Impact of Citing Papers for Summarisation of Clinical Documents

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

Abeed Sarker
Arizona State University
Tempe, AZ, USA
abeed.sarker@asu.edu

Abstract

In this paper we show that information from citing papers can help perform extractive summarisation of medical publications, especially when the amount of text available for development is limited. We used the data of the TAC 2014 biomedical summarisation task. We report several methods to find the reference paper sentences that best match the citation text from the citing papers ("citances"). We observed that methods that incorporate lexical domain information from UMLS, and methods that use extended training data, perform best. We then used these ranked sentences to perform extractive summarisation and observed a dramatic improvement of ROUGE-L scores when compared with methods that do not use information from citing papers.

1 Introduction

Text-based summarisation is a well-established area of research that aims to automatically produce condensed text representations of the original text. Text-based summarisation is useful in an increasing number of application domains where people cannot afford to spend time to read all the relevant information. This is certainly the case in the medical domain, and several approaches for the automated summarisation of medical text have been proposed, e.g. as surveyed by Afantenos *et al.* (2005).

Information from citing texts has been used in decades-old studies (Garfield et al., 1964). More recently, Nakov *et al.* (2004) proposed the use of citations for the semantic interpretation of bioscience text. They used the text surrounding the citations, which they named "citances", to summarise the original text. Further research focused on the extraction of the citances and surrounding text (Qazvinian and Radev, 2010) and on the use of these citances to gather information about the original text, which could be used as a surrogate of, or in addition to, a summary of the text (Mohammad et al., 2009; Abu-Jbara and Radev, 2011).

The Biomedical Summarization Track of the 2014 Text Analysis Conference (TAC 2014 BiomedSumm Track)[1] was designed as a set of shared tasks that focus on the use of the citances to build summaries of biomedical documents. The track organisers provided a small data set of 20 biomedical documents for training and fine-tuning. Each paper of the data set (henceforth "reference paper") has 10 citing papers, and the data are annotated with the citances found in the citing papers. For each citance, four annotators appointed by the National Institute of Standards and Technology (NIST) identified various pieces of information related to the track tasks. Three tasks were defined:

Task 1a Identify the text spans from the reference paper that most accurately reflect the text from the citance.

Task 1b Classify what facet of the paper a text span belongs to. There are 6 fixed facets: *hypothesis*, *method*, *results*, *implication*, *discussion*, and *data-set-used*.

Task 2 Generate a structured summary of the reference paper and all of the community discussion of the paper represented in the citances.

We have used the data from the TAC 2014 BiomedSumm Track to explore the hypothesis that using the information from citing papers can improve the results of an extractive summarisation

[1] http://www.nist.gov/tac/2014/BiomedSumm/

Diego Molla, Christopher Jones and Abeed Sarker. 2014. Impact of Citing Papers for Summarisation of Clinical Documents. In *Proceedings of Australasian Language Technology Association Workshop*, pages 79−87.

system. Whereas in prior work the information from the citing papers is presented as the summary of the reference paper to form what has been called citation-based summarisation, in this paper we will use the information from the citing papers as a step to select the most important sentences from the reference paper. This way the resulting summaries are directly comparable with standard extractive summarisation methods, and they suffer less from problems of balance, coherence and readability that typically affect summarisation systems based on multiple papers.

In this paper we present experiments with several settings to address task 1a of the TAC 2014 BiomedSumm track, and how these can help for task 2 and build extractive summaries. We have not explored yet how to incorporate the facet of the citances (task 1b) into task 2 and therefore we will not discuss task 1b in the remaining of this paper.

2 Finding the Best Fit to a Citance

Task 1a of the TAC 2014 BiomedSumm track assumes that the citances are known, and the goal is to identify the text spans from the reference paper that most accurately reflect the text from the citance. Figure 1 shows an example of the data for one citance.

To identify the sentences in the reference paper that best fit a citance we have tried several methods, all of which are based on computing the similarity between a sentence in the reference paper and the citance text. In all cases we have modelled each sentence as a vector, and used cosine similarity as the means to determine the closest match. Our methods vary on how the sentence vectors have been built, and how the similarity scores have been used to select the final sentences.

In our initial experiments we obtained best results after lowercasing, but without removing stop words or stemming, so all experiments conducted in this paper preprocess the text in this manner.

2.1 Oracle

We tried an oracle approach in order to have an upper bound. The oracle consists of the output given by one of the four annotators. The evaluation of the oracle was based on comparing each annotator against the other three annotators. The evaluation results are technically not an upper bound because the evaluation method is slightly different for two reasons: first, the annotator that is being used to select the target sentences is dropped from the gold data; and second, each original run is converted into multiple oracle runs, one per annotator, and the final result is the average among these runs. But it gives an idea of how much room for improvement is left by the automatic methods.

2.2 *tf.idf* and SVD

A straightforward method to build the sentence vectors is to compute the *tf.idf* scores of the sentence words. For each reference paper we computed a separate *tf.idf* matrix where the rows are the set of all sentences in the reference paper plus all sentences that appear in the citance text.

We also applied a variant that performs Singular Value Decomposition (SVD) on the *tf.idf* matrix, with the aim to reduce the size of the matrix and hopefully detect possible latent word relations. We tried with 100, 500, and 1000 components. In this paper we show the results for 500 components since it obtained the best results in our preliminary experiments.

2.3 Additional Data

Traditional uses of *tf.idf* rely on relatively large corpora. Given the very small amount of text used to compute *tf.idf* in our scenario (just the reference paper sentences and the set of citances for that reference paper), we expanded the data as follows.

Topics. Given a reference paper, we used the paper sentences plus all sentences of all documents that cite the reference paper, not just the sentences in the citance text. In the TAC data set, each reference paper had 10 citing papers.

Documents. In each reference paper we used all sentences of all documents of the TAC2014 set. This included the documents citing the reference paper, and all other documents.

Abstracts. We added the sentences of a separate collection of 2,657 abstracts extracted from PubMed and made available by Mollá and Santiago-Martínez (2011).[2] There was no guarantee that these abstracts were from any topic related to the reference paper. The resulting dataset may therefore contain noise but the additional text may help determine the important words of a document.

[2] http://sourceforge.net/projects/ebmsumcorpus/

Reference article: Voorhoeve.txt

Citance text: In this context, while the development of TGCTs would be allowed by a partial functional inactivation of p53 (see [53], [54]), such mechanism would be insufficient to counteract the pro-apoptotic function of p53 induced by a persistent damage, causing a rapid cell death

Target reference text: These miRNAs neutralize p53-mediated CDK inhibition, possibly through direct inhibition of the expression of the tumor-suppressor LATS2. We provide evidence that these miRNAs are potential novel oncogenes participating in the development of human testicular germ cell tumors by numbing the p53 pathway, thus allowing tumorigenic growth in the presence of wild-type p53 ... Altogether, these results strongly suggest that the expression of miR-372/3 suppresses the p53 pathway to an extent sufficient to allow oncogenic mutations to accumulate in TGCTs ... However, whereas in the majority of the cases neoplastic transformation will require inactivation of p53 (for example by expression of HPV E6, HDM2, or mutant p53), miR372&3 uniquely allowed transformation to occur while p53 was active

Figure 1: Extract of the data for task 1a. The goal is to identify the target reference text, which is an extract of the reference article. In this example, three extracts are indicated in the target reference text, separated with "...".

2.4 Additional Context

Conventional methods for the calculation of *tf.idf* assume that each document contains a reasonable amount of words. In our case we use sentences, not full documents, and therefore the information is much sparser. It is conceivable that better results may be achieved by expanding each sentence. In our experiments, we expanded each sentence by adding text from neighbouring sentences. A context window of n sentences centred on the original sentence was used. We experimented with context windows of 5, 10, 20 and 50 sentences. In our preliminary experiments we observed best results for a context window of 50 sentences but it was marginally better than 20 sentences and at the expense of computation time so in this paper we use a context window of 20.

2.5 Maximal Marginal Relevance

Maximal Marginal Relevance (Carbonell and Goldstein, 1998) uses a greedy algorithm to approximate the selection of sentences that maximises the similarity between the sentences and a query, while at the same time penalising similarity among the chosen sentences. The algorithm uses a parameter λ that adjusts the contribution of each of these two optimisation criteria, giving the definition shown in Figure 2.

For the experiments reported here we use $\lambda = 0.97$ since it gave the best results in our preliminary experiments.

2.6 UMLS and WordNet

As mentioned above, we used SVD as a means to detect latent word relations. In addition we used domain knowledge to detect explicit word relations. In particular, for every word we used all of its synsets as defined by WordNet (Fellbaum, 1998) to leverage synonymy information. We also used each word's most salient Unified Medical Language System (UMLS) concept ID and corresponding semantic types by means of the MetaMap tool (Aronson, 2001), using MetaMap's default word-sense disambiguation process. We tried several ways of using WordNet and UMLS, including the following:

1. Replace the word with the WordNet or UMLS IDs or semantic types, and apply *tf.idf* as before.

2. Keep the word and add the WordNet or UMLS IDs or semantic types and apply *tf.idf* as before. This way the data contain specific word information, plus information about word relations.

3. Apply the *tf.idf* similarity metrics separately for each information type and return a linear combination of all. We tried several combinations and settled with this one:

   $$0.5 \times w + 0.2 \times c + 0.3 \times s$$

   where w stands for the *tf.idf* of the original words, c stands for the *tf.idf* of the UMLS

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

Where:

- Q is the question sentence. In this paper, we used the citance text as Q.
- R is the set of sentences in the reference paper.
- S is the set of sentences that haven been chosen in the summary so far.

Figure 2: Maximal Marginal Relevance (MMR)

concepts, and s stands for the *tf.idf* of the UMLS semantic types. We did not observe any improvement of the results when incorporating the WordNet synsets in our preliminary experiments and therefore we did not use them in the experiments reported in this paper.

2.7 Results

To evaluate the methods we have used the ROUGE-L F1 score. ROUGE (Lin, 2004) is a popular evaluation method for summarisation systems that compares the output of the system against a set of target summaries. For each citance, we used the target reference text provided by the annotators, except for the Oracle setting, as described above, where the reference text of one annotator was compared against the reference text of the other three annotators.

Table 1 summarises the results of our experiments. The results of the oracle approach are relatively poor. This indicates relatively low agreement among the annotators.

We can observe an improvement of the results when using additional text to compute the *tf.idf* scores. When adding related documents (the 10 citing papers) the results clearly improved. When using a fairly large set of unrelated abstracts alone the results worsened dramatically, but when using the unrelated abstracts *in addition* to the related documents the results improved marginally *wrt.* using related documents. This seems to point that adding more data to form the *tf.idf* models helps up to a point. Ideally, we should add data that are on the topic of the reference paper.

There was also a small improvement of the results when the original sentences were extended within a context window.

The approaches giving the best results have overlapping confidence intervals. This is not surprising, given that prior work has observed that it is very difficult for two different extractive summarisation systems to produce ROUGE F1 scores with non-overlapping confidence intervals due to the long-tailed nature of the distribution of ROUGE F1 scores among different systems (Ceylan et al., 2010). In our case, in addition, the amount of data is fairly small. Nevertheless, it appears that using UMLS improves the results, and whereas MMR gives better results than UMLS, the difference is so small that it might not be worth incorporating MMR. SVD appears to improve the results over plain *tf.idf*, but again the improvement is small and the computation time increased dramatically. None of the methods approached the results of the oracle, so there is room for improvement. Still, as we will show below, these techniques are useful for extractive summarisation.

Note that the best result overall is plain *tf.idf* where the data have been expanded with the citing papers and the sentences have been expanded with a large context window (50, instead of 20). The computation time of this approach far exceeded that of the other approaches in the table, so there is still room for further exploring the use of UMLS, or smart forms to determine the related documents and extending the sentence context.

3 Building the Final Summary

Whereas the goal of task 1a of the TAC 2014 BiomedSumm track was to find the text from the reference paper that most accurately reflects the text from the citances, the goal of task 2 was to build a summary of the reference paper and all of the community discussion of the paper represented in the citances. This task was set as tentative by the organisers of the track. Figure 3 shows the target summary produced by one of the annotators.

It is reasonable to accept that information from the citances will help produce a community-based summary such as the one in Figure 3. It is not

System	R	P	F1	F1 95% CI
Abstracts	0.190	0.230	0.193	0.179 - 0.208
tf.idf	0.331	0.290	0.290	0.276 - 0.303
MMR $\lambda = 0.97$	0.334	0.293	0.293	0.279 - 0.307
SVD with 500 components	0.334	0.295	0.295	0.281 - 0.308
Topics	0.344	0.311	*0.307*	0.293 - 0.321
$0.2c + 0.3s + 0.5w$	0.364	0.294	*0.309*	0.297 - 0.320
MMR $\lambda = 0.97$ on topics	0.345	0.314	*0.311*	0.296 - 0.325
Topics + context 20	0.333	0.334	*0.312*	0.297 - 0.326
$0.2c + 0.3s + 0.5w$ on topics + context 20	0.356	0.307	*0.312*	0.299 - 0.325
Documents + context 20	0.334	0.336	*0.314*	0.299 - 0.327
Documents	0.347	0.325	*0.316*	0.303 - 0.330
Documents + abstracts	0.347	0.327	*0.317*	0.302 - 0.332
MMR $\lambda = 0.97$ on topics + context 20	0.336	0.340	*0.317*	0.303 - 0.331
Topics + context 50	0.341	0.336	**0.318**	0.302 - 0.332
Oracle	0.442	0.484	0.413	0.404 - 0.421

Table 1: ROUGE-L results of TAC task 1a, sorted by F1. The best result is in **boldface**, and all results within the 95% confidence interval range of the best result are in *italics*.

In the article A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors, Voorhoeve et al., performed genetic screens of miRNA to investigate its novel functions; which has implicated two of them as oncogenes. They demonstrated that miRNA-372&3 participate in proliferation and tumorigenesis of primary human cells along with oncogenic RAS and active wild-type p53 by numbing the p53 pathway.

The authors created expression library by cloning most annotated human miRNAs into their vector and made a corresponding microarray for barcode detection. Guo et al, contradicted this by stating that bead-based platform is more flexible and cost-effective for detecting barcode changes.

Voorhoeve et al., observed that in response to mitogenic signals like RAS primary human cells undergo growth arrest; in contrast cells lacking p53 overcame this arrest. They demonstrated that expression of miRNA-372&3 enabled cells to continue proliferating thus causing a selective growth advantage.

Voorhoeve et al., established that miRNA 371-3 cluster suppresses an inhibitor of CDK activity which is essential for development of TGCTs. On further investigation they observed that 3UTR of LATS2 is indeed the direct target of miRNA-372&3.

This article has a huge impact on society as Voorhoeve et al., have indicated that deregulated expression of miRNA-372&3 predisposes cells to carcinogenic events and these miRNA expressions must be carefully controlled during differentiation to prevent progression to cancer. Their expression library has helped in the functional annotation of miRNA encompassing other regulatory functions that result in DNA damage response, differentiation, sensitivity to growth factors, resistance to anti-cancer drugs etc. It remains to be seen how widespread oncogenic miRNAs are; nevertheless, their study has provided a system for uncovering the roles of other miRNAs in tumorigenesis.

Figure 3: Sample target summary for task 2 as given by an annotator

so straightforward to accept that such information would help produce a summary that does not explicitly incorporate the contribution from the citing papers. To test whether information from citing papers is useful even for a standard, non-community-based summary, we altered the data from TAC as follows: We removed the abstract from the reference paper, and used the abstract from the reference paper as the target summary. In other words, we produced training and test data such as is often done in standard text summarisation settings. Figure 4 shows the target summary for the reference paper in our example.

One of the 20 papers from the training set did not include an abstract and it was removed for the modified task.

Below we described our approaches to solve task 2 and its modified version.

3.1 Oracle

The oracle version compared the data of one annotator against that of the other annotators. Again, by building this oracle we have an idea of the difficulty of the task. The oracle was used for the unmodified task 2 but it was not possible for the modified task because only one target abstract was available for each reference paper.

3.2 Using Reference Text Alone

Our first set of experiments used the reference text alone. These methods basically used some of the most common single-document summarisation techniques. In particular we tried the following extractive summarisation techniques, which calculated a score of sentence importance and selected the top sentences:

1. *tf.idf*, SVD: compute the sum of the *tf.idf*, or *tf.idf*+SVD values of the candidate sentence.

2. Additional data, additional context: extend the data or sentence context prior to the computation of *tf.idf* as described in Section 2.

3. UMLS, WordNet: incorporate UMLS and WordNet information as described in Section 2.

3.3 Using Citing Text

We incorporated the citing text in a very straightforward way. For every citance, we used the methods described in Section 2 to rank the sentences. We then scored each sentence i by using rank(i,c), which has values between 0 (first sentence) and n (last sentence) and represents the rank of sentence i in citance c:

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

3.4 Results

Table 2 shows the result of the unmodified task 2, and Table 3 shows the results of the modified version. For the unmodified version we set an upper limit of 250 words per summary, as originally specified in the shared task. For the revised data set we kept the same upper limit of 250 words because it appeared to give the best results.

We observe that the confidence intervals of the best results of the version that uses the TAC data approached the results of the oracle, which is very encouraging, especially given the relatively simple approaches tried in this paper. Overall, we observed that using the scores of task 1a produced much better results than using the information from the reference paper alone. The difference was statistically significant, and given the above mentioned observation that it is generally difficult to obtain ROUGE F1 scores that have a difference that is statistically significant among different extractive summarisers (Ceylan et al., 2010), we have good evidence to the validity of approaches that leverage human knowledge of the paper through the exploitation of the citation links between papers.

Of the traditional methods, that is, the methods that did not incorporate the data of task1a, we observed no significant improvements over a simple *tf.idf* approach. Even adding additional context or documents did not help. This was the case both for the version that used the TAC data and the version that used the abstracts as the target summaries.

We can also observe that the ROUGE scores are higher for the original TAC task than for our modified task. This is compatible with the original goal of the TAC shared task, since the annotators were instructed to build sample summaries that incorporate the information from the citances. In contrast, the target summaries for our modified task were written before the citing papers.

It was interesting to observe that parameters that led to better results in Section 2 did not necessarily achieve best results now. This might be due to random effects, since the results among those settings

Endogenous small RNAs (miRNAs) regulate gene expression by mechanisms conserved across metazoans. While the number of verified human miRNAs is still expanding, only few have been functionally annotated. To perform genetic screens for novel functions of miRNAs, we developed a library of vectors expressing the majority of cloned human miRNAs and created corresponding DNA barcode arrays. In a screen for miRNAs that cooperate with oncogenes in cellular transformation, we identified miR-372 and miR-373, each permitting proliferation and tumorigenesis of primary human cells that harbor both oncogenic RAS and active wild-type p53. These miRNAs neutralize p53-mediated CDK inhibition, possibly through direct inhibition of the expression of the tumor-suppressor LATS2. We provide evidence that these miRNAs are potential novel oncogenes participating in the development of human testicular germ cell tumors by numbing the p53 pathway, thus allowing tumorigenic growth in the presence of wild-type p53.

Figure 4: Original abstract as the new sample target summary

System	R	P	F1	F1 95% CI
Oracle	0.459	0.461	0.458	0.446 - 0.470
tf.idf	0.260	0.264	0.260	0.226 - 0.290
SVD with 500 components	0.264	0.247	0.254	0.236 - 0.272
Topics	0.260	0.265	0.261	0.226 - 0.292
Documents	0.259	0.265	0.260	0.224 - 0.290
Topics + context 5	0.259	0.265	0.261	0.226 - 0.291
Topics + context 20	0.252	0.261	0.255	0.220 - 0.285
task1a (*tf.idf*)	0.384	0.375	0.378	0.350 - 0.408
task1a (MMR $\lambda = 0.97$ on topics)	0.398	0.396	*0.396*	0.372 - 0.421
task1a (MMR $\lambda = 0.97$ on topics + context 20)	0.420	0.407	**0.412**	0.385 - 0.438
task1a ($0.2c + 0.3s + 0.5w$)	0.398	0.392	*0.394*	0.369 - 0.419
task1a ($0.2c + 0.3s + 0.5w$ on topics)	0.405	0.399	*0.401*	0.378 - 0.423
task1a ($0.2c + 0.3s + 0.5w$ on topics + context 20)	0.417	0.404	*0.409*	0.387 - 0.431

Table 2: Rouge-L results of task 2 using the TAC 2014 data. The summary size was constrained to 250 words. In **boldface** is the best result. In *italics* are the results within the 95% confidence intervals of the best result.

System	R	P	F1	F1 95% CI
tfidf	0.293	0.192	0.227	0.190 - 0.261
SVD with 500 components	0.291	0.181	0.218	0.197 - 0.239
Documents	0.289	0.192	0.226	0.188 - 0.260
$0.2c + 0.3s + 0.5w$	0.314	0.210	0.247	0.207 - 0.284
task1a (tfidf)	0.425	0.264	0.*320*	0.293 - 0.353
task1a (MMR $\lambda = 0.97$)	0.418	0.275	*0.324*	0.299 - 0.351
task1a (MMR $\lambda = 0.97$ on topics)	0.436	0.272	*0.330*	0.300 - 0.363
task1a ($0.2c + 0.3s + 0.5w$)	0.439	0.276	*0.333*	0.308 - 0.358
task1a ($0.2c + 0.3s + 0.5w$ on topics)	0.428	0.276	*0.330*	0.304 - 0.357
task1a ($0.2c + 0.3s + 0.5w$ on topics + context 20)	0.451	0.279	**0.338**	0.312 - 0.366

Table 3: Rouge-L results of task 2 using the document abstracts as the target summaries. The summary size was constrained to 250 words. In **boldface** is the best result. In *italics* are the results within the 95% confidence intervals of the best result.

were within confidence intervals.

4 Summary and Conclusions

We have experimented with approaches to incorporate information from the citing papers in an extractive summarisation system. We observed that ranking the sentences of the reference paper by comparing them against the citances improved results over methods that did not incorporate such information. In other words, the information introduced by the community of authors citing a paper is useful to produce an extractive summary of the reference paper. The improvement of results was considerable, and it suggests that a good strategy to build summaries is to focus on finding citing papers and use that information in the summariser.

Given the small amount of data available we did not try supervised methods. It is conceivable that, if further data are available, better results might be achieved by applying classification-based approaches. It would be interesting to test whether supervised methods that rely on larger volumes of annotated training data would also benefit from information from the citing papers. Alternatively, the additional data could be used to produce an additional development set to fine-tune the parameters in the approaches that we have explored in this paper.

Further research includes performing a new evaluation that uses as target summaries annotations from people who have not read the abstract, since it is conceivable that authors of citing papers used text from the abstract of the original paper, and that could explain our comparatively good results.

We observed a general improvement of results when we included additional information at the stage when the underlying models for *tf.idf* were created. Both adding additional sentences, and expanding the existing sentences by adding a context window, helped produce better results. This suggests an additional strategy to improve the quality of summarisation systems: find related documents, and use their information to create better-informed language models of the reference paper.

At the time of submission of this paper the results of the TAC 2014 BiomedSumm track were not available. The organisers of TAC 2014 proposed a different approach to evaluate task 1a, based on a direct comparison of the string offsets of the extracts from the reference papers. We anticipate that such evaluation metrics is probably too strict since it does not accommodate cases where the extract has similar information to the annotated text span. It will be interesting to contrast the TAC evaluation results with our evaluation and observe whether the same conclusions still apply.

5 Acknowledgements

This research was possible thanks to a winter internship by the second author funded by the Department of Computing at Macquarie University.

References

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent Citation-Based Summarization of Scientific Papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 500–509.

Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005. Summarization from Medical Documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177.

A R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21, January.

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.

Hakan Ceylan, Rada Mihalcea, Umut Özertem, Elena Lloret, and Manuel Palomar. 2010. Quantifying the Limits and Success of Extractive Summarization Systems Across Domains. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, number June, pages 903–911. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. Language, Speech, and Communication. MIT Press, Cambrige, MA.

Eugene Garfield, Irving H Sher, and Richard J Torpie. 1964. The Use of Citation Data in Writing the History of Science. Technical Report 64, Institute for Scientific Information, Philadelphia, PA, USA.

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

Saif Mohammad, Bonnie Dorr, and Melissa Egan. 2009. Using citations to generate surveys of scientific paradigms. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (June):584–592.

Diego Mollá and Maria Elena Santiago-Martínez. 2011. Development of a Corpus for Evidence Based Medicine Summarisation. In *Proceedings of the Australasian Language Technology Workshop*.

Preslav I Nakov, Ariel S Schwartz, and Marti A Hearst. 2004. Citances : Citation Sentences for Semantic Analysis of Bioscience Text. In *SIGIR 2004*.

Vahed Qazvinian and Dragomir R Radev. 2010. Identifying Non-explicit Citing Sentences for Citation-based Summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, number 1996.