Vector Space and Language Models for Scientific Document Summarization

John M. Conroy
IDA Center for Computing Sciences
17100 Science Drive
Bowie, MD 20708, USA
conroy@super.org

Sashka T. Davis
IDA Center for Computing Sciences
17100 Science Drive
Bowie, MD 20708, USA
stdavi3@super.org

Abstract

In this paper we compare the performance of three approaches for estimating the latent weights of terms for scientific document summarization, given the document and a set of citing documents. The first approach is a term-frequency (TF) vector space method utilizing a nonnegative matrix factorization (NNMF) for dimensionality reduction. The other two are language modeling approaches for predicting the term distributions of human-generated summaries. The language model we build exploits the key sections of the document and a set of citing sentences derived from auxiliary documents that cite the document of interest. The parameters of the model may be set via a minimization of the Jensen-Shannon (JS) divergence. We use the OCCAMS algorithm (Optimal Combinatorial Covering Algorithm for Multi-document Summarization) to select a set of sentences that maximizes the term-coverage score while minimizing redundancy. The results are evaluated with standard ROUGE metrics, and the performance of the resulting methods achieve ROUGE scores exceeding those of the average human summarizer.

1 Introduction

The volume of the scientific literature is vast and increasing. It is commonly impossible for researchers to read all the papers published even in their own specialty, thus it is natural to apply text summarization methods to scientific literature. The problem we consider is to summarize a scientific paper that has been cited multiple times, given the paper (*reference paper*) and a set of citing papers. Note that the citing papers give additional insights into the impact of the results presented in the original paper and also how the paper is perceived by colleagues. Following the approach of Qazvinian et al. (Qazvinian and Radev, 2008), we use the citing papers to help inform the summary, but also build a language model to cover the major sections of the paper such as the abstract and the results sections. Thus, we form a summary pooling information from the paper and how other authors citing the paper view the contributions of the paper.

The summarization system we consider for this task consists of the following components:

1. **Data Preprocessing and Segmentation**
   The reference document is processed, the individual sections of the paper (when present) are isolated and extracted, and the document is then sentence split.

2. **Term Selection**
   Terms are formed with stemmed word bigrams whose mutual information is significantly high.

3. **Latent Term Weight Estimation**
   We explore two distinct approaches:

   (a) **A Vector Space Model** based on a term-frequency (TF) matrix representation of the document and a nonnegative matrix factorization (NNMF) approximation for rank reduction.

   (b) **A Bigram Language Model** built on the selected bigrams for each document section. The global language model is a

Proceedings of NAACL-HLT 2015, pages 186–191,
Denver, Colorado, May 31 – June 5, 2015.

convex combination of the section models. Each term is given a weight, which is an estimate of the probability that a term would occur in a human-generated summary.

4. **Sentence Selection**
   We use the OCCAMS algorithm (Optimal Combinatorial Coverage Algorithm for Multi-Document Summarization) to select the sentences.

In this paper, we use the Biomedical Summarization data[1] recently released by the National Institute of Standards (NIST) to evaluate our approaches. The Biomedical data consists of 20 documents (*reference* papers), each with 10 documents that cite it (*citation* papers), a human-generated summary, and set of citation sentences extracted from the citation papers. (The citation sentences are the sentences of the citation papers that refer to the reference documents.)

In section 3 we describe in details how these data were preprocessed for the summarization task.

2 Related Work

In (Teufel and Moens, 2002) the authors use rhetorical status of statements in a scientific article to produce a summary. They use machine learning to identify rhetorical structure and produce extracted sentences from the source document by filling in a template to produce a summary. In (Qazvinian and Radev, 2008) the authors use the citation network to produce a summary of a scientific article and thereby put the focus on what other authors wrote about a paper as the prominent information to include in a summary of a paper in the scientific literature. They thus summarize a scientific document by what other authors have written about the document. In contrast with (Teufel and Moens, 2002) and (Qazvinian and Radev, 2008), in this paper, we use the sections of the document and machine learning to estimate the relative importance of the sections of the document as well as the citing sentences that reflect what other authors write about the referenced document. Jensen-Shannon (JS) divergence correlate as well or better with human judgments of a summary's quality than ROUGE scoring for multidocument summarization (Lin et al., 2006). In (Louis and Nenkova, 2009) the authors demonstrated that JS divergence between automatically generated summaries and the distribution of terms of the document yields a metric for evaluating summaries without the need for human-generated summaries. These results suggest that JS-divergence is correlated with the intrinsic quality of a summary. We therefore employ JS-divergence and use both the provided human summaries, and optionally the extracted text, to learn the distribution of terms as would be used in a human-generated summary.

3 Data Segmentation and Preprocessing

Each document from the NIST Biomedical collection contains one reference paper, a set of citation papers, and an annotation file containing citation sentences (*citances*). The vast majority of the biomedical papers (both reference and citation papers) had a common structure. All but two of the papers contained a well-defined *Abstract* or a *Summary* section; most papers contained a *Results* section; and all concluded with a *Reference Bibliography* section. We removed the Reference Bibliography because its content is inappropriate for summarization. The body of the reference paper was partitioned into three parts: abstract, results and other. The abstract and the result parts contained the body of the *Abstract* (or *Summary*) and the *Results* sections, respectively, if they were present. All remaining sections[2] of the paper were extracted and were used for forming the other part of the paper. We also extracted the citances from the annotation file into a separate part, called citations. We used the entire body of the reference paper except the *Reference Bibliography* and the citations part to build the vector space model, and we used the abstract, results, other, and citations parts for the language model approach for latent term weight estimation.

We trimmed the sentences of the abstract, results, other, and citations parts to remove quoted, parenthetical or citation text. Doing this trimming im-

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

[2] Some papers contained subject specific sections or *Methodology*, or *Discussion* sections but they were not common across the entire collection.

proves the fidelity of the summarization. Finally, all parts of the paper were segmented into individual sentences.

4 Term Selection

Term selection first begins by finding a good background model and then using it to select a set of stemmed word bigrams that occur *significantly more often than expected*, specifically, we calculate the frequency of each stemmed bigram in a document set as well as the frequency of the stemmed bigram in a background corpus of biomedical abstracts from PubMed (National Institute of Health, 2014). We employ the $G-$statistic, which is equivalent to a mutual information statistic. This statistic was first suggested by Ted Dunning (Dunning, 1993) to identify "surprise words" and in context of summarization, Lin and Hovy (Lin and Hovy, 2000) referred to them as "signature terms." The statistic computes likelihood ratio and the p-value is computed under the assumption null hypothesis that a given term occurs with the same probability in the background as the document set.

Here, instead of finding a small set of topic signature terms as proposed by (Lin and Hovy, 2000), we use Dunning's statistic to remove terms for which the p-value is 0.001 or larger. For this threshold about 40% of the bigrams will remain in lieu of a topic signature model where $10-50$ bigrams remain.

5 Sentence Selection

We use the OCCAMS algorithm (Davis et al., 2012) to select sentences for the final summary. OCCAMS uses techniques from combinatorial optimization (approximation schemes for budgeted maximal coverage and the knapsack problems) to select a set of minimally overlapping sentences, whose combined weight of terms covered is maximized. OCCAMS views the document as a set of sentences. Each sentence is viewed as a set of terms. The input to the algorithm is the sentences of the document; the lengths of the sentences, measured as the number of words; and the latent weights of the terms, computed in sections 6 and 7, for the two models we study. (Conroy et al., 2013) gives an improved version of the original OCCAMS algorithm (Davis et al., 2012) that computes four potential summaries and chooses the one of maximal combined term-weight coverage as the final summary. OCCAMS has a minimal sentence length parameter that one can use to discard sentences whose lengths (number of words) falls below the specified minimal-length threshold. The biomedical documents we summarized had a higher-than-average sentence length, and we used a threshold of 10 words per sentence, to generate our result summaries. Summaries containing longer sentences improve readability of the summaries generated (sentences containing nine words or less were discarded), while shorter sentences improve scores computed by automatic metrics.

6 Vector Space Model

The paper (Conroy et al., 2013) investigated the performance of a variety of vector space models together with a variety of algebraic dimensionality reduction techniques (LSA, LDA, and NNMF) to summarize multi-lingual documents. In this paper we consider a simple and well known vector space model for text, namely the term frequency model, and explore the use of NNMF to derive improved term weights for scientific document summarization. To build a term-by-sentence matrix we use the abstract, results, citances, and other parts of the reference paper, which is equivalent to taking the entire body of the paper excluding the *Reference Bibliogrpahy*. We use only bigram terms with high mutual information to form the terms of the document. The (i, j)th component of the matrix A is the frequency of the ith term in the jth sentence. We use the MATLAB$^{\text{TM}}$ *nnmf()* function, with 100 random restarts and the alternating least squares option, to compute a rank k approximation of the column stochastic matrix derived from the term-sentence matrix A by scaling the columns to sum to 1. Let $\tilde{A}$ be this column stochastic matrix and the NNMF of this matrix gives $\tilde{A} \approx WH$, where W and H are nonnegative and W has k columns and H has k rows. The weights of the terms given to the OCCAMS algorithm are chosen to be the row sums of WH.

Table 1 shows ROUGE 1, 2, 3, and 4 scores of OCCAMS summaries given estimates of the latent weights of terms for values of the rank $k = 2, 4$, and 35. In our experiments we computed NNMF

System	R1	R2	R3	R4
TF	0.511	0.166	0.065	0.030
NNMF_2	0.509	0.172	0.073	**0.036**
NNMF_4	0.504	0.171	**0.074**	0.036
NNMF_35	**0.518**	**0.176**	0.070	0.033
Avg Human	**0.528**	**0.179**	**0.075**	**0.036**
Best Human	0.572	0.219	0.110	0.071

Table 1: Vector Space Model based on TF and NNMF

approximations for all values of $k \in [1, 50]$ but did not observe improvements of the score beyond the scores shown in Table 1. Our experiments show that the NNMF rank approximation of the matrix results in improved ROUGE scores compared to the baseline TF. It is also worth noting that the ROUGE scores of the best rank approximations are close to those of the averaged human ROUGE scores.

7 A Language Modeling Approach for Scientific Document Summarization

We consider a language modeling approach to estimate the importance of the terms in the referenced document to be summarized. This model is designed to estimate the probability that a term will occur in a human-generated summary of the referenced document. As described in section 3 a referenced document may be divided into abstract, results, and other parts and these in addition to the citances represent the "components" of the document, which are used to build the language model for the referenced document.

Specifically, we let $p_{i,j}^{(d)}$ be the estimate of the probability that term i occurs in document component j for the referenced document d. The estimate $\hat{p}_{i,j}^{(d)}$ is computed by the maximum likelihood estimate using the counts, and we then have

$$\sum_i \hat{p}_{i,j}^{(d)} = 1.$$

The probability that term i will occur in a human-generated summary of document d is given by $q_i^{(d)}$ and we estimated it for the purposes of training it in one of two ways. The first is simply the maximum likelihood model which sums the frequency observed in the human-generated summaries and then normalizes to form a probability distribution. We denote this estimate as $\hat{q}_i^{(d)}$.

The second estimate for $q_i^{(d)}$ uses a discount model and has a free parameter $\lambda^{(d)}$ with $0 \leq \lambda^{(d)} \leq 1$, which is used to compute a convex combination of the estimates $\hat{q}$ and the estimated probability distribution formed by human selected "referenced sentences." The referenced sentences are those sentences in the reference document that best support the information given in the citances and were selected by the humans as they were gathering information to create their summaries. We let $\hat{r}_i^{(d)}$ be the maximum likelihood estimate that term i occurs in a reference sentence. The discount model estimate of $q^{(d)}$ is then given by

$$\tilde{q}^{(d)} = \lambda \hat{r}^{(d)} + (1 - \lambda) \hat{q}^{(d)}.$$

While discount model provides an opportunity to smooth estimates, we defer its study to a later paper since 10 document sets proved insufficient to demonstrate a significant improvement.

7.1 Training the Language Model

We model the distribution of the terms in the human summaries as a simple mixture of the document components. As such, we expect that for a given document set d there exists a set of parameters α_i for $i = 1, 2, ..., k$ where k is the number of components in the document set, such that

$$\tau(\alpha, \lambda) = \sum_i \alpha_i p_{i,j}^{(d)} \approx \tilde{q}^{(d)}. \qquad (7.1.1)$$

Solving this equation in the least squares sense for both the parameters α and λ is the classical method of canonical correlation (Seber, 2004). Alternatively, we could seek to minimize a divergence function such as JS, i.e.,

$$[\alpha, \lambda] = \text{argmin JSD}(\sum_i \alpha_i p_{i,j}^{(d)}, \tilde{q}^{(d)}).$$

As first observed by (Lin et al., 2006) JS-divergence predicts as well as ROUGE, and it has continuous derivatives. The result of the training gives an optimal values of α and λ, and for each term t a term weight $\tau_t(\alpha, \lambda)$, which is an estimate of the probability that a term will occur in a human-generated summary given the component decomposition and the training data. The optional smoothing parameter, λ used for mixing human summaries with the set of human extracts can be forced to 1 and the optimal α can be computed from approximating the human abstract bigram distribution alone giving the term-weights for the unsmoothed probabilities. In addition to the term-weights, recall that a background model is used to discard bigrams with low mutual information as was described in section 4. The low mutual information terms are thus given a weight of 0.

7.2 Two Simple Language Models

Before discussing results of an optimized language model, we consider the special case of equal weighting of each document section. Recall that the sections of interest for summarization were limited to the abstract, results, other parts, and the citing sentences from the documents that reference the document to be summarized. We consider the following two simple language models on the set of significantly high mutual information bigrams. In the first model we build a language model by combining the four sections into one ($\mathrm{L}M_1$), and in the second model we compute the maximum likelihood for each section of the document combine them with equal weighting ($\mathrm{L}M_4^{equal}$). The ROUGE-1, 2, 3, and 4 scores of $\mathrm{L}M_1$ and $\mathrm{L}M_4^{equal}$ as well as the best and average of the nine human summarizer scores are given in Table 7.2. The lengths of both the machine and human generated summaries were limited to 250 words. Each of the differences between the ROUGE scores of $\mathrm{L}M_1$ and $\mathrm{L}M_4^{equal}$ are statistically significant at or above the 99% confidence level. We note that $\mathrm{L}M_1$ performs comparably with TF baseline, whose term weights differ only by a scale factor. Furthermore, the equal weighting model scores higher than the average human and significantly better than even the NNMF methods given in section 6. Finally, we note that $\mathrm{L}M_4^{opt}$, the language model that results via the JS optimization described in section 7.1, gives a slight improvement in each of the ROUGE scores, but there is no statistically significant difference between the ROUGE scores for $\mathrm{L}M_4^{opt}$ and $\mathrm{L}M_4^{equal}$.

System	R1	R2	R3	R4
$\mathrm{L}M_1$	0.511	0.169	0.067	0.031
$\mathrm{L}M_4^{equal}$	0.559	0.210	0.095	0.052
$\mathrm{L}M_4^{opt}$	0.562	0.216	0.100	0.055
Avg Human	0.528	0.179	0.075	0.036
Best Human	0.572	0.219	0.110	0.071

Table 2: ROUGE Results for Three Language Models and a Comparison to Human Performance

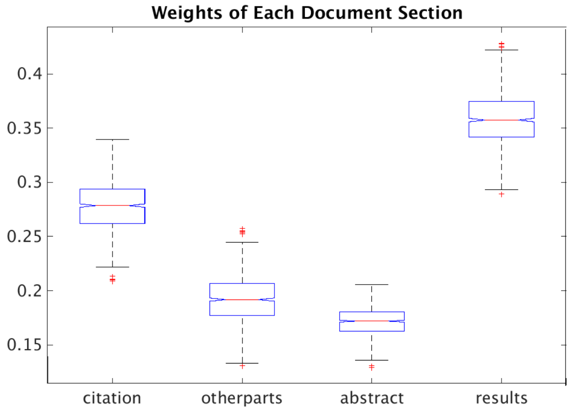

Figure 1: Language Model Coefficients for Document Sections

To measure the stability of the weighting coefficients learned by the optimization we performed 1000 trials of the optimization. In each trial a random subset of 10 of the reference papers and citances were chosen to perform the JS optimization. Figure 1 gives a notched box plot of the result of the experiment. The experiment demonstrates that the result section is given a significantly higher weight than the citations. Surprisingly the abstract is given the lowest weight. Note that the median abstract length of a doucment is about 145 words while human generated summaries are 250 words. Clearly, the human summarizers, having the freedom to write a summary longer than the median abstract length, chose to focused on the results section and the citances and did not draw mainly from the abstract.

8 Conclusions

In this paper we compared the performance of a simple vector space model and two language modeling approaches for estimating the latent weights of the terms for scientific document summarization[3] that exploit the underlying structure of the document. Our vector space model uses the TF representation of the text and a low rank approximation of the term-sentence matrix using NNMF. The TF vector space model is a good basic model, but we showed that it benefits from low rank NNMF approximation. The ROUGE scores of the summaries computed with NNMF exceeded those of the basic TF and were close to the average human ROUGE score. However given the humanly generated segmentation of a scientific paper (the sections abstract, result, other, and citances of the document) gives rise to a stronger language model that we show achieves a performance exceeding that of the average human in four ROUGE measures.

Acknowledgments

The authors would like to thank Jeff Kubina for his selection of the background corpus, PubMed, which was used in this work. In addition, we like to thank Dianne P. O'Leary and Judith D. Schlesinger for fruitful discussion and feedback on the training summaries. Finally, we thank Julie M. Conroy for her editing of this paper and for the suggestions by Jeff Kubina and Libby Beer, all of which made for a clearer and stronger paper.

References

John M. Conroy, Sashka T. Davis, Jeff Kubina, Yi-Kai Liu, Dianne P. O'Leary, and Judith D. Schlesinger. 2013. Multilingual Summarization: Dimensionality Reduction and a Step Towards Optimal Term Coverage. In *MultiLing Workshop*, pages 55–63.

Sashka T. Davis, John M. Conroy, and Judith D. Schlesinger. 2012. OCCAMS - An Optimal Combinatorial Covering Algorithm for Multi-document Summarization. In Jilles Vreeken, Charles Ling, Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu, editors, *ICDM Workshops*, pages 454–463. IEEE Computer Society.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501, Morristown, NJ, USA. Association for Computational Linguistics.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 463–470, Stroudsburg, PA, USA. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2009. Automatically evaluating content selection in summarization without human models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 306–314, Singapore, August. Association for Computational Linguistics.

The National Institute of Health. 2014. Pubmed. *http://www.ncbi.nlm.nih.gov/pubmed*.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 689–696, Stroudsburg, PA, USA. Association for Computational Linguistics.

G.A.F. Seber. 2004. *Multivariate observations*. Wiley series in probability and statistics. Wiley-Interscience.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28:2002.

[3] The models were applied to the NIST Biomedical Summarization data.