Integrating Importance, Non-Redundancy and Coherence in Graph-Based Extractive Summarization

Daraksha Parveen and Michael Strube
NLP Group and Research Training Group AIPHES
Heidelberg Institute for Theoretical Studies gGmbH
Schloß-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
{daraksha.parveen—michael.strube}@h-its.org

Abstract

We propose a graph-based method for extractive single-document summarization which considers importance, non-redundancy and local coherence simultaneously. We represent input documents by means of a bipartite graph consisting of sentence and entity nodes. We rank sentences on the basis of importance by applying a graph-based ranking algorithm to this graph and ensure non-redundancy and local coherence of the summary by means of an optimization step. Our graph based method is applied to scientific articles from the journal PLOS Medicine. We use human judgements to evaluate the coherence of our summaries. We compare ROUGE scores and human judgements for coherence of different systems on scientific articles. Our method performs considerably better than other systems on this data. Also, our graph-based summarization technique achieves state-of-the-art results on DUC 2002 data. Incorporating our local coherence measure always achieves the best results.

1 Introduction

Summaries should contain the most important information from input documents. Summaries should not contain redundant information. Finally, summaries should be readable, hence they should be grammatical and coherent. Many summarization approaches focus on extracting important sentences from input documents while ensuring that the extracted information is non-redundant (e.g., methods based on maximum marginal relevance [Carbonell and Goldstein, 1998]). Grammaticality does not concern extractive summarization as complete sentences, which are assumed to be grammatical, are extracted from input documents. However, there has been surprisingly little research on including discourse processing techniques into extractive summarization. When relating discourse processing techniques to automatic summarization, [Nenkova and McKeown, 2011] mention only work based on discourse relations [Marcu, 2000; Louis et al., 2010] which has been used for selecting important infomation, but which has not been applied to improve the coherence of the summaries. [Clarke and Lapata, 2007] use discourse constraints for sentence compression.

[Barzilay and Lapata, 2008] apply their local coherence model, the entity grid, to summary coherence evaluation. However, to our knowledge, the entity grid has not been used directly in extractive summarization to ensure summary coherence. Our work is based on the graph-based extractive summarization technique developed by [Parveen and Strube, 2014]. It is intuitively plausible to extend this technique by the entity graph [Guinaudeau and Strube, 2013]. Therefore, in our work, computing importance, non-redundancy and coherence is tightly integrated and taken care of simultaneously. Our graph-based summarization technique has the further advantage of being completely unsupervised.

We apply our graph-based summarization technique to scientific articles from the journal PLOS Medicine, a high-impact open-access journal from the medical domain. Articles in this journal are not only accompanied by authors’ abstracts but also by a summary written by an editor. We propose to use editors’ summaries as gold-standard for our evaluation. PLOS Medicine has the further advantage of being available in XML-format, which saves us the expense of noisily extracting information from PDF documents and the ambiguities of HTML parsing. PLOS Medicine is distributed by means of a Creative Commons Attribution License allowing us to publish the dataset. On the PLOS Medicine data our graph-based approach outperforms several baselines and the graph-based method TextRank [Mihalcea and Tarau, 2004]. We also show that our technique works significantly better than a version without the coherence constraint. We do not only report ROUGE scores, but also evaluate the coherence by having our summaries judged by human subjects. Finally, we also apply our technique to the DUC 2002 data, where the input documents are shorter by an order of magnitude compared to the scientific articles. On these data, our technique reaches a performance comparable to the state-of-the-art despite being unsupervised.

In Section 2, we discuss related work in the field of summarization. Section 3 provides a detailed description of our method. The datasets used, the experimental setup and the results are described in Section 4 and discussed in Section 5.
2 Related Work

We focus on extractive summarization of scientific articles. Extractive summarization involves computing the importance of sentences which is used for deciding whether to include a sentence in the summary. The importance of sentences may depend on different factors: term frequency, position in text, cue words, and lexical chains among other factors. The importance of sentences may also depend on the document representation. In a graph-based document representation [Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Parveen and Strube, 2014], nodes, i.e. sentences, are ranked by means of ranking algorithms. The importance of sentences can also be determined by supervised classification [Amini and Gallinari, 2002], unsupervised clustering [Nomoto and Parveen and Strube, 2014], nodes, i.e. sentences, are ranked by means of ranking algorithms. The importance of sentences may also depend on the document representation. In a graph-based document representation [Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Parveen and Strube, 2014], nodes, i.e. sentences, are ranked by means of ranking algorithms. The importance of sentences can also be determined by supervised classification [Amini and Gallinari, 2002], unsupervised clustering [Nomoto and Matsumoto, 2001], HMM and CRF based methods [Conroy and O’Leary, 2001; Shen et al., 2007], and support vector regression based methods [Galanis et al., 2012]. [Carbonell and Goldstein, 1998] use a greedy method to select sentences for the final summary. The system MEAD [Radev et al., 2004] uses a clustering method to find a centroid. It considers the centroid as words which are highly similar to the topic. Ranking using centroids and redundancy are taken care of simultaneously. Similarly, the CLASSY system [Conroy et al., 2004], the best performing system at DUC 2004, uses the tf-idf score to calculate the importance of sentences. However, these systems do not consider the global frame of reference. Therefore, [McDonald, 2007] has proposed an optimization method based on integer linear programming (ILP) which considers text summarization as a knapsack problem. Similarly, [Berg-Kirkpatrick et al., 2011], [Galanis et al., 2012] also used ILP optimization methods for their tasks. [Wan and Xiao, 2010] assume that nearest neighbour documents provide additional knowledge and improve single-document summarization and keyphrase extraction. [Hirao et al., 2013] consider coherence by using the rhetorical structure for single-document summarization.

For summarizing scientific articles, [Teufel and Moens, 2002] present an algorithm based on rhetorical status. More recently, citation-based summarization received a great deal of attention [Teufel et al., 2006; Qazvinian and Radev, 2008]. [Abu-Jbara and Radev, 2011] provide a method based on sentence clustering and ranking, which also takes care of the coherency in the citation based summary.

3 Our Method

The control flow of our method is shown in Figure 1.

3.1 Document Representation

[Erkan and Radev, 2004] and [Mihalcea and Tarau, 2004] employ graphs where nodes are sentences which are connected by weighted edges. The weights represent sentence similarity. While these graphs are of one mode type, we here build on [Parveen and Strube, 2014] who represent documents by a graph $G = (V_s, V_c, E_{c,s})$. $G$ has two different sets of nodes $V_s$ and $V_c$. There are no edges connecting nodes within the same set. Edges $E_{c,s}$ connect only nodes from different sets. Here, $e$ represents an entity, and $s$ represents a sentence in the document. This bipartite graph, the entity graph, has been introduced by [Guinaudeau and Strube, 2013] as an alternative, graph-based representation of the entity grid [Barzilay and Lapata, 2008]. While [Guinaudeau and Strube, 2013] employ the entity graph for computing local coherence of documents, we use it here additionally for determining the importance of sentences. In Figure 2 we show an abstract from a PLOS Medicine article, its representation in an entity grid, and its transformation into a bipartite entity graph.

3.2 Sentence Ranking

We use the HITS [Kleinberg, 1999, Hyperlink Induced Topic Search] algorithm to rank nodes in bipartite graphs. In HITS, webpages are divided into a set of hub pages and set of authority pages. We represent each document as a bipartite graph. Hence, we follow [Parveen and Strube, 2014] in using the HITS algorithm for sentence ranking.

The HITS algorithm requires to associate nodes in the graph with an initial rank. [Parveen and Strube, 2014] apply the HITS algorithm to rank sentences in topic-based multidocument summarization. For initialization they rank entities based on their presence in the topic. In our work, we instead provide initial ranks for sentences and entities. Because we do not have topics (neither the DUC 2002 data nor the PLOS Medicine articles have topics) we use the document title. Initial ranks for sentences are based on the similarity between sentences and the title. Initial ranks for entities are based on their presence in the title. So the initial rank of entity $e_i$ is $Ranks_{e_i} = 1 + \text{tf}(e_i, \text{article}) + \text{occurrence}(e_i, \text{title})$. Here, $\text{tf}(e_i, \text{article})$ is the term frequency of $e_i$ in the document. $	ext{occurrence}(e_i, \text{title})$ shows the occurrence of $e_i$ in the document title. Hence, if $e_i$ is not present in the title then $\text{occurrence}(e_i, \text{title}) = 0$. If it is present then $\text{occurrence}(e_i, \text{title}) = 1$. We also initialize the sentence rank by considering its similarity to the document title. So, $Ranks_s = 1 + \text{sim}(s_i, \text{title})$ is an initialization of sentences. $\text{sim}(s_i, \text{title})$ is the cosine similarity between sentence $s_i$ and the document title. After initialization, we apply the HITS algorithm on the bipartite graph.

3.3 Coherence Measure

[Barzilay and Lapata, 2008] introduce the entity grid which operationalizes the linguistic intuition that entities shared by
Haemorrhage is a common cause of death in trauma patients.

Although transfusions are extensively used in the care of bleeding trauma patients, there is uncertainty about the balance of risks and benefits and how this balance depends on the baseline risk of death.

Our objective was to evaluate the association of red blood cell (RBC) transfusion with mortality according to the predicted risk of death.

A secondary analysis of the CRASH-2 trial (which originally evaluated the effect of tranexamic acid on mortality in trauma patients) was conducted.

Figure 2: Abstract from *PLOS Medicine*, entity grid, bipartite entity graph

Figure 3: One mode projection of the bipartite graph

Until now entity grid and entity graph have been used only to evaluate summaries according to their coherence. Since we represent the input documents also by means of the bipartite entity graph, we can incorporate the coherence measure by performing the one mode projection directly in our summarization system. This provides additional guidance in selecting sentences for the summary.

\[
\text{coherence}(s_i, P) = \frac{\text{Outdegree}(s_i, P)}{\text{position}(s_i)}
\]  

Equation 2 calculates the outdegree of every sentence from the projection graph. Then in Equation 3 the coherence value of a sentence is penalized according to its position in a document. If we want to use only the coherence then \(\text{position}(s_i) = 1\) in Equation 3, and if we want to use the position only then \(\text{coherence}(s_i, P) = 1\) in Equation 3.

We use this coherence value while selecting sentences for a summary in the optimization phase. Only sentences maximizing the coherence value will be selected.
3.4 Optimization

[McDonald, 2007] views summarization as optimization task, since a summarization system has to consider the tradeoff between importance and redundancy. Our system selects information by importance. Because the scientific articles we deal with are very long and redundant information is spread out over several sections, we have to deal with redundancy. In addition, we take into account the coherence value of a sentence. Hence, we optimize importance, non-redundancy and coherence simultaneously. The importance of a sentence is a rank calculated by applying the HITS algorithm as discussed in Section 3.2. The coherence value of sentences is calculated using their outdegree in a projection graph as explained in Section 3.3. For computing non-redundancy we consider entities in a document. If more new entities are included in the final summary, the least redundant the final summary will be [Parveen and Strube, 2014].

The objective function shown in Equation 5 contains three measures: importance, coherence value and redundancy.

Variables:

\[ x_i \quad \text{and} \quad y_j \]

Objective function:

\[ f(X,Y) = \max \left( \sum_{i=1}^{n} \text{Rank}(s_i) \cdot x_i + \sum_{i=1}^{n} f(s_i) \cdot x_i + \sum_{j=1}^{m} y_j \right) \]  

Constraints:

\[ \sum_{i=1}^{n} x_i \leq \text{Len(summary)} \]  

\[ \sum_{j \in E_i} y_j \geq \text{Entities}_{x_i}, \quad \text{for} \quad i = 1, \ldots, n \]  

\[ \sum_{i \in S_j} x_i \geq y_j, \quad \text{for} \quad j = 1, \ldots, m \]

The last constraint in Equation 8 provides us with information about the entities. If entity \( y_j = 1 \) then at least one sentence containing this entity must be selected. \( \sum_{i \in S_j} x_i \geq 1 \), hence the constraint holds. If entity \( y_j \) is not selected then none of the sentences containing that entity may also not be selected, \( y_j = 0 \) and \( \sum_{i \in S_j} x_i = 0 \).

4 Experiments

Experiments are performed on the PLOS Medicine dataset and the DUC 2002 single-document summarization dataset.

4.1 Datasets

We introduce a new dataset for summarizing scientific articles. It consists of 50 articles from the high impact open access journal PLOS Medicine. Articles in this journal are much longer than documents in most standard datasets used in research on automatic summarization (see Table 1). Articles in PLOS Medicine can be as long as the whole set of documents related to one query in multi-document summarization.

PLOS Medicine publishes 10-15 articles per month since 2004. We chose PLOS Medicine as source for our experiments, because articles are accompanied not only by an abstract written by the authors but also by a summary written by an editor of the journal. We report experiments using both editors’ summaries and authors’ abstracts as gold standard for evaluation. The editor’s summary consists usually of 15 to 20 sentences whereas the abstract consist of 10 to 15 sentences. Hence both editor’s summary and authors’ abstract are relatively long compared to summaries usually used in research on automatic summarization. When applying our summarization system we remove editor’s summary, abstract, figures, tables and references from the paper.

As a second point of reference we also perform experiments on the single-document summarization data from DUC 2002 which contains rather short news articles.

4.2 Experimental Setup

We use the XML formatted scientific articles from PLOS Medicine. We extract the content of a paper excluding figures, table and references. After this, 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 this full form in the final 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 apply the Brown coherence

\( ^2 \)http://www.smartschat.de/software/

Table 1: PLOS Medicine and DUC 2002 datasets

<table>
<thead>
<tr>
<th></th>
<th>PLOS</th>
<th>DUC 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. of docs.</td>
<td>50</td>
<td>567</td>
</tr>
<tr>
<td>avg. no. of sents. per doc.</td>
<td>154</td>
<td>25</td>
</tr>
<tr>
<td>avg. no. of words per doc.</td>
<td>4756</td>
<td>627</td>
</tr>
</tbody>
</table>
toolkit [Elsner and Charniak, 2011]3 to the articles to convert
document into an entity grid which then is transformed
into the bipartite entity graph [Guinaudeau and Strube, 2013].
Entities in the bipartite graph are the head nouns of noun
phrases. Afterwards, the HITS algorithm is applied on the
bipartite graph for computing the importance of sentences.
We calculate the coherence values of sentences using the one
mode projection. The importance and coherence of a sen-
tence is used in the optimization phase4. The optimization
phase returns a binary value associated with each sentence.
The sentence is included in the summary if its value is 1.

4.3 Human Coherence Judgements
Automatic summarization has frequently been evaluated by
computing ROUGE scores [Lin, 2004]. Beyond n-grams
ROUGE does not have a notion of order, and therefore does
not account for coherence. In order to complement ROUGE
scores which we report in Section 4.4, we perform an evalu-
ation with human judgements for coherence. We asked five
PhD. students in Natural Language Processing (the authors
were not among them) to comparatively rank the output of
our system on the basis of coherence. We randomly selected
ten scientific articles from PLOS medicine. We used three
different systems to generate summaries: the Lead baseline
(S1), Our System + Coh. + Pos. (S2) and TextRank (S3).
Our human judges were asked to assign rank 1 to the best
summary, rank 2 to the second best, rank 3 to the worst. By
computing the average over the ranks given by all five judges
we compute an overall rank: S1 gets an overall rank of 1.34,
2 gets 1.82, and S3 gets 2.84.

Unsurprisingly S1 performed best among the three sys-
tems. S1 is the Lead baseline which consists of the first five
sentences from the article. Since these five sentences are ex-
tracted en bloc, they are as coherent as the original authors
intended them to be. Still, the difference in average rank be-
tween S1 and S2 is not very substantial. In three of our ten
documents S2 was ranked higher than S1 on average. The
difference between S2 and S3 however is substantial.

We apply the Kendall concordance coefficient (W) [Siegel
and Castellan, 1988] to measure whether our human subjects
agree in ranking the three systems. With $W = 0.64$ the corre-
lation between the human subjects is relatively high. Apply-
ing the $\chi^2$ test shows that W is significant at the 95% level.
Hence, we interpret the rankings provided by our human sub-
jects reliable and informative.

4.4 Results
Results on PLOS Medicine data are shown in Tables 2 and
3. Evaluation metrics are ROUGE-SU4 and ROUGE-2 [Lin,
2004]. ROUGE-SU4 calculates the co-occurrence of skip
bigrams between system summary and human summary.
ROUGE-2 calculates bigram co-occurrence. We limit the
summaries to five sentences. We compare our system with
four different baselines. Lead selects the top five sentences,
Random selects five sentences randomly from the scientific

<table>
<thead>
<tr>
<th>Systems</th>
<th>R-SU4</th>
<th>R-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead</td>
<td>0.067</td>
<td>0.055</td>
</tr>
<tr>
<td>Random</td>
<td>0.048</td>
<td>0.031</td>
</tr>
<tr>
<td>MMR</td>
<td>0.069</td>
<td>0.048</td>
</tr>
<tr>
<td>TextRank</td>
<td>0.068</td>
<td>0.048</td>
</tr>
<tr>
<td>Our System</td>
<td>0.121</td>
<td>0.090</td>
</tr>
<tr>
<td>Our System + Coh.</td>
<td>0.130</td>
<td>0.096</td>
</tr>
<tr>
<td>Our System + Pos.</td>
<td>0.129</td>
<td>0.093</td>
</tr>
<tr>
<td>Our System + Coh. + Pos.</td>
<td>0.131</td>
<td>0.098</td>
</tr>
</tbody>
</table>

Table 2: Results on PLOS Medicine, editors’ summaries

<table>
<thead>
<tr>
<th>Systems</th>
<th>R-SU4</th>
<th>R-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead</td>
<td>0.105</td>
<td>0.077</td>
</tr>
<tr>
<td>Random</td>
<td>0.093</td>
<td>0.589</td>
</tr>
<tr>
<td>MMR</td>
<td>0.118</td>
<td>0.098</td>
</tr>
<tr>
<td>TextRank</td>
<td>0.134</td>
<td>0.101</td>
</tr>
<tr>
<td>Our System</td>
<td>0.200</td>
<td>0.170</td>
</tr>
<tr>
<td>Our System + Coh.</td>
<td>0.219</td>
<td>0.175</td>
</tr>
<tr>
<td>Our System + Pos.</td>
<td>0.218</td>
<td>0.174</td>
</tr>
<tr>
<td>Our System + Coh. + Pos.</td>
<td>0.224</td>
<td>0.189</td>
</tr>
</tbody>
</table>

Table 3: Results on PLOS Medicine, authors’ abstracts

article. MMR is an implementation of maximal marginal rele-
vance [Carbonell and Goldstein, 1998]. TextRank is the graph

Our system outperforms all baselines substantially as shown
in Tables 2 and 3. In the experiments reported in Table
2 we use the editor’s summary as a gold summary, in Table
3 the authors’ abstracts. We observe improvements when
including coherence and taking position of the extracted sen-
tences into account. We obtain best results when combining
our system with coherence and position. The improvements
of our system are consistent across the editors’ summaries
vs. the authors’ abstracts conditions. In the latter case the
absolute numbers returned by ROUGE are higher, because the
abstracts are shorter than the editors’ summaries.

For editors’ summaries the difference between Our System
and Our System + Coh. + Pos. is statistically significant
at the 0.01-level for ROUGE-SU4 and at the 0.05-level for
ROUGE-2. For abstracts the difference between Our System
and Our System + Coh. + Pos. is statistically significant at
the 0.01-level for both metrics. The difference between TextRank
and Our System is statistically significant at the 0.01-level for
both editors’ summaries and abstracts for both metrics. Dif-
fences between Our System + Coh. + Pos. and Our System
+ Coh. or Our System + Pos. are not statistically significant.

We also compare results on DUC 2002 to check against
the state-of-the-art on a well-known dataset (Table 4, some
cells are empty, because ROUGE-SU4 was not reported by
all systems). Lead selects the first few sentences from the
document with a length of up to a 100 words. The Lead
baseline performs very well on the DUC 2002 data which are
composed of news articles, because it exploits characteristics
of the news genre. DUC 2002 Best is the result reported by

3http://www.ling.ohio-state.edu/~melsner/#software
4We use Gurobi, http://www.gurobi.com
5https://kenai.com/projects/textsummarizer/sources/mercurial-
textsummarizer
the top performing system at DUC 2002. This system actually obtained better results than TextRank [Mihalcea and Tarau, 2004] and the more recent system UniformLink [Wan and Xiao, 2010]. Our System + Coh. + Pos. is the only system which outperforms DUC 2002 Best.

5 Discussion

Our system including coherence works well across different domains, genres, and compression rates. It does not depend on any parameters and training data. Hence it is fully unsupervised. We compared results of different system versions on PLOS Medicine data. We also did experiments with more than 5 sentences, e.g. with 10, 15 and 20. ROUGE scores increase with summary length. The system with coherence works always better than the system without and performs best when combined with positional information.

6 Conclusions and Future Work

While previous work uses a measure of local coherence only to evaluate the coherence of summaries, we integrate it closely with determining importance and avoiding redundancy in an unsupervised graph-based method for extractive single-document summarization. We evaluate our method on long scientific articles from the journal PLOS Medicine and short news articles from the DUC 2002 single-document summarization data. On both datasets our system obtains good results. Including coherence always improves results. We also used human subjects to judge our system’s readability among three given systems. These judgements show that our system takes care of coherence. We publish the PLOS Medicine dataset, which conveniently contains editors’ summaries which we propose to use as gold summaries for evaluation. In future work we would like to include more linguistic information into the entity graph. We plan to obtain judgements by domain experts to see whether the editors’ summaries in fact can be used as gold summaries.

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 Benjamin Heinzerling, Yufang Hou, Sebastian Martschat, Alex Judea, Nafise Moosavi and Mohsen Mesgar who became human subjects and commented on earlier drafts.

References


[Erkan and Radev, 2004] G¨unes Erkan and Dragomir R. Radev. LexRank: Graph-based lexical centrality as

### Table 4: DUC 2002, single-document summarization

<table>
<thead>
<tr>
<th>Systems</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead</td>
<td>0.459</td>
<td>0.180</td>
<td>0.201</td>
</tr>
<tr>
<td>DUC 2002 Best</td>
<td>0.480</td>
<td>0.228</td>
<td></td>
</tr>
<tr>
<td>TextRank</td>
<td>0.470</td>
<td>0.195</td>
<td>0.217</td>
</tr>
<tr>
<td>UniformLink (k = 10)</td>
<td>0.471</td>
<td>0.201</td>
<td></td>
</tr>
<tr>
<td>Our System + Coh. + Pos.</td>
<td>0.485</td>
<td>0.230</td>
<td>0.253</td>
</tr>
</tbody>
</table>


