Lin. 2004. {ROUGE}: A Package for Automatic Evaluation of Summaries. In *ACL Workshop on Tech Summarisation Branches Out*.

Diego Mollá, Christopher Jones, and Abeed Sarker. 2014. Impact of Citing Papers for Summarisation of Clinical Documents. In *Proc. ALTA 2014*.

[1] Table 4 has a row "Average" instead of two rows for micro- and macro-average because both micro- and macro-average have the same values.

Paper ID	*tf.idf*	task1a *tf.idf*	task1a MMR
C90-2039_TRAIN	0.347	0.315	0.293
C94-2154_TRAIN	0.095	0.123	0.120
E03-1020_TRAIN	0.189	0.189	0.196
H05-1115_TRAIN	0.134	0.306	0.321
H89-2014_TRAIN	0.294	0.319	0.320
J00-3003_TRAIN	0.221	0.382	0.367
J98-2005_TRAIN	0.221	0.216	0.233
N01-1011_TRAIN	0.187	0.268	0.284
P98-1081_TRAIN	0.241	0.210	0.206
Average	0.214	0.259	0.260

Table 4: ROUGE-L F1 results for individual SciSumm reference papers for task 2