Topical Coherence for Graph-based Extractive Summarization

Daraksha Parveen[†] **Hans-Martin Ramsl**[‡] **Michael Strube**[†]

†NLP Group and Research Training Group AIPHES
Heidelberg Institute for Theoretical Studies gGmbH
Heidelberg, Germany
{daraksha.parveen|michael.strube}@h-its.org

‡SAP SE
Walldorf, Germany
hans-martin.ramsl@sap.com

Abstract

We present an approach for extractive single-document summarization. Our approach is based on a weighted graphical representation of documents obtained by topic modeling. We optimize importance, coherence and non-redundancy simultaneously using ILP. We compare ROUGE scores of our system with state-of-the-art results on scientific articles from *PLOS Medicine* and on DUC 2002 data. Human judges evaluate the coherence of summaries generated by our system in comparision to two baselines. Our approach obtains competitive performance.

1 Introduction

Summarization systems take a long document as input and generate a concise document as output. Several summarization variants exist such as generic, query-based, multi-document and single document, but the basic requirements for summarization remain the same. Summaries should contain salient information so that the reader will not miss anything from the original document. Also, the reader is not interested in repetitive information, so summaries should not include redundant information. Finally, summaries should be coherent and of high readability.

We introduce a completely unsupervised graph-based summarization using latent drichlet allocation (LDA, Blei and Lafferty (2009)). LDA is a simple model for topic modeling where topic probabilities are assigned words in documents. The probabilities can be used to measure the semantic relatedness between words and hence the topical coherence of a document. We use topical coherence as a means to ensure the coherence of extractive single-document summaries. Remus and Biemann (2013) apply LDA to compute lexical chains while Gorinski and Lapata (2015) also develop a graph-based summarization system which takes coherence into account.

Our work is based on the bipartite entity graph introduced by Guinaudeau and Strube (2013). However, in their graph one set of nodes corresponds to entities whereas in our graph it corresponds to topics. The entity graph has already been used by Parveen and Strube (2015) for summarization. Their graph is unweighted and sparse, whereas our topical graph is weighted and dense.

We apply our topical graph on the dataset introduced by Parveen and Strube (2015). This dataset contains scientific articles from the journal *PLOS Medicine*[1]. Every *PLOS Medicine* article is accompanied by an editor's summary and an authors' abstract. We use both as gold summaries for evaluation. Results obtained on the *PLOS Medicine* dataset using the topical graph are as good as using the entity graph and significantly better than several baselines and the graph-based system *TextRank* (Mihalcea and Tarau, 2004). We use the DUC 2002 dataset to compare our results with state-of-the-art techniques. In contrast to the *PLOS Medicine* data the DUC 2002 dataset contains very small articles. Still, our technique gives comparable results to the state-of-the-art. This shows that our technique is flexible and scalable despite being unsupervised.

2 Our Method

2.1 Document Representation

A graph-based representation has been used by well known summarization systems such as LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004). The graph used by both is of one mode type where sentences are nodes which are connected by weighted edges.

[1] http://journals.plos.org/plosmedicine/

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1949–1954,
Lisbon, Portugal, 17-21 September 2015.

Weights express sentence similarity.

We use a bipartite graph representation of documents (Figure 1). The bipartite graph, $G = (V_s, V_t, E_{t,s})$, has two sets of nodes where V_s represents sentences and V_t topics. The two sets of nodes are connected with edge $E_{t,s}$, if a word in a sentence s is present in a topic t. If multiple words are present in topic t of sentence s, then the edge weight is the logarithmic sum of probabilities of words in topic t. We normalize the edge weight by dividing them by the length of the sentence. Hence long sentences will not benefit from their lengths. We call the resulting graph *topical graph*.

2.2 Sentence Ranking

The final summary should contain only important sentences. Therefore, we give a score to every sentence in a document to obtain important sentences. Following Parveen and Strube (2015) we apply the HITS (Hyperlink Induced Topic Search) (Kleinberg, 1999) algorithm for ranking sentences by importance, since our graph is a bipartite graph. It puts nodes of a graph in two sets: *hub nodes* and *authority nodes*.

For the HITS algorithm the rank of nodes needs to be initialized. We initialize the topic rank $Rank_{t_i} = 1$ and the sentence rank $Rank_{s_i} = 1 + sim(s_i, title)$. The $title$ in the sentence rank is the title of the article. $sim(s_i, title)$ is the cosine similarity between the sentence s_i and the title of the article. After initialization of all nodes in the weighted topical graph, the HITS algorithm is applied to obtain ranks of sentences.

2.3 Coherence Measure

Guinaudeau and Strube (2013) represent a document by the entity graph, a bipartite graph consisting of sentence and entity nodes. They perform a one-mode projection on sentence nodes, compute the coherence of a document on the basis of the one-mode projections and use the coherence measure for summary coherence rating. Building upon this work, Parveen and Strube (2015) integrate this coherence measure to directly generate coherent summaries. Instead of the entity graph we here use the topical graph to incorporate the coherence measure. Parveen and Strube (2015) use an unweighted projection graph whereas we use a weighted projection graph of a topical graph to compute the coherence. The weighted one mode projection of the topical graph is shown in Figure 1, bottom right.

$$weighted\_coh(s_i, P) = weighted\_Outdegree(s_i, P) \tag{1}$$

$$norm\_weighted\_coh(s_i, P) = \frac{weighted\_coh(s_i, P)}{\sum weighted\_coh(s_i, P)} \tag{2}$$

Equation 1 calculates the outdegree of every sentence from the weighted projection graph. However $weighted\_coh(s_i, P)$ in Equation 1 is not a normalized value. The normalized coherence value is in Equation 2. Afterwards, we use this coherence value in the optimization phase for the selection of sentences.

2.4 Optimization

McDonald (2007) introduces summarization as an optimization task which takes care of importance, redundancy and coherence simultaneously. In this paper, we also propose a model for single document summarization which is based on integer linear programming (ILP). We consider ranks obtained by the HITS algorithm as sentence importance. The weighted coherence measure is calculated using Equation 1 and Equation 2. *PLOS Medicine* articles are very long and contain repetitive information, so we have to deal with redundancy even in single-document summarization. Therefore we model non-redundancy as topic coverage in the final summary: the more topics in a summary, the less redundant the summary will be. The ILP objective function is shown in Equation 3. $f_i(X)$ is the function which maximizes importance, $f_c(X)$ maximizes coherence, and $f_t(Y)$ maximizes topic coverage.

$$Objective\,function : \max_{X,Y}(f_i(X) + f_c(X) + f_t(Y)) \tag{3}$$

X is a variable for sentences which contains boolean variables x_i, where $0 < i < n$ is the number of sentences. Y is a variable for topics which contains boolean variables y_j, where $0 < j < m$ is the number of topics.

$$f_t(Y) = \sum_{j=1}^{m} y_j \tag{4}$$

Constraints ensure that the system satisfies additional requirements such as summary length:

$$\sum_{i=1}^{n} x_i \leq Len(summary) \tag{5}$$

$$\sum_{j \in T_i} y_j \geq |Topics_{x_i}| \cdot x_i, \quad \text{for } i = 1, \ldots, n \tag{6}$$

S_1 WHO recommends prompt diagnosis and quinine plus clindamycin for treatment of uncomplicated malaria in the first trimester and artemisinin-based combination therapies in subsequent trimesters.

S_2 We undertook a systematic review of women's access to and healthcare provider adherence to WHO case management policy for malaria in pregnant women.

S_3 Data were appraised for quality and content.

S_4 Determinants of women's access and providers' case management practices were extracted and compared across studies.

	Topic1 (t1)	Topic2 (t2)	Topic3 (t3)	Topic4 (t4)	Topic5 (t5)	Topic6 (t6)	Topic7 (t7)	Topic8 (t8)	Topic9 (t9)	Topic10 (t10)	Topic11 (t11)	.....	Topic500 (t500)
S_1	-	0.3	0.46	0.50	0.23	0.55	0.89	0.56	0.09	-	-	.....	0.73
S_2	0.3	0.29	0.3	-	0.66	0.35	0.36	0.63	-	-	0.55	.....	-
S_3	-	0.28	-	0.43	-	0.39	-	0.42	0.65	-	0.12	.....	0.58
S_4	0.22	0.43	0.29	0.6	-	-	0.14	0.26	-	-	-	.....	0.41

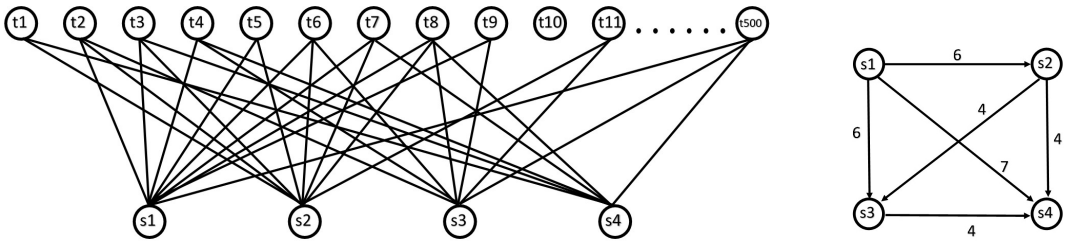

Figure 1: Abstract from *PLOS Medicine*, topical grid, bipartite topical graph, one-mode projection

$$\sum_{i \in S_j} x_i \geq y_j, \quad \text{for } j = 1, \ldots, m \tag{7}$$

The final summary should be shorter than the original text and it should also have a length limit (Equation 5). The results on *PLOS medicine* data (Section 3) are limited to 5 sentences. We have also experimented with multiple lengths. Increasing the summary length increases ROUGE scores. DUC 2002 summaries are limited to 100 words.

Equation 6 shows that topics present in sentence x_i are selected, when sentence x_i is selected. Therefore, $x_i = 1$ and $T_i = Topics_{x_i}$. The constraint holds, because $\sum_{j \in T_i} y_j = |Topics_{x_i}|$. Furthermore, if sentence $x_i = 0$, i.e., it is not selected, then there must be topics which are already present in selected sentences. Hence, the constraint holds, $\sum_{j \in T_i} y_j \geq 0$.

Equation 7 constrains the selection of topics. If topic $y_j = 1$, then at least one sentence containing this topic has been selected. Therefore $\sum_{i \in S_j} x_i \geq 1$, and the constraint holds. If topic $y_j = 0$, then sentences containing this topic are not selected, so $\sum_{i \in S_j} x_i = 0$, and the constraint holds.

3 Experiments

Following Parveen and Strube (2015), we evaluate on the science genre, i.e. *PLOS Medicine* articles, and on the news genre, i.e. DUC 2002 data.

3.1 Datasets

PLOS Medicine articles are considerably longer than DUC 2002 documents. The average number of sentences per document is 154 in *PLOS Medicine* and 25 in DUC 2002. Benefits of using *PLOS Medicine* articles for experiments are:

- They are accompanied by an authors' abstract.
- They have a summary written by an editor.
- They are formatted in XML.
- They contain explicit full forms of abbreviations.

Editor's summaries have a different perspective, writing style and length than authors' abstracts. We use both as gold summaries for evaluation. Following Parveen and Strube (2015) we report the results using editor's summaries and author's abstracts independently. To compare with the state-of-the-art in single-document summarization, we also evaluate on DUC 2002 data.

3.2 Experimental Setup

We use the XML version of *PLOS Medicine* articles. We extract the contents excluding figures, tables and references. Editor's summary and authors' abstract are separated from the content for evaluation. The *PLOS Medicine* XML provides explicit full forms when abbreviations are introduced. We replace abbreviations with their full form in the summary. We then remove non-alphabetical characters. After this we parse articles using the Stanford parser (Klein and Manning, 2003). We perform pronoun resolution using the coreference resolution system by Martschat (2013)[2]. We use *gensim* to generate the topics. For generating topics we use a dataset containing scientific articles from biology, which contains 221,385 documents and about 50 million sentences[3]. We also use Wikipedia to compare with topics from a general domain.

The HITS algorithm is applied on the bipartite graph for computing sentence importance. We calculate the coherence values of sentences on weighted one-mode projection graphs. The importance and coherence of a sentence is used in the optimization phase[4] which returns a binary value for each sentence.

3.3 Results

Results on *PLOS Medicine* are shown in Tables 1 and 2. We evaluate using ROUGE-SU4 and ROUGE-2 (Lin, 2004). We limit the length of the summaries to five sentences and the number of topics to 2000 in the topical graph. We also experimented with varying numbers of topics, i.e. 500, 1000 and 2000, and varying summary length limits. The results changed only marginally. The general trends remained the same.

We compare our system with four different baselines and two versions of the entity graph. *Lead* selects the top five sentences, *Random* five sentences randomly. *MMR* is an implementation of maximal marginal relevance (Carbonell and Goldstein, 1998). *TextRank* is the graph-based system by Mihalcea and Tarau (2004)[5]. *Egraph* is the entity graph based system by Parveen and Strube (2015). *Egraph + Coh.* is their system which includes a coherence measure, which is calculated by using the unweighted projection graph. *Egraph + Coh. + Pos.* combines the coherence measure and positional information.

Systems	R-SU4	R-2
Lead	0.067	0.055
Random	0.048	0.031
MMR	0.069	0.048
TextRank	0.068	0.048
Egraph	0.121	0.090
Egraph + Coh.	0.130	0.096
Egraph + Coh. + Pos.	0.131	0.098
Tgraph (n=2000)	0.123	0.091
Tgraph (n=2000) + Coh.	0.129	0.095
Tgraph (n=2000) + Coh. + Pos.	0.125	0.092

Table 1: *PLOS Medicine*, editor's summaries

Systems	R-SU4	R-2
Lead	0.105	0.077
Random	0.093	0.589
MMR	0.118	0.098
TextRank	0.134	0.101
Egraph	0.200	0.170
Egraph + Coh.	0.219	0.175
Egraph + Coh. + Pos.	0.224	0.189
Tgraph (n=2000)	0.217	0.173
Tgraph (n=2000) + Coh.	0.221	0.179
Tgraph (n=2000) + Coh. + Pos.	0.215	0.174

Table 2: *PLOS Medicine*, authors' abstracts

Our system outperforms all baselines substantially, as shown in Tables 1 (editor's summaries) and 2 (authors' abstracts). We observe improvements in the results when including coherence in the topical graph. We obtain best results with Tgraph + Coh., where the number of topics is 2000. In Tgraph, penalizing coherence measures with positional information lowers ROUGE scores. While including positional information into the entity graph obtains the best results on the *PLOS Medicine* dataset, positional information does not appear to be beneficial for the topical graph. Absolute ROUGE scores are higher when using abstracts as gold summaries, because the abstracts are shorter than editor's summaries.

We compare results using biology journals (Table 3) and Wikipedia (Table 4) to generate topics. The topical graph is denser when using biology journals compared to the graph generated from Wikipedia. Results using the in-domain biology journals as data to generate topics are better than using general domain Wikipedia data. The scores are highest with 2000 topics. For Bio topic the differences are negligible, however.

We also compare results on DUC 2002 to

[2] http://www.smartschat.de/software/

[3] http://www.datawrangling.com/some-datasets-available-on-the-web/

[4] We use Gurobi, http://www.gurobi.com

[5] https://kenai.com/projects/textsummarizer

Topics	R-1	R-2	R-SU4
Tgraph (n=500) + Coh.	0.279	0.090	0.125
Tgraph (n=1000) + Coh.	0.289	0.093	0.128
Tgraph (n=2000) + Coh.	0.291	0.095	0.129

Table 3: *PLOS Medicine*, editor's summ., Bio topic

Topics	R-1	R-2	R-SU4
Tgraph (n=500) + Coh.	0.208	0.060	0.098
Tgraph (n=1000) + Coh.	0.258	0.073	0.106
Tgraph (n=2000) + Coh.	0.283	0.086	0.121

Table 4: *PLOS Medicine*, editor's summ., Wiki topic

check against the state-of-the-art on a well-known dataset. *Lead* performs well on DUC 2002 as shown in Table 5, because important information appears in the initial lines of news articles. *DUC 2002 Best* is the result reported by the top performing system at DUC 2002. This system actually obtains better results than *TextRank* (Mihalcea and Tarau, 2004) and the more recent system *UniformLink* (Wan and Xiao, 2010). Our system *Tgraph + Coh.* performs better than the well known best systems on DUC 2002 and slightly better than *Egraph + Coh.* However the difference between the results of Tgraph and Egraph are not significant. In contrast to the entity graph based system, the coherence measure in our system is calculated by using a topic-based weighted projection graph, which is denser and hence more informative.

3.4 Human Coherence Judgements

In addition to ROUGE scores, we use human judgements for evaluating the coherence of our summaries. We asked four PhD students in natural language processing to evaluate our summaries on the basis of coherence. We randomly selected ten summaries of scientific articles from three different systems, *TextRank*, *Lead* and *Tgraph + Coh.* We asked the human judges to rank the summaries according to their coherence. So, the summary which is best in coherence gets rank 1, second best gets rank 2, and worst gets rank 3. We calculated the Kendall concordance coefficient (W) (Siegel and Castellan, 1988) to measure the judges' agreement. We obtain $W = 0.61$, which indicates a relatively high agreement.

Systems	R-1	R-2	R-SU4
Lead	0.459	0.180	0.201
DUC 2002 Best	0.480	0.228	
TextRank	0.470	0.195	0.217
UniformLink (k = 10)	0.471	0.201	
Egraph + Coh.	0.479	0.238	0.230
Tgraph (n=2000) + Coh.	0.481	0.243	0.242

Table 5: DUC 2002, single-document summarization

To compare the three systems, we take the average over the ranks. The overall rank of *TextRank* is 2.625, *Lead* is 1.675 and *Tgraph + Coh.* is 1.8. *Lead* performs best, because it selects the top five consecutive sentences, which are coherent as the original authors intended them to be. However, the overall ranks of *Lead* and *Tgraph + Coh.* are not significantly different, whereas *TextRank*'s overall rank is significantly worse than both. Hence, *Tgraph + Coh.* performs very well in our human judgement coherence experiment.

4 Discussion

In this paper we introduced the topical graph for single document summarization. We experimented with multiple numbers of topics on the scientific article dataset. Our system performs well when including the weighted coherence measure in the optimization phase. The results are comparable with the entity graph. However, the entity graph is less informative and very sparse as compared to the topical graph. Our system does not need annotated training data and, except for the number of topics, no optimization of parameters. Hence, we consider it unsupervised.

Acknowledgments

This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a Heidelberg Institute for Theoretical Studies Ph.D. scholarship. This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1. We would like to thank our colleagues Sebastian Martschat, Nafise Moosavi, Alex Judea and Mohsen Mesgar who served as human subjects and commented on earlier drafts.

References

David M. Blei and John D. Lafferty. 2009. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall, Boca Raton, Flo.

Jaime G. 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,* Melbourne, Australia, 24–28 August 1998, pages 335–336.

Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Denver, Col., 31 May – 5 June 2015, pages 1066–1076.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Sofia, Bulgaria, 4–9 August 2013, pages 93–103.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July 2003, pages 423–430.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out Workshop at ACL '04,* Barcelona, Spain, 25–26 July 2004, pages 74–81.

Sebastian Martschat. 2013. Multigraph clustering for unsupervised coreference resolution. In *51st Annual Meeting of the Association for Computational Linguistics: Proceedings of the Student Research Workshop,* Sofia, Bulgaria, 5–7 August 2013, pages 81–88.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Proceedings of the European Conference on Information Retrieval,* Rome, Italy, 2-5 April 2007.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* Barcelona, Spain, 25–26 July 2004, pages 404–411.

Daraksha Parveen and Michael Strube. 2015. Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence,* Buenos Aires, Argentina, 25–31 July 2015, pages 1298–1304.

Steffen Remus and Chris Biemann. 2013. Three knowledge-free methods for automatic lexical chain extraction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Atlanta, Georgia, 9–14 June 2013, pages 989–999.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, 2nd edition.

Xiaojun Wan and Jianguo Xiao. 2010. Exploiting neighborhood knowledge for single document summarization and keyphrase extraction. *ACM Transactions on Information Systems*, 28(2):8 pages.