Generating Coherent Summaries of Scientific Articles Using Coherence Patterns

Daraksha Parveen **Mohsen Mesgar** **Michael Strube**

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

Abstract

Previous work on automatic summarization does not thoroughly consider coherence while generating the summary. We introduce a graph-based approach to summarize scientific articles. We employ coherence patterns to ensure that the generated summaries are coherent. The novelty of our model is twofold: we mine coherence patterns in a corpus of abstracts, and we propose a method to combine coherence, importance and non-redundancy to generate the summary. We optimize these factors simultaneously using Mixed Integer Programming. Our approach significantly outperforms baseline and state-of-the-art systems in terms of coherence (summary coherence assessment) and relevance (ROUGE scores).

1 Introduction

The growth in the scientific output of many different fields makes the task of automatic summarization imperative. Automatic summarizers assist researchers to have an informative and coherent gist of long scientific articles. An automatic summarizer produces summaries considering three properties:
Importance: The summary should contain the important information of the input document.
Non-redundancy: The summary should contain non-redundant information. The information should be diverse in the summary.
Coherence: Though the summary should comprise diverse and important information of the input document, its sentences should be connected to one another such that it becomes coherent and easy to read.

If we do not ensure that a summary is coherent, its sentences may not be properly connected. This results in an obscure summary. In previous work coherence has not been thoroughly considered. Parveen and Strube (2015) use single sentence connectivity in the input document as a coherence measure. They measure coherence by calculating the outdegree of a sentence in a graph representation of an input document. This has two disadvantages: first, since it is computed only based on one sentence, it is not sufficient to generate coherent summaries; second, it is obtained based on sentence connectivity in the input document rather than in the summary.

In this work, we focus on the coherence aspect of summarization. We use discourse entities as the unit of information that relate sentences. Here, discourse entities are referred to as head nouns of noun phrases (see Section 2). The main goal is to extract sentences which refer to those entities which are important and unique, and also to entities which connect the extracted sentences in a coherent manner. Entities in connected sentences can be used to create linguistically motivated coherence patterns (Daneš, 1974). Recently, Mesgar and Strube (2015) modeled these coherence patterns by subgraphs of the graph representation (nodes represent sentences and edges represent entity connections among sentences) of documents. They show that the frequency of coherence patterns can be used as features for coherence.

The key idea of this paper is to apply coherence patterns to long scientific articles to extract (possibly) non-adjacent sentences which, however, are already coherent. Based on the assumption that ab-

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 772–783,
Austin, Texas, November 1-5, 2016.

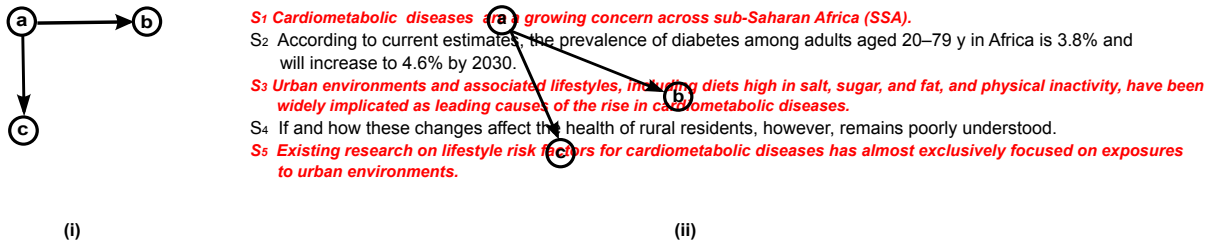

Figure 1: (i) A sample of mined coherence patterns from abstracts; nodes are sentences and edges are entity connections; (ii) Sentences S_1, S_3 and S_5 constitute the pattern in an input document.

stracts of scientific articles are similar in style to coherent summaries, we obtain coherence patterns by analyzing a corpus of abstracts of articles from bio-medicine (*PubMed* corpus). Then we apply the most frequent coherence patterns to input documents, i.e. long scientific articles from bio-medicine (*PLOS Medicine* dataset), extract corresponding sentences to generate coherent summaries, and evaluate them by comparing with summaries written by a *PLOS Medicine* editor. Figure 1 illustrates the extraction of sentences from an input document (Figure 1, (ii)) which constitute a coherence pattern (Figure 1, (i)). If we overlay the input document with coherence patterns and extract the sentences which constitute those patterns, then the extracted sentences are already coherent. We also take into account importance and non-redundancy. We capture all three factors in an objective function maximized by Mixed Integer Programming (MIP) (Section 2).

We evaluate our method on two different datasets: *PLOS Medicine* (Parveen and Strube, 2015) and *DUC 2002*. We extract frequent coherence patterns from all abstracts in the *PubMed* corpus, and generate summaries of unseen scientific articles of the *PLOS Medicine* dataset (Section 3.1). For DUC 2002 we extract coherence patterns from the human summaries of DUC 2005 (Dang, 2005). We evaluate our model on DUC 2002 to compare with state-of-the-art systems.

Our experimental results show that using coherence patterns for summarization produces more informative (but not redundant) and coherent summaries as compared to several baseline methods and state-of-the-art methods based on ROUGE scores and human judgements.

2 Method

We solve the task of creating coherent summaries by employing coherence patterns. We tightly integrate determining importance, non-redundancy and coherence by applying global optimization, i.e., MIP.

2.1 Document Representation

We use the entity graph (Guinaudeau and Strube, 2013) to represent scientific articles. The entity graph is a bipartite graph which consists of entities and sentences as two disjoint sets of nodes (Figure 2, ii). Entity nodes are connected only with sentence nodes and not among each other. An entity node is connected with a sentence node if and only if the entity is present in the sentence. Entities are the head nouns of noun phrases.

We perform a one-mode projection on sentence nodes to create a directed one-mode projection graph (Figure 2, iii). Two sentence nodes in the one-mode projection graph are connected if they share at least one entity in the entity graph. Edge directions encode the sentence order in the input document.

2.2 Mining Coherence Patterns

We use one-mode projection graphs of abstracts in the *PubMed* corpus (see Section 3.1) to mine coherence patterns. The weight of a coherence pattern, $weight(pat_u)$, is its frequency in the *PubMed* corpus normalized by the maximum number of its occurrence in abstracts in the *PubMed* corpus (Equation 1).

$$weight(pat_u) = \frac{\sum_{k=1}^{q} freq(pat_u, g_k)}{max_{k=1}^{q} freq(pat_u, g_k)}, \quad (1)$$

where q is the number of graphs associated with abstracts in the corpus, and g_k represents the graph of the k^{th} abstract in the *PubMed* corpus.

S_1 The overall $[rates]_{e_1}$ of cesarean $[delivery]_{e_2}$ are increasing significantly in the $[world]_{e_3}$.

S_2 In $[parts]_{e_4}$ of $[England]_{e_5}$ in 2010, the $[proportion]_{e_6}$ of total $[births]_{e_7}$ by cesarean $[section]_{e_8}$ was almost 25%, compared with just 2% in the 1950s.

S_3 In the United States and Australia rates of greater than 33% have been reported and in $[China]_{e_9}$ and $[parts]_{e_4}$ of South $[America]_{e_{10}}$, including Brazil and $[Paraguay]_{e_{11}}$, cesarean $[rates]_{e_1}$ of between 40% and 50% are common.

S_4 $[Concerns]_{e_{12}}$ have been expressed regarding the $[impact]_{e_{13}}$ of a cesarean $[section]_{e_8}$ on subsequent $[pregnancy]_{e_{14}}$ $[outcome]_{e_{15}}$ particularly the $[rate]_{e_{16}}$ of subsequent $[stillbirth]_{e_{17}}$, $[miscarriage]_{e_{18}}$, and ectopic pregnancy.

S_5 Hypothesized biological $[mechanisms]_{e_{19}}$ include placental $[abnormalities]_{e_{20}}$, prior $[infection]_{e_{21}}$, and adhesion $[formation]_{e_{22}}$ due to cesarean $[section]_{e_8}$.

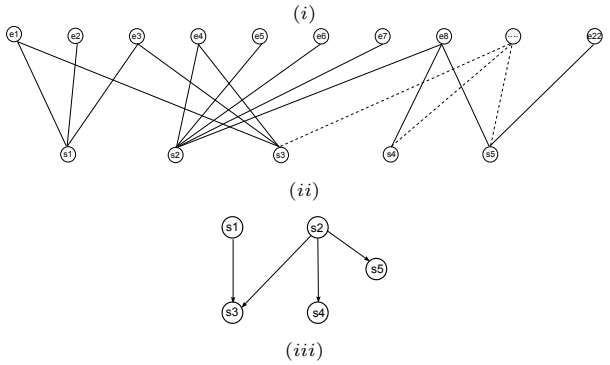

Figure 2: (i) A sample text from *PLOS Medicine*; (ii) entity graph; (iii) projection graph of the text.

The weights of the coherence patterns are not on the same scale. We normalize the weights using the standard score $\left(\frac{x-\mu}{\sigma}\right)$, where μ is the mean and σ is the standard deviation. A sigmoid function scales weights to the interval $[0, 1]$.

2.3 Summary Generation

We maximize importance, non-redundancy and pattern-based coherence with their respective weights λ to generate coherent summaries. The objective function is:

$$max(\lambda_I f_I(S) + \lambda_R f_R(E) + \lambda_C f_C(P)), \quad (2)$$

where S is a set of binary variables for sentences in an article, E is a set of binary variables for entities and P is a set of binary variables for coherence patterns.

Importance ($\mathbf{f_I(S)}$): The importance function quantifies the overall importance of information in the summary, which is calculated by considering the ranks of selected sentences for the summary:

$$f_I(S) = \sum_{i=1}^{n} Rank(sent_i) \cdot s_i. \quad (3)$$

In Equation 3, $Rank\,(sent_i)$ represents the rank of sentence $sent_i$ and s_i is the binary variable of sentence $sent_i$. n is the number of sentences.

Kleinberg (1999) develops the Hubs and Authorities algorithm (HITS) to rank web pages. He divides web pages into two sets: Hubs, pages which contain links to informative web pages, and Authorities, informative web pages. Here, Hubs are entities and Authorities are sentences. We calculate the rank of sentences using the HITS algorithm (Parveen and Strube, 2015). Initial ranks for sentences and entities are computed by Equations 4 and 5 in an entity graph:

$$Rank_{init}(sent_i) = 1 + sim(sent_i, title), \quad (4)$$

$$Rank_{init}(ent_j) = 1. \quad (5)$$

In Equation 4, $sim(sent_i, title)$ is the cosine similarity between the scientific article's title and sentence $sent_i$. In Equation 5, ent_j refers to the j^{th} entity in the entity graph. After applying the HITS algorithm on the entity graph using the above initialization, the final rank of a sentence is its importance.

Non-redundancy ($\mathbf{f_R}$ (E)): In the objective function, $f_R(E)$ represents the non-redundancy of information in the summary. Intuitively, if the summary has unique information in every sentence then the summary is non-redundant. We measure non-redundancy as follows:

$$f_R(E) = \sum_{j=1}^{m} e_j, \quad (6)$$

where m is the number of entities and e_j is a binary variable for each entity. The summary becomes non-redundant if we include only unique entities.

On the basis of $f_I(S)$ and $f_R(E)$ we define the following optimization constraints:

$$\sum_{i=1}^{n} |Sent_i| \cdot s_i \leq l_{max}, \quad (7)$$

$$\sum_{j \in E_i} e_j \geq |E_i| \cdot s_i \quad \text{for } i = 1, \ldots, n, \quad (8)$$

$$\sum_{s_i \in S_j} s_i \geq e_j \quad \text{for } j = 1, \ldots, m. \tag{9}$$

The constraint in Equation 7 limits the length of the summary. l_{max} is the maximal length of the summary and $|Sent_i|$ is the length of sentence $sent_i$.

In Equation 8, the constraint ensures that if sentence $sent_i$ is selected ($s_i = 1$), then all entities E_i present in sentence $sent_i$ must also be selected. In Equation 9, S_j represents the set of binary variables of sentences which contain entity ent_j. This constraint prescribes that if entity ent_j is selected ($e_j = 1$), then at least one of the sentences in S_j must be selected, too.

Coherence ($\mathbf{f_C(P)}$): We use the mined patterns to extract sentences from the input document of *PLOS Medicine* to create a coherent summary. We extract sentences, if the connectivity among nodes in their projection graph matches the connectivity among nodes in a coherence pattern. In Figure 3 we overlay the projection graph from Figure 2, ii with the coherence pattern from Figure 1, i. This results in three instances of this coherence pattern. However, we select only one since we simultaneously optimize for importance and non-redundancy.

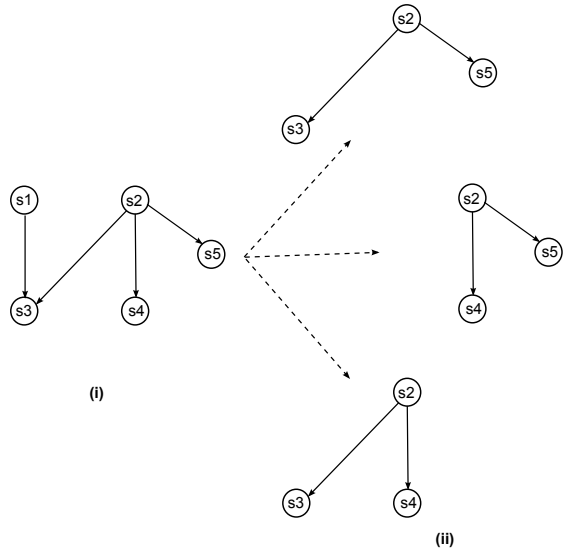

Figure 3: (i) A projection graph; (ii) several instances of a coherence pattern in Figure 1, ii.

In the objective function, $f_C(P)$ measures the coherence of the summary based on the weights of the coherence patterns occurring in it (Section 2.2):

$$f_C(P) = \sum_{u=1}^{U} weight(pat_u) \cdot p_u, \tag{10}$$

where p_u is a boolean variable associated with coherence pattern pat_u.

The optimization considers pattern pat_u for summarizing the input article, if pat_u is a subgraph of the projection graph of the article. To find the coherence pattern in a projection graph we apply a graph matching algorithm (Lerouge et al., 2015).

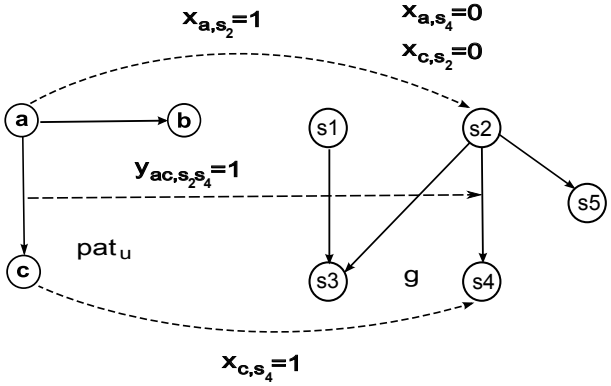

Figure 4: An illustration of mapping variables to overlay graph g with coherence pattern pat_u.

To model the graph matching problem between projection graph $g = (V_g, E_g)$ and patterns $pat_u = (V_{pat_u}, E_{pat_u})$, two kinds of mapping binary variables are used: $x_{i,k}$ for the node map, and $y_{ij,kl}$ for the edge map. $x_{i,k,} = 1$, if vertices $i \in V_{pat_u}$ and $k \in V_g$ match. $y_{ij,kl} = 1$, if for each pair of edges $ij \in E_{pat_u}$ and $kl \in E_g$ match (Figure 4). Constraints for graph matching are as follows:

- Every node of the pattern matches at most one unique node of the graph:

$$\sum_{k \in V_g} x_{i,k} \leq 1 \quad \forall i \in V_{pat_u}. \tag{11}$$

- Every edge of the pattern matches at most one unique edge of the graph:

$$\sum_{kl \in E_g} y_{ij,kl} \leq 1 \quad \forall ij \in E_{pat_u}. \tag{12}$$

- Every node of the graph matches at most one node of the pattern:

$$\sum_{i \in V_{pat_u}} x_{i,k} \leq 1 \quad \forall k \in V_g. \tag{13}$$

- A node of pattern pat_u matches a node of graph g if an edge originating from the node of pat_u matches an edge originating from the node of g:

$$\sum_{kl \in E_g} y_{ij,kl} = x_{i,k} \ \forall k \in V_g, \forall ij \in E_{pat_u}. \tag{14}$$

- A node of pattern pat_u matches a node of graph g if an edge targeting the node of pat_u matches an edge targeting the node of g:

$$\sum_{kl \in E_g} y_{ij,kl} = x_{j,l} \ \forall l \in V_g, \forall ij \in E_{pat_u}. \tag{15}$$

- We need a constraint to extract *induced* patterns[1]:

$$\sum_{i \in V_{pat_u}} x_{i,k} + \sum_{j \in V_{pat_u}} x_{j,l} - \sum_{ij \in E_{pat_u}} y_{ij,kl} \leq 1 \quad \forall kl \in E_g. \tag{16}$$

The constraints in Equations $11-16$ are defined to find pattern pat_u in projection graph g of the input article. However these constraints do not ensure that the pattern is in the summary. For this, we define constraints in Equations $17-19$ to assure that an existing pattern in an article is selected if there are some sentences in the summary which constitute the pattern.

- The constraint in Equation 17 ensures that if sentences s_k and s_l are selected for the summary then the edge between them is selected ($z_{kl} = 1$), too:

$$s_k \cdot s_l = z_{kl} \quad \forall k, l \in V_g. \tag{17}$$

- Pattern pat_u is present in the summary ($p_u = 1$) if and only if one of its instances in the projection graph is included in the summary, i.e., some of the selected sentence nodes must be present in an instance of pattern pat_u. $|V_{pat_u}|$ is the number of nodes in pattern pat_u, and $|E_{pat_u}|$ is the number of edges in pattern pat_u. This constraint is shown below:

$$\sum_{i \in v_{pat_u}} \sum_{k \in v_g} s_k \cdot x_{i,k} + \sum_{ij \in e_{pat_u}} \sum_{kl \in e_g} z_{kl} \cdot y_{ij,kl} = p_u(|V_{pat_u}| + |E_{pat_u}|). \tag{18}$$

- If a sentence is selected then it has to match a node of at least one of the patterns:

$$\sum_{pat_u \in P} \sum_{i \in V_{pat_u}} x_{i,k} \geq s_k \quad \forall k \in V_g. \tag{19}$$

3 Experiments

In this section we discuss the datasets and the experimental setup. We evaluate our model using ROUGE scores and human judgements.

3.1 Datasets

PLOS Medicine: This dataset contains 50 scientific articles. In this dataset every scientific article is accompanied by a summary written by an editor of the month. This editor's summary has a broader perspective than the authors' abstract. We use the editor's summary as a gold summary for calculating the ROUGE scores. We use 700 different *PLOS Medicine* articles from the PubMed[2] corpus to mine coherence patterns from their abstracts and to calculate patterns' weights.

DUC: The DUC 2002 dataset has been annotated for the Document Understanding Conference 2002. It contains 567 news articles for summarization. Every article is accompanied by at least two gold summaries. DUC 2002 articles are shorter than *PLOS Medicine* articles (25 vs. 154 sentences average length). We use all (300) DUC 2005 human summaries to mine coherence patterns and to calculate their weights.

3.2 Experimental Setup

First, we extract the text of an article. We remove figures, tables, references and non-alphabetical characters. Then we use the Stanford parser (Klein and

[1] Pattern pat_u is an induced subgraph of graph g if pat_u contains all possible edges which appear in g.

[2] http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

Manning, 2003) to determine sentence boundaries. We apply the Brown coherence toolkit (Elsner and Charniak, 2011) to convert the articles into entity grids (Barzilay and Lapata, 2008) which then are transformed into entity graphs. We use gSpan (Yan and Han, 2002) to extract all subgraphs from the projection graphs of the abstracts of the *PubMed* corpus.

It is possible that patterns with a large number of nodes are not at all present in the projection graph. Hence, we use coherence patterns with 3 and 4 nodes, referred to as CP_3 and CP_4, respectively. We use Gurobi (Gurobi Optimization, Inc., 2014) to solve the MIP problem. We use a pronoun resolution system (Martschat, 2013) to replace all pronouns in the summary with their antecedents.

We determine the best values for λ_I, λ_R, and λ_c on the development sets. $\lambda_I = 0.4$, $\lambda_R = 0.3$, and $\lambda_c = 0.3$ are the best weights for the *PLOS Medicine* development set. Weights for the DUC 2002 development set are $\lambda_I = 0.5$, $\lambda_R = 0.2$ and $\lambda_c = 0.3$.

3.3 Results

We evaluate our model in two ways. First, we use ROUGE scores to compare our model with other models. Second, we explicitly evaluate the coherence of the summaries by human judgements.

3.3.1 ROUGE Assessment

The ROUGE score (Lin, 2004) is a standard evaluation score in automatic text summarization. It calculates the overlap between gold summary and system summary. In automatic text summarization ROUGE 1, ROUGE 2 and ROUGE SU4 are usually reported (see Graham (2015) for an assessment of evaluation metrics for summarization).

We compare our system (CP_3 and CP_4) with four baselines: *Lead, Random, Maximal Marginal Relevance (MMR)* and *TextRank*. *Lead* selects adjacent sentences from the beginning of an input article. *Random* selects sentences randomly. *MMR* (Carbonell and Goldstein, 1998) uses a trade-off between relevance and redundancy. *TextRank* is a graph-based system using sentences as nodes and edges weighted by cosine similarity between sentences (Mihalcea and Tarau, 2004).

We compare our system with three state-of-the-art systems: E_{Coh} (Parveen and Strube, 2015), T_{Coh} (Parveen et al., 2015), and *Mead* (Radev et al., 2004). E_{Coh} uses entity graphs which consists of entities and sentences, and T_{Coh} uses topical graphs where entities are replaced by the topics. They both use the outdegree of sentence nodes in the unweighted and the weighted projection graph, respectively, as the coherence measure of each sentence. *Mead* employs a linear combination of three features: centroid score, position score and overlap score. The linear combination is used to add sentences to the summary up to the required length. The centroid score gives the highest score to the most central sentence in the cluster of sentences, the position score gives a higher score to the sentences which are in the beginning of the document, and the overlap score computes the similarity between the sentences of a document. All three features do not take care of the coherence of a summary as they do not have any notion of the order and the structure of a summary.

Systems	R-SU4	R-2
Baselines		
Lead	0.067	0.055
Random	0.048	0.031
MMR	0.069	0.048
TextRank	0.068	0.048
State-of-the-art		
E_{Coh}	0.131	0.098
T_{Coh}	0.129	0.095
Mead	0.084	0.068
Our Model		
$\mathbf{CP_3}$	**0.135**	**0.103**

Table 1: *PLOS Medicine*, editor's summaries with **5 sentences**.

To compare with the state-of-the-art systems on PLOS Medicine, E_{Coh} (Parveen and Strube, 2015) and T_{Coh} (Parveen et al., 2015), we limit the length of summaries to 5 sentences. Table 1 reports ROUGE scores of different systems. Our system outperforms baselines and state-of-the-art systems.

Since the word length limit of a summary is more meaningful than the sentence length limit of a summary, we limit the length of a summary to the average length of editor's summaries in the dataset (750 words). Table 2 shows the performance of different systems with 750 words limit for a summary. In Table 2, we use different versions of ROUGE-SU4 and ROUGE-2 where *W/WO* stands

PLOS Medicine Editor's summaries	WO_{Stop} W_{Stem}	WO_{Stop} WO_{Stem}	W_{Stop} W_{Stem}	W_{Stop} WO_{Stem}	WO_{Stop} W_{Stem}	WO_{Stop} WO_{Stem}	W_{Stop} W_{Stem}	W_{Stop} WO_{Stem}
	ROUGE SU4 (*$p_{value} < 0.05$)				ROUGE 2 (*$p_{value} < 0.01$)			
Upper Bound	0.423	0.354	0.519	0.470	0.344	0.304	0.430	0.399
Baselines								
Lead	0.191	0.158	0.246	0.222	0.158	0.140	0.185	0.171
Random	0.140	0.113	0.169	0.153	0.102	0.088	0.125	0.116
MMR	0.183	0.149	0.240	0.215	0.141	0.125	0.171	0.157
TextRank	0.148	0.104	0.161	0.159	0.115	0.084	0.126	0.118
State-of-the-art								
E_{Coh}	0.204*	0.167	0.254	0.228	0.160*	0.145	0.187	0.173
T_{Coh}	0.195	0.161	0.231	0.206	0.157	0.140	0.169	0.165
Mead	0.197	0.165	0.246	0.222	0.156	0.139	0.186	0.172
Our Model								
CP_3	0.215*	0.178	0.268	0.241	0.172*	0.153	0.200	0.184
CP_4	0.218	0.179	0.270	0.245	0.175	0.156	0.201	0.187

Table 2: ROUGE scores on *PLOS Medicine* with **750 words**.

for *With/Without*. Here, WO_{Stop} means without considering stopwords while calculating ROUGE scores, and WO_{Stem} means without applying the Porter Stemmer on summaries while calculating ROUGE scores. Our models outperform baseline and state-of-the-art systems (Table 2). We compute statistical significance between E_{Coh} and CP_3 on both scores, ROUGE SU4 is significantly different by 95%. ROUGE 2 is significantly different by 99%.

Upper Bound in Table 2 represents maximum ROUGE scores that can be achieved in extractive summarization on the *PLOS Medicine* dataset. It is calculated by considering the whole scientific article as a summary and the corresponding editor's summary as the gold standard. The *Upper Bound* scores are not very high showing that a significant improvement in ROUGE scores on the *PLOS Medicine* dataset is difficult. Thus, the performance achieved by our systems, CP_3 and CP_4, is a considerable improvement on the *PLOS Medicine* dataset.

Furthermore, we apply CP_3 on the dataset introduced by Liakata et al. (2013). The dataset consists of 28 scientific articles from the chemistry domain. The state-of-the-art system on this dataset is *CoreSC*, which is developed by Liakata et al. (2013). *CoreSC* considers discourse information while summarizing a scientific article. The ROUGE-1 score of CP_3 (0.96) is significantly better than *CoreSC* (0.75) and *Microsoft Office Word 2007 AutoSummarize* (0.73) (García-Hernández et al., 2009), in respect of abstracts. This shows that our system performs well in other domains.

We further calculate the average number of sentences per summary obtained by *Mead* and CP_3. On average *Mead* produces 17.5 sentences per summary whereas CP_3 produces 27.2 sentences per summary. The possibility of longer sentences containing more topic irrelevant entities is higher than shorter sentences (Jin et al., 2010).

We calculate the average percentage of sentences selected from the sections Introduction, Method, Results and Discussion by different systems. CP_3 extracts sentences mainly from Introduction (32.5%) and Method (38.5%), but also a considerable number of sentences from Results (17.67%) and Discussion (11.33%). The distribution is quite similar to *TextRank* and *MMR*. *Lead*, obviously, extracts only from Introduction (80.59%) and Method (19.41%). *Mead* extracts maximum sentences from the beginning of the document using its positional feature. The sentences in a summary extracted by CP_3 are evenly distributed indicating that they are not biased to any sections. This clearly represents that coherence patterns not only seeks for nearby sentences but also for any distant sentences of a scientific article.

Table 3 shows the results on DUC 2002 to compare the results with state-of-the-art systems. There is no significant difference between the ROUGE scores of using CP_3 and CP_4 on DUC 2002. Thus, we only report the results of using CP_3 on DUC 2002.

In Table 3, *LREG* is a baseline system us-

Systems	R-1	R-2	R-SU4
Baselines			
Lead	0.459	0.180	0.201
DUC 2002 Best	0.480	0.228	
TextRank	0.470	0.195	0.217
LREG	0.438	0.207	
State-of-the-art			
Mead	0.445	0.200	0.210
ILP_{phrase}	0.454	0.213	
URANK	0.485	0.215	
UniformLink (k = 10)	0.471	0.201	
E_{Coh}	0.485	0.230	0.253
T_{Coh}	0.481	0.243	0.242
NN-SE	0.474	0.230	
Our Model			
$\mathbf{CP_3}$	**0.490**	**0.247**	**0.258**

Table 3: ROUGE scores on DUC 2002.

ing logistic regression and hand-made features (Cheng and Lapata, 2016). We compare our model to previously published state-of-the-art systems. These systems show reasonable performance on the DUC 2002 summarization task. ILP_{phrase} is a phrase-based extraction model, which selects important phrases and combines them via integer linear programming (Woodsend and Lapata, 2010). *URANK* utilizes a unified ranking process for single-document and multi-document summarization tasks (Wan, 2010). *UniformLink (k=10)*, considers similar documents for document expansion in the single-document summarization task (Wan and Xiao, 2010). The more recent system, *NN-SE*, utilizes a neural network hierarchical document encoder and an attention-based extractor to extract sentences from a document for a summary (Cheng and Lapata, 2016). ROUGE scores of our approach on this dataset are better than baselines and state-of-the-art systems. This shows that our system performs well even in a different genre (robust) and with considerably shorter input documents (scalable).

3.3.2 Coherence Assessment

ROUGE scores do not evaluate summary coherence, since ROUGE only calculates overlapping recall scores and does not consider the structure of the summary. Haghighi and Vanderwende (2009), Celikyilmaz and Hakkani-Tür (2010) and Christensen et al. (2013) evaluate the overall summary quality by asking human subjects to rank system generated summaries. Parveen and Strube (2015) and Parveen et al. (2015) assess the coherence by asking human assessors to rank system generated summaries and compare their system with baseline systems.

We perform summary coherence assessment by asking one Postdoc, two PhD students and one Masters student from the field of natural language processing. We provide them with the output summaries of four different systems for ten articles. We ask them to rank the summaries, i.e., the best summary gets rank 1, the second best gets rank 2, the third best gets rank 3, and the worst gets rank 4.

The four systems assessed are CP_3, E_{Coh}, *TextRank*, and *Lead*. We apply the Kendall concordance coefficient (W) (Siegel and Castellan, 1988) to measure whether the human assessors agree in ranking the four systems. With W $= 0.6725$ the correlation between the human assessors is high. Applying the χ^2 test shows that W is significant at least at the 99% level indicating that the ranks provided by the human assessors are reliable and informative. Table 4 shows the overall average rank of a system given by the four human assessors. The lower the value of average human scores the more coherent the summary. Unsurprisingly *Lead* gets the best overall average rank. *Lead* extracts adjacent sentences from the beginning of the document. Hence, these summaries are as coherent as the author intends them to be, but they are not informative. However, CP_3 is very close in coherence to *Lead* indicating that our strategy is successful. It also performs substantially better than *TextRank* and E_{Coh}. This confirms that using coherence patterns for sentence extraction yields more coherent summaries.

PLOS Medicine System	Average Human Score
TextRank	3.950
E_{Coh}	2.325
$\mathbf{CP_3}$	**1.875**
Lead	1.625

Table 4: The average human scores.

4 Related Work

Summarizing scientific articles is as difficult as multi-document summarization because scientific articles are tend to be long and the important infor-

mation is spread all over the article unlike information in news articles (Teufel and Moens, 2002).

There are various approaches for summarizing scientific articles. Citations have been used by many researchers for summarization in this domain (Elkiss et al., 2008; Mohammad et al., 2009; Qazvinian and Radev, 2008; Abu-Jbara and Radev, 2011). Nanba and Okumura (2000) develop rules for categorizing citations by analyzing citation sentences. Newman (2001) analyzes the structure using a citation network. Similarly, Siddharthan and Teufel (2007) discover scientific attributions using citations. Discourse structure (but not necessarily coherence) has been used by Teufel and Moens (2002), Liakata et al. (2013) and others for summarizing scientific articles.

Several state-of-the-art extractive summarization systems implement summarization as maximizing an objective function using constraints. McDonald (2007) interprets text summarization as a global inference problem, where he is maximizing the importance score of a summary by considering the length constraint. Similarly, various approaches for summarization are based on optimization using ILP (Gillick et al., 2009; Nishikawa et al., 2010; Galanis et al., 2012; Parveen and Strube, 2015).

Until now, only few works have considered coherence while summarizing scientific articles. Abu-Jbara and Radev (2011) work on citation based summarization. They preprocess the citation sentences to filter out irrelevant sentences or sentence fragments, then extract sentences for the summary. Eventually, they refine the summary sentences to improve readability. Jha et al. (2015) consider Minimum Independent Discourse Contexts (MIDC) to solve the problem of non-coherence in extractive summarization. However, none of them deals with the problem of coherence within the task of sentence selection. Sentence selection and ensuring the coherence of summaries are not tightly integrated in their techniques. They model coherence in summarization by only considering adjacent sentences.

There are few methods (Hirao et al., 2013; Parveen and Strube, 2015; Gorinski and Lapata, 2015) which integrate coherence in optimization. These methods do not take into account the overall structure of the summary. Unlike earlier methods, we incorporate coherence patterns in optimization.

5 Conclusion

We introduce a novel graph-based approach to generate coherent summaries of scientific articles. Our approach takes care of coherence distinctively by coherence patterns. We have experimented with *PLOS Medicine* and DUC 2002. The results show that the approach is robust, works on both scientific and news documents and with input documents of different length. It considerably outperforms state-of-the-art systems on both datasets. We collected human assessments to evaluate the coherence of summaries. Our system substantially outperforms baselines and state-of-the-art systems, i.e., incorporating coherence patterns produces more coherent summaries. The results show that our approach performs well in human summary coherence assessment and relevance evaluation (ROUGE scores).

Acknowledgments

This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first and second authors have 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 Alexander Judea, Isabell Wolter, Mark-Christoph Müller and Nafise Moosavi who became human subjects for coherence assessment evaluation.

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 (Volume 1: Long Papers),* Portland, Oreg., 19–24 June 2011, pages 500–509.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

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.

Asli Celikyilmaz and Dilek Hakkani-Tür. 2010. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pages 815–824.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Berlin, Germany, 7–12 August 2016, pages 484–494.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multi-document summarization. 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 1163–1173.

František Daneš, editor. 1974. *Papers on Functional Sentence Perspective*. Academia, Prague.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the 2005 Document Understanding Conference held at the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing,* Vancouver, B.C., Canada, 9–10 October 2005.

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir Radev. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1):51–62.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* Portland, Oreg., 19–24 June 2011, pages 125–129.

Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of the 24th International Conference on Computational Linguistics,* Mumbai, India, 8–15 December 2012, pages 911–926.

René Arnulfo García-Hernández, Yulia Ledeneva, Griselda Matías Mendoza, Ángel Hernández Dominguez, Jorge Chavez, Alexander Gelbukh, and José Luis Tapia Fabela. 2009. Comparing commercial tools and state-of-the-art methods for generating text summaries. In *Proceedings of Advances in Artificial Intelligence, 8th Mexican International Conference on Artificial Intelligence,* Guanajuato, Mexico, 9-13 November 2009, pages 92–96.

Daniel Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. 2009. A global optimization framework for meeting summarization. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing,* Taipei, Taiwan, 19–24 June 2009, pages 4769–4772.

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.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, 17–21 September 2015, pages 128–137.

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.

Gurobi Optimization, Inc. 2014. Gurobi optimizer reference manual.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pages 362–370.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Seattle, Wash., 18–21 October 2013, pages 1515–1520.

Rahul Jha, Reed Coke, and Dragomir Radev. 2015. Surveyor: A system for generating coherent survey articles for scientific topics. In *Proceedings of the 29th Conference on the Advancement of Artificial Intelligence,* Austin, Texas, 25–30 January 2015, pages 2167–2173.

Feng Jin, Minlie Huang, and Xiaoyan Zhu. 2010. A comparative study on ranking and selection strategies for multi-document summarization. In *Proceedings of Coling 2010: Poster Volume,* Beijing, China, 23–27 August 2010, pages 525–533.

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.

Julien Lerouge, Pierre Le Bodic, Pierre Héroux, and Sébastien Adam. 2015. GEM++: A tool for solving substitution-tolerant subgraph isomorphism. In C.-L. Liu, B. Luo, W.G. Kropatsch, and J. Cheng, editors, *Graph-Based Representations in Pattern Recognition*, pages 128–137. Springer, Heidelberg, Germany.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Seattle, Wash., 18–21 October 2013, pages 747–757.

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.

Mohsen Mesgar and Michael Strube. 2015. Graph-based coherence modeling for assessing readability. In *Proceedings of STARSEM 2015: The Fourth Joint Conference on Lexical and Computational Semantics,* Denver, Col., 4–5 June 2015, pages 309–318.

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.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics,* Boulder, Col., 31 May – 5 June 2009, pages 584–592.

Hidetsugu Nanba and Manabu Okumura. 2000. Producing more readable extracts by revising them. In *Proceedings of the 18th International Conference on Computational Linguistics,* Saarbrücken, Germany, 31 July – 4 August 2000, pages 1071–1075.

Mark E.J. Newman. 2001. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):016131.

Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo, and Genichiro Kikui. 2010. Opinion summarization with integer linear programming formulation for sentence extraction and ordering. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 23–27 August 2010, pages 910–918.

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.

Daraksha Parveen, Hans-Martin Ramsl, and Michael Strube. 2015. Topical coherence for graph-based extractive summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, 17–21 September 2015, pages 1949–1954.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics,* Manchester, U.K., 18–22 August 2008, pages 689–696.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celibi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. MEAD – a platform for multidocument multilingual text summarization. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26–28 May 2004.

Advaith Siddharthan and Simone Teufel. 2007. Whose idea was this, and why does it matter? Attributing scientific work to citations. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, N.Y., 22–27 April 2007, pages 316–223.

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

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

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.

Xiaojun Wan. 2010. Towards a unified approach to simultaneous single-document and multi-document summarizations. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 23–27 August 2010, pages 1137–1145.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pages 565–574.

Xifeng Yan and Jiawei Han. 2002. gSpan: Graph-based substructure pattern mining. In *Proceedings of the International Conference on Data Mining,* Maebashi City, Japan, 9–12 December 2002, pages 721–724.