A Proposition-Level Clustering Approach for Multi-Document Summarization

Ori Ernst[1], Avi Caciularu[1*], Ori Shapira[1*], Ramakanth Pasunuru[2], Mohit Bansal[2], Jacob Goldberger[1], and Ido Dagan[1]

[1]Bar-Ilan University [2]UNC Chapel Hill
{oriern, avi.c33, obspp18}@gmail.com
{ram, mbansal}@cs.unc.edu
{jacob.goldberger@, dagan@cs.}biu.ac.il

Abstract

Text clustering methods were traditionally incorporated into multi-document summarization (MDS) as a means for coping with considerable information repetition. Clusters were leveraged to indicate information saliency and to avoid redundancy. These methods focused on clustering *sentences*, even though closely related sentences also usually contain non-aligning information. In this work, we revisit the clustering approach, grouping together *propositions* for more precise information alignment. Specifically, our method detects salient propositions, clusters them into paraphrastic clusters, and generates a representative sentence for each cluster by fusing its propositions. Our summarization method improves over the previous state-of-the-art MDS method in the DUC 2004 and TAC 2011 datasets, both in automatic ROUGE scores and human preference.[1]

1 Introduction

Common information needs are most often satisfied by multiple texts rather than by a single one. Accordingly, there is a rising interest in Multi-Document Summarization (MDS) — generating a summary for a set of topically-related documents. Inherently, MDS needs to address, either explicitly or implicitly, several subtasks embedded in this summarization setting. These include salience detection, redundancy removal, and text generation. While all these subtasks are embedded in Single-Document Summarization (SDS) as well, the challenges are much greater in the multi-document setting, where information is heterogeneous and dispersed, while exhibiting substantial redundancy across linguistically divergent utterances. Indeed, compared to recent impressive progress in SDS, MDS quality is still lagging.

• *The trial opened Thursday of 29* mostly Moroccan *suspects charged with involvement in the 2004 Madrid train bomb attacks*, which killed 191 people and injured 1,824 in the worst terror strike to hit Spain.
• *Of the 29 people who go on trial Thursday for the March 2004 Madrid train bombings*, seven face some 40,000 years in jail if found guilty.
• But the man, Rabei Osman Sayed Ahmed – expected to be the first of *29 defendants to take the stand when the bombing trial begins Thursday in Madrid* – also said in the recordings that the attack was carried out according to his plan.

Table 1: An example of a cluster of *propositions*, shown within their source sentence context, from TAC 2011 (topic D1103). Clustering these as sentences would yield noisy unaligned information, however grouping together only the marked propositions keeps information alignment clean.

An intuitive summarization approach that copes with these challenges, and is especially relevant for MDS, is clustering-based summarization. In such an approach, the goal is to cluster together redundant paraphrastic pieces of information which roughly convey the same meaning. Repetition of information across texts, a common property of MDS that is extracted by paraphrastic clustering, typically indicates its importance, and can be leveraged for salience detection. Moreover, a cluster of paraphrases may facilitate generating a corresponding summary text that eliminates repetitions while fusing together complementing information pieces within the cluster.

Traditionally, clustering-based approaches were widely used for summarization, mostly in an extractive and unsupervised manner. One such approach clustered topically-related sentences, after which cluster properties were leveraged for rating sentence salience (Radev et al., 2004; Wang et al., 2008; Wan and Yang, 2008). Another approach rated sentence salience and clustered sentences simultaneously, iteratively improving the two objectives (Cai et al., 2010; Wang et al., 2011; Cai and Li, 2013; Zhang et al., 2015). Recently, however, clustering methods have been gradually marginalized out, being replaced by neural techniques, mostly

[*] Equal contribution.
[1]Our code and system summaries are publicly available at https://github.com/oriern/ClusterProp

end-to-end. More recently, some approaches (Nayeem et al., 2018; Fuad et al., 2019) presented abstractive clustering-based summarization, where topically-related sentences in each cluster are fused together to generate a summary sentence candidate. Yet, all these works generated sentence-based clusters that tend to be noisy, since a sentence typically consists of several units of information that only partially overlap with other cluster sentences. As a result, such clusters often capture topically related sentences rather than high quality paraphrases. Table 1 exemplifies such a noisy cluster that contains paraphrastic propositions (marked in *blue*) with their full sentences as context (marked in black). As can be seen, considering the full sentence diverts the focus from a single information unit to a wider scope of topically-related information. Consequently, another line of research in summarization looked into the use of sub-sentential units for the summarization process. For example, Li et al. (2016) summarize with elementary discourse units (EDUs), and Ernst et al. (2021) endorse the use of OpenIE-based propositions (Stanovsky et al., 2018) for summarizing.

In this paper, we revisit and combine these two earlier design choices which were proposed for MDS and explored only individually and rather scarcely in recent years: clustering related and redundant information, and basing the summarization process on sub-sentential propositions. Specifically, we extend clustering-based summarization to apply at the more fine-grained *propositional* level, which avoids adding non-aligning pieces of information and provides accurate paraphrastic clusters. Working with standalone propositions yields cleaner clusters, providing a clear content scope for each cluster. This supports more accurate detection of redundancy, better salience ranking, and heightened control over the generated summary sentences – as the generation component is only required to fuse similar propositions.

To that end, our model (§3) leverages a dedicated supervised proposition similarity metric (with fine-tuned CDLM (Caciularu et al., 2021) and SuperPAL (Ernst et al., 2021)) as a basis for an agglomerative clustering algorithm (Ward, 1963), then rank all clusters by salience, and finally generate a coherent abstractive summary sentence per cluster using a fine-tuned BART model (Lewis et al., 2020). This process produces a bullet-style summary of concise and coherent sentences, each containing roughly one proposition.

Overall, our experiments (§5) show that this multi-step model outperforms strong recent end-to-end solutions, which do not include explicit modeling of propositions and information redundancy. To the best of our knowledge, our approach achieves state-of-the-art results in our setting on the DUC 2004 and TAC 2011 datasets, with an improvement of more than 1.5 and 4 ROUGE-1 F1 points respectively, over the previous best approach. Additionally, our proposed method lays the foundation for directly addressing supplemental aspects of the summarization process, like sentence planning and surface realization, which will likely further improve summary quality.

Finally, we also suggest (§6) that clustering-based methods provide “explanations", or supporting evidence, for each generated sentence, in the form of the source propositions in the cluster from which the sentence was generated (see an example in Table 2). In applied settings, these supportive clusters can be leveraged interactively to expand on a specific sub-topic. In addition, one can use these explanations to validate the faithfullness of each generated sentence to its source, or detect hallucinations. This is not trivial in the MDS setup, where source documents should be read in full to validate each summary sentence. In fact, as far as we know, we are the first to suggest a feasible annotation approach for fact validation (i.e, faithfulness) for multi-document summarization.

2 Background and Related Work

Sub-sentence unit based summarization. While most summarization approaches extract full document sentences, especially for extractive summarization, there are methods that work on the sub-sentential level. Li et al. (2016) produced extractive summaries consisting of Elementary Discourse Units (EDUs) – clauses comprising a discourse unit according to Rhetorical Structure Theory (RST). Such extractive approaches usually focus on content selection, possibly disregarding the inferior coherence arising from the concatenation of sub-sentence units. Accordingly, Arumae et al. (2019) established the highlighting task, where salient sub-sentence units are marked within their document to provide context around the salient units. Recently, Cho et al. (2020) proposed self-contained sub-sentence units, obtained heuristically by a language model score for adding

an EOS token at the begining and the end of the text unit.

Conversely, abstractive approaches extract sub-sentence units as a preliminary step for generation. Text units range from words (Lebanoff et al., 2020; Gehrmann et al., 2018), to noun or verb phrases (Bing et al., 2015), to full sentences (Song et al., 2018). In this work, we follow the same extract-then-generate pipeline, using Open Information Extraction (OpenIE) spans (Stanovsky et al., 2018) as proposition units. Since propositions are meant to contain single standalone facts, they are beneficial for grouping paraphrases with reduced dissimilarities. In addition, propositions, extracted with OpenIE, can be noncontiguous, while alternative options, like EDUs, are contiguous sequences.

Multi Document Summarization. The DUC and TAC[2] datasets are popularly employed for the MDS task, and are considered to be of high quality. However, their relatively small sizes, a few hundreds of multi-document instances total, may be insufficient for training MDS models. Accordingly, previous works overcame the shortage of data by supplementing external training datasets, usually of SDS, to pretrain their model (Mao et al., 2020; Cho et al., 2019). Others avoided using data-hungry neural methods and applied optimization methods, such as Determinantal Point Processes (DPP) (Cho et al., 2019) and submodular methods (Lin and Bilmes, 2010), or unsupervised methods (Nayeem et al., 2018; Zhao et al., 2020). In this work, we finetuned several models for summarization subtasks with DUC and TAC datasets only.

3 Method

This section details our clustering-based summarization pipeline. First, we extract all propositions from the input set of documents (§3.1) and filter out non-salient propositions with a salience-detection model (§3.2). Then, all salient propositions are clustered into groups based on their semantic similarity (§3.3). The largest clusters, i.e, those containing information that is more redundant across the documents, are selected to participate in the summary (§3.4). Finally, each cluster is fused to form a sentence for the abstractive summary (§3.5). The full pipeline is presented in Figure 1.

[2] https://{duc,tac}.nist.gov

3.1 Proposition Extraction

Aiming to generate proposition-based summaries, as mentioned in §2, we first extract all propositions from the source documents using Open Information Extraction (OpenIE) (Stanovsky et al., 2018), following Ernst et al. (2021).

3.2 Proposition Salience Model

To enable filtering of non-salient propositions, we fine-tuned the Cross-Document Language Model (CDLM) (Caciularu et al., 2021) as a binary classifier for predicting whether a proposition is salient or not. Propositions with a salience score below a certain threshold were filtered out. The threshold was optimized with the full pipeline against the final ROUGE score on the validation set. CDLM is pretrained with sets of related documents, and was hence shown to operate well over several downstream tasks in the multi-document setting (e.g, cross-document coreference resolution and multi-document classification).

As input, the finetuned CDLM model is fed with a proposition within its document and the other documents in the set. Specifically, since CDLM's input size is limited to 4,096 tokens, it is unfeasible to feed the full document set. Therefore, following Lebanoff et al. (2019), only the first 20 sentences of each document are considered. Accordingly, a candidate proposition is input within its full document (up to 20 sentences), while other documents, ordered by their date, are truncated evenly and concatenated to fill the remaining space (9 sentences per document on average).

For training data, we obtain gold labels for proposition salience by means of containment within oracle extractive summaries. An oracle summary is generated by greedily appending propositions that maximize ROUGE-$1_{F\text{-}1}$ + ROUGE-$2_{F\text{-}1}$ against the corresponding reference summaries (Nallapati et al., 2017; Liu and Lapata, 2019). Only propositions included in the oracle summary are marked as salient.

3.3 Clustering

Next, all salient propositions are clustered according to their semantic similarity. Paraphrastic clusters are advantageous for summarization as they can assist in avoiding selection of redundant information for an output summary. Furthermore, paraphrastic clustering offers an additional indicator for salience of propositions. The salience model de-

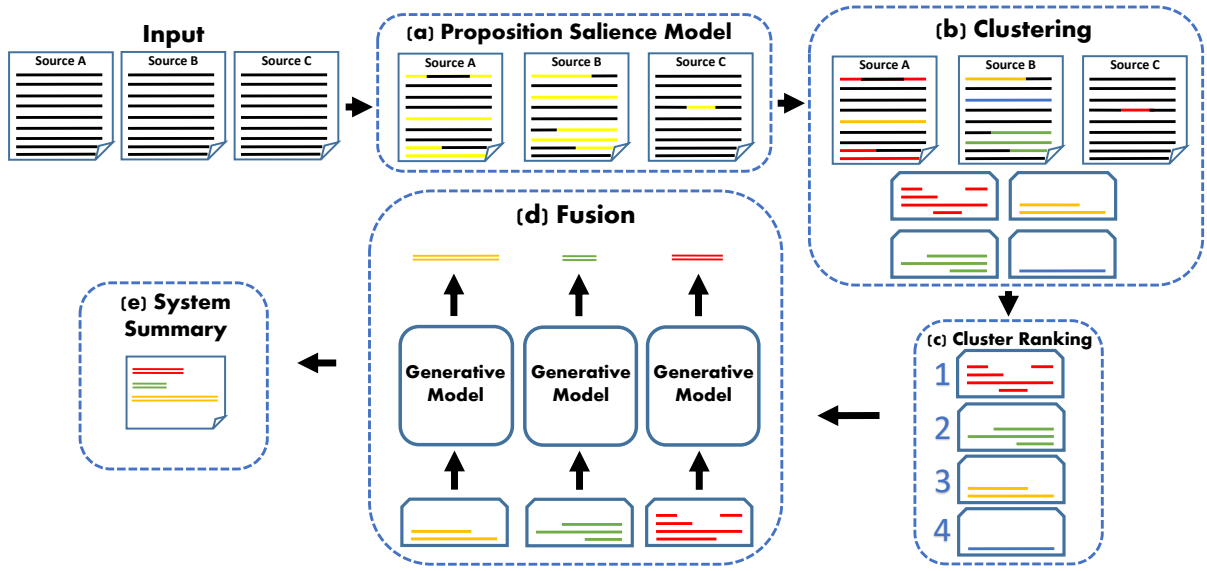

Figure 1: Our multi-step multi-document summarization process. (a) All propositions are extracted (OpenIE: Stanovsky et al., 2018) from the documents and issued a salience score (fine-tuned CDLM: Caciularu et al., 2021). (b) Salient propositions are clustered (fine-tuned SuperPAL: Ernst et al., 2021), forming groups of paraphrastic information units. (c) Clusters are ranked, as an indicator for information importance. (d) For each cluster, its propositions are fused (fine-tuned BART: Lewis et al., 2020) to generate a concise and coherent abstractive sentence. (e) The output summary is obtained as a bullet-style ranked list of the concise sentences.

scribed in Section 3.2 accounts for general context, while features of the paraphrastic clusters suggest the importance of a proposition in reference to its repetitions.

To cluster propositions we adopt SuperPAL (Ernst et al., 2021), a binary classifier that measures similarity between two propositions. All salient proposition pairs are scored with SuperPAL, on which standard agglomerative clustering (Ward, 1963) is applied. Examples of generated clusters from the test set are presented in Table 2.

3.4 Ranking

The resulting proposition clusters are now ranked according to cluster-based features. We examined various features, listed in Table 3, on our validation sets. For each feature, (1) clusters were ranked according to the feature, (2) the proposition with the highest salience model score (Section 3.2) was selected from each cluster as a cluster representative, (3) the representatives from the highest ranked clusters were concatenated to obtain a system summary. The resulting ROUGE scores of these summaries on validation sets are presented in Table 3.[3]

[3]We also tried training a regression model on a mixture of features that should predict the ROUGE score of a proposition, but results were comparable. Bettering the ranking process is left for future work.

The features examined include: average of ROUGE scores between all propositions in a cluster (*'Avg. inner ROUGE'*), average of SuperPAL scores between all propositions in a cluster (*'Avg. inner SuperPAL'*), average of the salience model scores of cluster propositions (*'Avg. salience'*), minimal position (in a document) of cluster propositions (*'Min. position'*), and cluster size (*'Cluster size'*). We also measured combinations of two features (*'Cluster size + Min. position'* for example), where the first feature is used for primary ranking, and the second feature is used for secondary ranking in case of a tie. We find that *'Cluster size'* yields the best ROUGE scores as a single feature, and *'Min. position'* further improves results as a secondary ranking feature.

Alongside the justifications posed above, intuitively, a large cluster represents a piece of information that is mentioned many times across documents, and hence is likely of higher importance to the topic. For all the reasons laid out, we accordingly rank the clusters by their size for approximating the overall salience of information.

3.5 Cluster Fusion

For each cluster formed, we next fuse together all of its contained propositions to generate a new coherent sentence. As previously mentioned, doing

Cluster A

- The agreement will make Hun Sen prime minister and Ranariddh president of the National Assembly.
- ...to a coalition deal...will make Hun Sen sole prime minister and Ranariddh president of the National Assembly.
- The deal, which will make Hun Sen prime minister and Ranariddh president of the National Assembly...ended more than three months of political deadlock
- Last week...Hun Sen's Cambodian People's Party and Ranariddh's FUNCINPEC party agreed to form a coalition that would leave Hun Sen as sole prime minister and make the prince president of the National Assembly.
- In a long-elusive compromise...opposition leader Prince Norodom Ranariddh will become president of the National Assembly

Cluster B

- ...opposition party leaders Prince Norodom Ranariddh and Sam Rainsy are out of the country
- Sam Rainsy and his then-ally Prince Norodom Ranariddh led an exodus of opposition lawmakers out of Cambodia
- Opposition leaders Prince Norodom Ranariddh and Sam Rainsy...said they could not negotiate freely in Cambodia
- Opposition leaders Prince Norodom Ranariddh and Sam Rainsy...citing Hun Sen's threats

Cluster C

- Hun Sen's Cambodian People's Party narrowly won the polls
- Hun Sen's ruling party narrowly won a majority in elections in July
- Hun Sen's Cambodian People's Party narrowly won the election.
- the ruling party narrowly won.

Cluster D

- A series of negotiations to forge a new government
- ...any...in deadlocked negotiations to form a government.
- A series of negotiations to forge a new government have failed.

Cluster E

- *Hun Sen accused him* of being behind a plot against his life.
- Sam Rainsy...to take refuge in a U.N. office in September to avoid arrest after Hun Sen accused him of
- Sam Rainsy...to avoid arrest after Hun Sen accused him of being behind a plot against his life.

Cluster F

- Hun Sen ousted Ranariddh in a coup.
- The men served as co-prime ministers until Hun Sen overthrew Ranariddh in a coup last year.
- Hun Sen overthrew Ranariddh in a coup last year.

ClusterProp summary

A. The deal will make Hun Sen prime minister and Ranariddh president of the National Assembly

B. The opposition party leaders Prince Norodom Ranariddh and Sam Rainsy are out of the country

C. Hun Sen's Cambodian People's Party narrowly won the election.

D. A series of negotiations to forge a new government failed.

E. *The U.N. accused him* of being behind a plot against his life.

F. Hun Sen ousted Ranariddh in a coup last year.

G. The opposition alleging widespread fraud and intimidation by the CPP

H. The parties have refused to enter into a coalition with Hun Sen until their allegations of election fraud have been thoroughly investigated.

Reference Summary

- Cambodia King Norodom Sihanouk praised formation of a coalition of the Countries top two political parties, leaving strongman Hun Sen as Prime Minister and opposition leader Prince Norodom Ranariddh president of the National Assembly.
- The announcement comes after months of bitter argument following the failure of any party to attain the required quota to form a government.
- Opposition leader Sam Rainey was seeking assurances that he and his party members would not be arrested if they return to Cambodia.
- Rainey had been accused by Hun Sen of being behind an assassination attempt against him during massive street demonstrations in September.

Table 2: The clusters and generated system summary for DUC 2004, topic D30001. Each summary sentence (lower left box) was fused from its corresponding cluster (top boxes). An example of an unfaithful abstraction is marked in *red*.

so avoids redundancy, as one sentence represents several repeating propositional paraphrases.

We fine-tuned a BART generation model (Lewis et al., 2020) to generate a proposition that consolidates the information represented in all propositions of a certain cluster. As input, the model receives cluster propositions, ordered by their predicted salience model score and separated with special tokens.

For training data, we adopt the SuperPAL model (Ernst et al., 2021), that was also separately employed in §3.3. This time, the model is used for measuring the similarity between each of the cluster propositions (that were extracted from the documents) and each of the propositions extracted from the reference summaries. The reference summary proposition with the highest average similarity score to all cluster propositions was selected as the aligned summary proposition of the cluster.

Cluster Feature	DUC 2004		TAC 2011	
	R1	R2	R1	R2
Avg. inner ROUGE	35.9	7.48	38.14	9.93
Avg. salience	35.5	7.98	41.18	12.55
Min. position	37.25	8.89	38.86	11.37
Avg. inner SuperPAL	37.41	8.90	41.22	12.59
Cluster size	37.58	9.01	41.35	12.49
Cluster size + Avg. inner SuperPAL	37.54	8.96	41.45	12.71
Cluster size + Avg. salience	37.77	9.09	41.44	12.62
Cluster size + Min. position	**38.05**	**9.21**	**41.68**	**12.78**

Table 3: ROUGE F1 results on validation sets when ranking clusters according to differing features (DUC 2004 is the validation set of TAC 2011 and vice versa). Two combined features means ranking on the first feature, and breaking ties with the second feature. In all options, a further ranking tie between clusters is resolved according to the maximal proposition salience score of each cluster.

This summary proposition was used as the target output for training the generation model.

The final bullet-style summary is produced by appending generated sentences from the ranked clusters until the desired word-limit is reached.

4 Experimental Setup

4.1 Datasets

We train and test our summarizer with DUC and TAC multi-document summarization benchmarks. Specifically, following standard convention (Mao et al., 2020; Cho et al., 2019), we test on DUC 2004 using DUC 2003 for training, and TAC 2011 using TAC 2008/2009/2010 for training. These sets contain between 30 and 50 document sets each. For validation sets, we used DUC 2004 for the TAC benchmark and TAC 2011 for the DUC benchmark.

4.2 Automatic Evaluation

Following common practice, we evaluate and compare our summarization system with ROUGE-1/2/SU4 F1 measures (Lin, 2004). Stopwords are not removed, and the output summary is limited to 100 words.[4] Note that methods evaluated with ROUGE recall (instead of F1) or limited to 665 bytes (instead of 100 tokens) are not directly comparable to our approach.

4.3 Compared Methods

We compare our method to several strong *extractive* baselines: *SumBasic* (Vanderwende et al., 2007) extracts phrases with words that appear frequently in the documents; *KLSumm* (Haghighi and Vanderwende, 2009) extracts sentences that optimize KL-divergence; *LexRank* (Erkan and Radev, 2004) is a graph-based approach where vertices represent sentences, the edges stand for word overlap between sentences, and sentence importance is computed by eigenvector centrality; *DPP-Caps-Comb* (Cho et al., 2019) balances between salient sentence extraction and redundancy avoidance by optimizing determinantal point processes (DPP); *HL-XLNetSegs* and *HL-TreeSegs* (Cho et al., 2020) are two versions of a DPP-based *span* highlighting approach that heuristically extracts candidate spans by their probability to begin and end with an EOS token; RL-MMR (Mao et al., 2020) adapts a reinforcement learning single document summarization (SDS) approach (Chen and Bansal, 2018) to the multi-document setup and integrates Maximal Margin Relevance (MMR) to avoid redundancy.

We additionally compare to some *abstractive* baselines: Opinosis (Ganesan et al., 2010) generates abstracts from salient paths in a word co-occurrence graph; Extract+Rewrite (Song et al.,

[4]ROUGE parameters: -c 95 -2 4 -U -r 1000 -n 4 -w 1.2 -a -l 100 -m.

	method	R1	R2	RSU4
abstractive	Opinosis	25.15	5.12	8.12
	Extract+Rewrite	29.07	6.11	9.20
	PG	31.44	6.40	10.20
	PG-MMR	37.17	10.72	14.16
	ClusterProp$_{abs}$	**41.45**	**12.75**	**16.16**
extractive	SumBasic	31.58	6.06	10.06
	KLSumm	31.23	7.07	10.56
	LexRank	33.10	7.50	11.13
	HL-XLNetSegs[5]	37.32	10.24	13.54
	HL-TreeSegs[5]	36.70	9.68	13.14
	DPP-Caps-Comb[5]	38.14	11.18	14.41
	RL-MMR	39.65	11.44	15.02
	ClusterProp$_{ext}$	**40.98**	**12.4**	**15.77**
	Oracle$_{prop}$	49.65	21.82	23.19

Table 4: Automatic ROUGE F1 evaluation scores on the TAC 2011 MDS test set. Our solutions (ClusterProp) improve over the previous state-of-the-art methods both in the abstractive and extractive settings. Notably, our *abstractive* approach also surpasses the best *extractive* ones.

	method	R1	R2	RSU4
abstractive	Opinosis	27.07	5.03	8.63
	Extract+Rewrite	28.9	5.33	8.76
	PG	31.43	6.03	10.01
	PG-MMR	36.88	8.73	12.64
	MDS-Joint-SDS	37.24	8.60	12.67
	ClusterProp$_{abs}$	**38.71**	**9.62**	**14.07**
extractive	SumBasic	29.48	4.25	8.64
	KLSumm	31.04	6.03	10.23
	LexRank	34.44	7.11	11.19
	HL-XLNetSegs[5]	36.73	9.10	12.63
	HL-TreeSegs[5]	38.29	**10.04**	13.57
	DPP-Caps-Comb[5]	38.26	9.76	13.64
	RL-MMR	38.56	10.02	13.80
	ClusterProp$_{ext}$	**38.73**	9.64	**13.89**
	Oracle$_{prop}$	46.49	16.16	18.76

Table 5: Automatic ROUGE F1 evaluation scores on the DUC 2004 MDS test set. Our solutions (ClusterProp) improve over the previous state-of-the-art methods both in the abstractive and extractive settings.

[5]The outputs of DPP-Caps-Comb (Cho et al., 2019), HL-XLNetSegs and HL-TreeSegs (Cho et al., 2020) were re-evaluated using author released output.

2018) selects sentences using LexRank and generates for each sentence a title-like summary; PG (See et al., 2017) runs a Pointer-Generator model that includes a sequence-to-sequence network with a copy-mechanism; PG-MMR (Lebanoff et al., 2018) selects representative sentences with MMR and fuses them with a PG-based model; MDS-Joint-SDS (Jin and Wan, 2020) is a hierarchical encoder-decoder architecture that is trained with SDS and MDS datasets while preserving document boundaries.[6]

5 Results

5.1 Automatic Evaluation

As seen in Tables 4 and 5, our model, denoted $\text{ClusterProp}_{\text{abs}}$, surpasses all abstractive baselines by a large margin in all measures both on TAC 2011 and DUC 2004 datasets. In addition, while abstractive system scores are ordinarily inferior to extractive system scores, $\text{ClusterProp}_{\text{abs}}$ notably outperforms all extractive baselines in both benchmarks. Indeed, some of the *extractive* approaches (HL-TreeSegs, DPP-Caps-Comb and RL-MMR) show good performance on DUC 2004 compared to our approach, but they used the large CNN/DailyMail dataset for training while we avoid external sources. Overall, our $\text{ClusterProp}_{\text{abs}}$ provides the new *abstractive* MDS state of the art score in this setting.

On grounds of the effectiveness of $\text{ClusterProp}_{\text{abs}}$ in both the abstractive and extractive settings, we implemented an analogous extractive version, $\text{ClusterProp}_{\text{ext}}$. In this version, for each cluster we extracted the proposition with the highest lexical overlap with the cluster's fused proposition (that was utilized for $\text{ClusterProp}_{\text{abs}}$). As expected, $\text{ClusterProp}_{\text{ext}}$ achieves similar scores, making it the new *extractive* MDS state of the art solution. For reference, we also present the proposition-based extractive upperbound for each dataset ($\textit{Oracle}_{\textit{prop}}$), where document propositions were selected greedily to maximize ROUGE with respect to the reference summaries.

5.2 Ablation Tests

To better apprehend the contribution of each of the steps in our pipeline, Table 6 presents results of the system when applying partial pipelines.

[6]For the MDS-Joint-SDS approach we present only DUC 2004 scores since neither TAC 2011 scores nor code are available for it.

	method	R1	R2	RSU4
TAC 2011	$\text{Oracle}_{\text{sent}}$	47.53	19.83	22.10
	$\text{Oracle}_{\text{prop}}$	49.65	21.82	23.19
	$\text{Oracle}_{\text{cluster-represent}}$	43.40	14.61	17.46
	$\text{Oracle}_{\text{ranking}}$	46.38	17.59	19.88
	$\text{Salience}_{\text{sent}}$	37.32	9.59	13.40
	$\text{Salience}_{\text{prop}}$	39.92	11.53	15.12
	$\text{Salience}_{\text{prop}}$ + Clustering	41.05	12.40	15.73
	$\text{ClusterProp}_{\text{abs}}$	41.45	12.75	16.16
DUC 2004	$\text{Oracle}_{\text{sent}}$	43.91	14.50	17.39
	$\text{Oracle}_{\text{prop}}$	46.49	16.16	18.76
	$\text{Oracle}_{\text{cluster-represent}}$	39.74	10.76	14.56
	$\text{Oracle}_{\text{ranking}}$	43.70	12.92	16.43
	$\text{Salience}_{\text{sent}}$	37.38	9.09	12.90
	$\text{Salience}_{\text{prop}}$	37.73	8.97	13.18
	$\text{Salience}_{\text{prop}}$ + Clustering	38.41	9.09	13.56
	$\text{ClusterProp}_{\text{abs}}$	38.71	9.62	14.07

Table 6: Ablation ROUGE F1 scores on TAC 2011 and DUC 2004. Each additional step in our multi-step method improves the output summaries. The Oracle results indicate the potential of our approach. Specifically, the benefit of summarizing on the proposition level is quite evident.

First, $\textit{Salience}_{\textit{prop}}$ generates summaries simply consisting of the highest scoring document propositions, according to the CDLM-based salience model (§3.2). We also trained the salience model on the sentence- rather than the proposition-level, and similarly generated summaries of salient sentences, denoted $\textit{Salience}_{\textit{sent}}$. The significant improvement of $\textit{Salience}_{\textit{prop}}$ over $\textit{Salience}_{\textit{sent}}$ in both datasets reveals the advantage of working on the proposition level for exposing salient information. This observation is also apparent when comparing the proposition-based oracle ($\textit{Oracle}_{\textit{prop}}$) to the sentence-based oracle method ($\textit{Oracle}_{\textit{sent}}$). The results show that proposition-based systems have a higher ROUGE upperbound across the board, supporting its merit for use in summarization.

Next for ablation, we additionally freeze the clustering stage of the original pipeline, i.e., applying $\text{Salience}_{\text{prop}}$ followed by clustering and ranking of clusters (Sections 3.2, 3.3 and 3.4). From each cluster we then select the proposition with the highest salience score, replacing the fusion step. In both datasets, the clustering stage provides added improvement, suggesting its contribution to our pipeline.

To further demonstrate the potential of our clustering-based summarization approach, we also present two additional oracle scores for extractive upperbound analysis. First, we examine the poten-

	System	unigram	bigram	trigram	sent.
TAC	PG-MMR	98.36	94.42	91.97	50.11
	ClusterProp$_{abs}$	99.08	91.40	81.07	24.39
	Ref. Summs.	90.27	53.17	29.66	1.48
DUC	PG-MMR	98.34	94.99	90.91	50.82
	ClusterProp$_{abs}$	98.86	89.72	78.28	23.50
	Ref. Summs.	88.41	44.27	18.65	0.13

Table 7: Percentage of n-gram/sentence overlap between summaries and source documents in TAC 2011 and DUC 2004. Compared to PG-MMR, our system has substantially less sequential overlap, indicating its increased abstractiveness. Reference summaries are naturally highly abstractive.

tial of optimally selecting cluster representatives for the summary. We greedily select a single representative from each cluster (while preserving the original cluster ranking order from §3.4) that optimizes the overall ROUGE score of all selected representatives with respect to the reference summaries (*Oracle*$_{cluster\text{-}represent}$). These results express the additional improvement that a better choice of cluster representatives could produce, i.e., up to ~2 ROUGE-2 points in TAC 2011 and ~1 point in DUC 2004.

Another aspect of our pipeline to examine is the potential of enhanced cluster ranking. To that end, we first selected the highest salience-scoring proposition as a representative from each cluster. Then we greedily selected representatives, one a time, that maximized the overall ROUGE againt the reference summaries. Effectively, this points to the clusters that would be best for producing a summary, which indicates an optimal cluster choice (*Oracle*$_{ranking}$). The potential improvement of better cluster ranking is hence up to ~5 ROUGE-2 points in TAC 2011 and ~3 points in DUC 2004.

Overall, we observe that our multi-step pipeline approach is indeed effective for MDS, and that their is great potential for furthering its success.

5.3 Human Evaluation

We assessed our system, ClusterProp$_{abs}$, with manual comparison against PG-MMR, a strong abstractive MDS baseline. Crowdworkers on Amazon Mechanical Turk[7] were shown the summaries of a topic from the two systems in arbitrary order along with a corresponding reference summary. They were asked to select the preferred system with respect to **Content** (*"Which of the system summaries has higher content overlap with the reference?"*)

[7] https://www.mturk.com

	method	Content	Readability
TAC	PG-MMR	18%	27%
	ClusterProp$_{abs}$	**82%**	**73%**
DUC	PG-MMR	35%	41%
	ClusterProp$_{abs}$	**65%**	**59%**

Table 8: Human preferences of system summaries, with respect to content overlap with reference summaries and overall readability, on TAC 2011 and DUC 2004.

and **Readability** (*"Which of the system summaries is more readable and well-understood?"*). This procedure was repeated for each of the four available reference summaries per topic, and each such triplet was evaluated by three workers. For the final preference choice we first took the majority vote for each triplet, and then summed up all the votes.

Table 8 shows that our summaries were favored in terms of both content and readability by a large margin in both datasets. As this work is aimed to select better salient content, the large gap in favor of ClusterProp$_{abs}$ in the content criterion is not surprising, and is consistent with the ROUGE scores in §5.1.

While our summaries are non-conventionally structured as bullet-style lists of propositions rather than a coherent paragraph, evaluators preferred our style of summarization in terms of readability. Moreover, as Table 7 points out, ClusterProp$_{abs}$ appears to be more abstractive than PG-MMR, as suggested by the reduced n-gram overlap with source documents. Specifically, about half of the system summary sentences of PG-MMR summaries are fully copied. While the intensified abstractiveness of our summaries could have potentially hindered readability, our system was nevertheless preferred. Our approach leaves fertile ground for further improving readability by fusing several clusters together to generate sentences containing multiple information units.

6 Paraphrastic Clusters as Summary Evidence

One of the unique advantages of a cluster-based summary is that each summary sentence is linked to a group of propositions from which the sentence was generated, in so providing an "explanation" for the output. We verified that clusters indeed "explain" their generated sentences, by assessing how many of the propositions within a cluster align with the respective output sentence. To that end, we

conducted a crowdsourced annotation procedure, where a worker marked whether a proposition and its corresponding generated sentence aligns. Each pair was examined by three workers, with the majority vote used for deciding on alignment. On a random selection of 25% of the clusters, we found that, on average, 89%/84% of a cluster's propositions in DUC 2004/TAC 2011 support their corresponding generated sentence, with an average cluster size of 3.4/4.8 propositions, respectively.

Given the strong alignment of a cluster to its generated sentence, a cluster facilitates effective verification of faithfulness of its corresponding generated abstractive sentence. Since the output sentence is based solely on its cluster propositions, the sentence's correctness can be verified against the cluster instead of the full document set. An example of an unfaithful abstraction is marked in red in Table 2. To the best of our knowledge, ours is the first attempt for efficient assessment of faithfulness in MDS. We conducted a respective evaluation process, through crowdsourcing, to assess the faithfulness of our system summaries. A worker saw a cluster and its generated sentence and marked whether hallucinations were evident in the sentence. Over the full test sets, the annotations showed that 80% and 90% of the DUC 2004 and TAC 2011 summary sentences, respectively, were faithful to their corresponding clusters.

The cluster explanations can additionally be leveraged for various downstream purposes. For example, a cluster of propositions that is linked to a summary sentence can provide complementary facts regarding the information unit. Such a feature can be incorporated in interactive summarization systems, as applied in (Shapira et al., 2017) where a user can choose to expand on the facts within a sentence of the presented summary.

7 Conclusion

We advocate the potential of proposition-level units as a cleaner and more accurate unit for summarization content selection. To that end, we present a new proposition-level pipeline for summarization that includes an accurate paraphrastic propositional clustering component followed by fusion of cluster propositions, to generate concise and coherent summary sentences. Our proposed method outperforms state-of-the-art baselines (in this setting) in both automatic and human evaluation on the DUC and TAC MDS benchmarks. We provide an ablation study that indicates the benefit of each of the steps in the pipeline, as well as the potential for further future improvement of the overall approach. Moreover, we demonstrate the utility of the clustering-based approach for validation of summary faithfulness.

References

Kristjan Arumae, Parminder Bhatia, and Fei Liu. 2019. Towards annotating and creating summary highlights at sub-sentence level. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 64–69, Hong Kong, China. Association for Computational Linguistics.

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1587–1597, Beijing, China. Association for Computational Linguistics.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

X. Cai and Wenjie Li. 2013. Ranking through clustering: An integrated approach to multi-document summarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 21:1424–1433.

Xiaoyan Cai, Wenjie Li, You Ouyang, and Hong Yan. 2010. Simultaneous ranking and clustering of sentences: A reinforcement approach to multi-document summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 134–142, Beijing, China. Coling 2010 Organizing Committee.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1027–1038, Florence, Italy. Association for Computational Linguistics.

Sangwoo Cho, Kaiqiang Song, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2020. Better highlighting: Creating sub-sentence summary highlights. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6282–6300, Online. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.

Tanvir Ahmed Fuad, Mir Tafseer Nayeem, Asif Mahmud, and Yllias Chali. 2019. Neural sentence fusion for diversity driven abstractive multi-document summarization. *Comput. Speech Lang.*, 58:216–230.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

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

Hanqi Jin and Xiaojun Wan. 2020. Abstractive multi-document summarization via joint learning with single-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2545–2554, Online. Association for Computational Linguistics.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Walter Chang, and Fei Liu. 2020. A cascade approach to neural abstractive summarization with content selection and fusion. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 529–535, Suzhou, China. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920, Los Angeles, California. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020. Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1737–1751, Online. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manag.*, 40:919–938.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ori Shapira, Hadar Ronen, Meni Adler, Yael Amsterdamer, Judit Bar-Ilan, and Ido Dagan. 2017. Interactive abstractive summarization for event news tweets. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 109–114, Copenhagen, Denmark. Association for Computational Linguistics.

Kaiqiang Song, Lin Zhao, and Fei Liu. 2018. Structure-infused copy mechanisms for abstractive summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1717–1729, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618. Text Summarization.

Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In *SIGIR '08*.

Dingding Wang, Tao Li, Shenghuo Zhu, and C. Ding. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *SIGIR '08*.

Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, and Yihong Gong. 2011. Integrating document clustering and multidocument summarization. *ACM Trans. Knowl. Discov. Data*, 5:14:1–14:26.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244.

Yang Zhang, Yunqing Xia, Yi Liu, and Wenmin Wang. 2015. Clustering sentences with density peaks for multi-document summarization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1262–1267, Denver, Colorado. Association for Computational Linguistics.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. SummPip: Unsupervised multi-document summarization with sentence graph compression. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.