Proposition-Level Clustering for Multi-Document Summarization

Anonymous ACL submission

Abstract

Text clustering methods were traditionally incorporated into multi-document summarization (MDS) as a means for coping with considerable information repetition. Particularly, clusters were leveraged to indicate information saliency as well as to avoid redundancy. Such prior methods focused on clustering sentences, even though closely related sentences usually contain also non-aligned parts. In this work, we revisit the clustering approach, grouping together sub-sentential propositions, aiming at more precise information alignment. Specifically, our method detects salient propositions, clusters them into paraphrastic clusters, and generates a representative sentence for each cluster via text fusion. 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.

An appealing summarization approach that copes with these challenges, and is especially relevant for MDS, is clustering-based summarization.

1Our code and system summaries will be release upon publication.

Figure 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. The first sentence is illustratively divided into propositions, where only one of them is aligned to those in the other sentences.

In such an approach, the goal is to cluster redundant paraphrastic pieces of information across the texts, which roughly convey the same meaning. Repetition of information across texts, as captured by paraphrastic clustering, typically indicates its importance, and can be leveraged for salience detection. Moreover, representing a paraphrastic cluster may facilitate generating a corresponding summary that eliminates repetitions while fusing together complementary details within the cluster.

Traditionally, clustering-based approaches were widely used for summarization, mostly in extractive and unsupervised settings (Radev et al., 2004; Zhang et al., 2015; Nayeem et al., 2018). Notably, most of these works generated sentence-based clusters, which 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 paraphrases. Figure 1 exemplifies such a noisy cluster, which does contain paraphrastic propositions (marked in blue) within their full sentences (marked in black). Another line of research in summarization coped with such noisy sentence-based setting, and looked into
the use of sub-sentential units for summarization, e.g., Li et al. (2016) summarizes with Elementary Discourse Units (EDUs), while Ernst et al. (2021) endorse using OpenIE-based propositions (Stanovský et al., 2018) for summarization.

In this paper, we revisit and combine the clustering-based approaches along with sub-sentential setting, two research lines that were explored only individually and rather scarcely in recent years. Specifically, we apply clustering-based summarization at the more fine-grained propositional level, which avoids grouping non-aligned texts, yielding accurate paraphrastic clusters. These clusters also provide better control over the generated summary sentences – as the generation component is only required to fuse similar propositions.

Our model (§3) leverages gold reference summaries to derive training datasets for several summarization sub-tasks. First, salient document propositions were extracted, to train a salience model, by greedily maximizing alignment with the reference summaries. Then, an available proposition similarity model, trained from summary-source alignments (Ernst et al., 2021), provides the basis for agglomerative clustering (Ward, 1963). Finally, we created training data for a BART-based model for sentence fusion (Lewis et al., 2020) by aligning reference summary propositions with source proposition clusters. Similar to many other works, we leave inter-sentence coherence and sentence planning and ordering outside the scope of the current paper. Accordingly, our process produces a bullet-style summary of individual concise and coherent sentences.

Overall, our experiments (§4) 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. Finally, we also suggest (§5) that clustering-based methods provide “explanations”, or supporting evidence, for each generated sentence, in the form of the source cluster propositions (see an example in Table 1).

2 Background and Related Work

Clustering-based summarization. Clustering-based summarization approaches typically involve salience detection while avoiding redundancy. 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. More recently though, 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. While most of previous clustering approaches operated at the noisy sentence level, in our work we present more accurate proposition-level clustering that eventually enhances summarization.

Sub-sentence units in summarization. While many summarization approaches extract full document sentences, either for extractive summarization or as an intermediate step for abstractive summarization, there are methods that operated 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) (Mann and Thompson, 1988). 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 surrounding context. Recently, Cho et al. (2020) proposed identifying heuristically self-contained sub-sentence units for the highlighting task.

Abstractive approaches have been extracting sub-sentence units as a preliminary step for generation. Such units range from words (Lebanoff et al., 2020; Gehrmann et al., 2018), to noun or verb phrases (Bing et al., 2015), to Open Information Extraction (OpenIE) propositions (Pasunuru et al., 2021). In our work, we follow the same extract-then-generate pipeline, using OpenIE spans (Stanovský et al., 2018) as proposition units. Since propositions are meant to contain single standalone facts consisting of a main predicate and its arguments, they are beneficial for grouping mostly overlapping para-
Figure 2: Our multi-document summarization process. (a) All propositions are extracted (OpenIE; Stanovsky et al., 2018) from the documents. (b) Propositions are classified by a salience score (fine-tuned CDLM; Caciularu et al., 2021). (c) Salient propositions are clustered (fine-tuned SuperPAL; Ernst et al., 2021), forming groups of paraphrastic information units. (d) Clusters are ranked, as an indicator for information importance. (e) For each cluster, its propositions are fused (fine-tuned BART; Lewis et al., 2020) to generate a concise and coherent abstractive sentence. (f) The output summary is obtained as a bullet-style ranked list of the concise sentences.

phrases (unlike sentential paraphrases). In addition, propositions extracted with OpenIE can be noncontiguous, while alternative options, like EDUs, are limited to contiguous sequences.

3 Method

This section first provides an overview of our method, followed by subsections describing its components. We follow previous clustering-based approaches, where text segments are first clustered into semantically similar groups, exploiting redundancy as a salience signal. Then, each group is fused to generate a merged sentence, while avoiding redundancy. As we operate at the proposition-level, we first extract all propositions from the input documents (§3.1). Then, to facilitate the clustering step, we filter out non-salient propositions using a salience model (§3.2). Next, salient propositions are clustered based on their semantic similarity (§3.3). The largest clusters, whose information was most repeated, are selected to be included in the summary (§3.4). Finally, each cluster is fused to form a sentence for a bullet-style abstractive summary (§3.5). In addition, to support extractive summarization, we provide an extractive version where a representative (source) proposition is selected from each cluster (§3.6). Overall, clustering explicit propositions induces a multi-step process that requires dedicated training data for certain steps. To that end, we derive new training datasets for the salience detection and the fusion models from the original gold summaries. The full pipeline is illustrated in Figure 2, where additional implementation details appear in §B in the Appendix.

3.1 Proposition Extraction

Aiming to generate proposition-based summaries, we first extract all propositions from the source documents using Open Information Extraction (OpenIE) (Stanovsky et al., 2018)\(^2\), following Ernst et al. (2021). To convert an OpenIE tuple containing a predicate and its arguments into a proposition string, we simply concatenate them by their original order, as illustrated in Figure 3 in the Appendix.

3.2 Proposition Salience Model

To facilitate the clustering stage, we first aim to filter non-salient propositions by a supervised model. To that end, we derive gold labels for proposition salience from the existing reference summaries. Specifically, we select greedily propositions that maximize ROUGE-1\(_F\)-1 + ROUGE-2\(_F\)-1 against their reference summaries (Nallapati et al., 2017; Liu and Lapata, 2019) and marked them as salient. Using this derived training data, we fine-tuned the Cross-Document Language Model (CDLM) (Caciularu et al., 2021) as a binary classifier for

\(^2\)https://demo.allenai.org/open-information-extraction
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. All propositions contained in the clusters in Table 1 are examples of predicted salient propositions. We chose to use CDLM as it was 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).

### 3.3 Clustering

Next, all salient propositions are clustered to semantically similar groups. Clusters of paraphrastic propositions are advantageous for summarization as they can assist in avoiding redundant information in an output summary. Furthermore, paraphrastic clustering offers redundancy as an additional indicator for saliency, while the former salience model (§3.2) does not utilize repetitions explicitly. To cluster propositions we utilize SuperPAL (Ernst et al., 2021), a binary classifier that measures paraphrastic similarity between two propositions. All pairs of salient propositions are scored with SuperPAL, over which standard agglomerative clustering (Ward, 1963) is applied. Examples of generated clusters are presented in Table 1.

#### 3.4 Cluster Ranking

The resulting proposition clusters are next ranked according to cluster-based properties. We examined various features, listed in Table 2, on our validation sets. The features examined include: average of ROUGE scores between all propositions in a cluster (‘Avg. ROUGE’), average of SuperPAL scores between all propositions in a cluster (‘Avg. SuperPAL’), entropy of the cluster according to cluster-based properties, and cluster size. The features were rank ordered according to their performance on the validation set. The features are applied with a cluster size lower than 10. The resulting ranking of the clusters is shown in Table 2.

Table 1: The proposition clusters and system and reference summaries for DUC 2004, topic D30001. Each summary sentence (lower left box) was fused from its corresponding cluster (top boxes) that also provides supporting source evidence. An example of an unfaithful abstraction is marked in red.
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’).

For each feature, (1) clusters were ranked according to the feature, (2) the proposition with the highest salience model score (§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. 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. In all options, if a tie is still remained, further ranking between clusters is resolved according to the maximal proposition salience score of each cluster. The resulting ROUGE scores of these summaries on validation sets are presented in Table 2.3 We found that ‘Cluster size’ yields the best ROUGE scores as a single feature, and ‘Min. position’ further improves results as a secondary tie breaking ranking feature. Intuitively, a large cluster represents redundancy of information across documents thus likely to indicate higher importance.

3.5 Cluster Fusion

Next, we would like to merge the paraphrastic propositions in each cluster, while consolidating complementary details, to generate a new coherent summary sentence. As mentioned, this approach helps avoiding redundancy, since redundant information is concentrated separately in each cluster.

To train a cluster fusion model, we derived training data automatically from the reference summaries, by leveraging the SuperPAL model (Ernst et al., 2021) (which was also 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. This summary proposition was used as the target output for training the generation model. Although these target OpenIE propositions may be ungrammatical or non-fluent, a human examination has shown that BART tends to produce full coherent sentences (mostly containing only a single proposition), even though it was finetuned over OpenIE extractions as target. Examples of coherent generated sentences can be seen in Table 1.

Accordingly, we fine-tuned a BART generation model (Lewis et al., 2020) with this dedicated training data. As input, the model receives cluster propositions, ordered by their predicted salience score (§3.2) and separated with special tokens. The final bullet-style summary is produced by appending generated sentences from the ranked clusters until the desired word-limit is reached.

3.6 Extractive Summarization Version

To support extractive summarization settings, for example when hallucination is forbidden, we created a corresponding extractive version of our method. In this version, we extracted a representative proposition for each cluster, which was chosen according to the highest word overlap with the sentence that was fused from this cluster by our abstractive version.

4 Evaluation

4.1 Experimental Setup

Datasets. We train and test our summarizer with the challenging DUC and TAC MDS benchmarks. Specifically, following standard convention (Mao et al., 2020; Cho et al., 2019), we test on DUC 2004 using DUC 2003 for training, and on TAC 2011 using TAC 2003 for training.

Table 2: 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.

<table>
<thead>
<tr>
<th>Cluster Feature</th>
<th>DUC 2004</th>
<th>TAC 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
</tr>
<tr>
<td>Avg. ROUGE</td>
<td>35.9</td>
<td>7.48</td>
</tr>
<tr>
<td>Avg. salience</td>
<td>35.5</td>
<td>7.98</td>
</tr>
<tr>
<td>Min. position</td>
<td>37.25</td>
<td>8.89</td>
</tr>
<tr>
<td>Avg. SuperPAL</td>
<td>37.41</td>
<td>8.90</td>
</tr>
<tr>
<td>Cluster size</td>
<td>37.58</td>
<td>9.01</td>
</tr>
<tr>
<td>Cluster size + Avg. SuperPAL</td>
<td>37.54</td>
<td>8.96</td>
</tr>
<tr>
<td>Cluster size + Avg. salience</td>
<td>37.77</td>
<td>9.09</td>
</tr>
<tr>
<td>Cluster size + Min. position</td>
<td>38.05</td>
<td>9.21</td>
</tr>
</tbody>
</table>

3We 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.

4For the Hi-MAP and MDS-Joint-SDS approaches we present only DUC 2004 scores since TAC 2011 scores are not available for them.

5The outputs of DPP-Caps (Cho et al., 2019), HL-XLNet and HL-Tree (Cho et al., 2020) were re-evaluated using author released output.
Table 3: Automatic ROUGE F1 evaluation scores on the TAC 2011 & DUC 2004 MDS test sets. Our solutions (ProCluster) 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>TAC 2011</th>
<th>DUC 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
</tr>
<tr>
<td><strong>abstractive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opinosis (Ganesan et al., 2010)</td>
<td>25.15</td>
<td>5.12</td>
</tr>
<tr>
<td>Extract+Rewrite (Song et al., 2018)</td>
<td>29.07</td>
<td>6.11</td>
</tr>
<tr>
<td>PG (See et al., 2017)</td>
<td>31.44</td>
<td>6.40</td>
</tr>
<tr>
<td>Hi-MAP⁴ (Fabbri et al., 2019)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PG-MMN (Lebanoff et al., 2018)</td>
<td>37.17</td>
<td>10.72</td>
</tr>
<tr>
<td>MDS-Joint-SDS² (Jin and Wan, 2020)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ProCluster</td>
<td><strong>41.45</strong></td>
<td><strong>12.75</strong></td>
</tr>
<tr>
<td><strong>extractive</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SumBasic (Vanderwende et al., 2007)</td>
<td>31.58</td>
<td>6.06</td>
</tr>
<tr>
<td>KLSumm (Haghighi and Vanderwende, 2009)</td>
<td>31.23</td>
<td>7.07</td>
</tr>
<tr>
<td>LexRank (Erkan and Radev, 2004)</td>
<td>33.10</td>
<td>7.50</td>
</tr>
<tr>
<td>HL-XLNetSegs⁵ (Cho et al., 2020)</td>
<td>37.32</td>
<td>10.24</td>
</tr>
<tr>
<td>HL-TreeSegs⁵ (Cho et al., 2020)</td>
<td>36.70</td>
<td>9.68</td>
</tr>
<tr>
<td>DPP-Caps-Comb⁵ (Cho et al., 2019)</td>
<td>38.14</td>
<td>11.18</td>
</tr>
<tr>
<td>RL-MMR (Mao et al., 2020)</td>
<td>39.65</td>
<td>11.44</td>
</tr>
<tr>
<td>ProClusterext</td>
<td><strong>40.98</strong></td>
<td><strong>12.40</strong></td>
</tr>
<tr>
<td><strong>Oracleprop</strong></td>
<td>49.65</td>
<td>21.82</td>
</tr>
</tbody>
</table>

Table 3: Automatic ROUGE F1 evaluation scores on the TAC 2011 & DUC 2004 MDS test sets. Our solutions (ProCluster) 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.

4.2 Automatic Evaluation

As seen in Table 3, our abstractive model, denoted ProClusterabs for Propositional Clustering, surpasses all abstractive baselines by a large margin in all measures on both TAC 2011 and DUC 2004. Moreover, while the abstractive system scores were typically inferior to extractive system scores, ProClusterabs notably outperforms all extractive baselines in both benchmarks. Overall, our ProClusterabs provides the new *abstractive* MDS state-of-the-art score in this setting.

As said in §3.6, we also developed an extractive version, denoted ProClusterext. As ProClusterext selects document propositions that have the highest overlap with ProClusterabs sentences, ProClusterext achieves similar scores to ProClusterabs, yielding the new extractive MDS state-of-the-art results. For comparison we selected strong baseline, including previous state-of-the-art in this setup, in both the extractive and abstractive settings. See in Appendix §C for more concise details over each baseline. For reference, we also present a proposition-based extractive upperbound for each dataset (Oracleprop), where document propositions were selected greedily to maximize ROUGE-1F₁ + ROUGE-2F₁ with respect to the reference summaries.

4.3 Ablation Analysis

To better apprehend the contribution of each of the steps in our pipeline, Table 4 presents results of the system when applying partial pipelines.

First, Salienceprop 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 Salience_sent. The notable improvement of Salienceprop over Salience_sent in both datasets reveals the advantage of working at the proposition level for exposing salient information. This observation is also apparent when comparing the proposition-based oracle (Oracleprop) to the sentence-based oracle method (Oracle_sent). The results indicate that proposition-based systems have a higher ROUGE upperbound across the board, supporting its merit for use in summarization.

---

⁴ROUGE parameters: -c 95 -2 4 -U -r 1000 -n 4 -w 1.2 -a -1 100 -m.

⁵Note 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.
Next, we would like to assess the contribution of the clustering step. Therefore, we applied Salience\textsubscript{prop} followed by clustering and ranking of clusters (Sections 3.2, 3.3 and 3.4), while leaving the fusion step aside. From each cluster we then select the proposition with the highest salience score to be in the system summary. In both datasets, the clustering stage provides added improvement, suggesting its contribution to our pipeline.

To further demonstrate the potential of our approach, we also present two additional oracle scores for extractive upperbound analysis. First, we examine the potential of optimally selecting cluster representatives for the summary. We greedily select a single representative per cluster following the original cluster ranking (§3.4) that optimizes the overall ROUGE-1\textsubscript{F-1} + ROUGE-2\textsubscript{F-1} score of all selected representatives with respect to the reference summaries (Oracle\textsubscript{cluster-rep}). These results express the improvement comparing to our final model (ProCluster\textsubscript{abs}), that a better cluster representative choice could produce, i.e., up to ~2 R-2 points in TAC 2011 and ~3 points in DUC 2004.

Another aspect 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 at a time, that maximized the overall ROUGE-1\textsubscript{F-1} + ROUGE-2\textsubscript{F-1} against the reference summaries. Effectively, this points to a greedily optimized cluster choice (Oracle\textsubscript{ranking}). The potential improvement of better cluster ranking compared to our final model (ProCluster\textsubscript{abs}) is hence up to ~5 R-2 points in TAC 2011 and ~3 points in DUC 2004. Indeed, our approach leaves cluster ranking improvement to future work.

Overall, we observe that all components of our multi-step approach are indeed effective for MDS, and that there is a great potential for further improvements within this architecture.

### 4.4 Human Evaluation

We further assessed our primary system, ProCluster\textsubscript{abs}, through manual comparison against PG-MMR, a strong abstractive MDS baseline. Crowworkers on Amazon Mechanical Turk\footnote{https://www.mturk.com} were shown the summaries of a given 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?") 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 choice we first took the majority vote for each triplet, and then summed up all the votes.

Table 5 shows that our summaries were favored in terms of both content and readability by a large margin in both datasets. As our work is focused on selecting better salient content, the large gap in favor of ProCluster\textsubscript{abs} in the content criterion supports the advantage of our approach, and is consistent with the ROUGE scores in §4.2.

While our summaries are (somewhat 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 6 points out, ProCluster\textsubscript{abs} appears to be more abstractive than PG-MMR, as suggested by the reduced n-gram and sentence overlap with source documents. Specifically, about half of the system summary sentences of PG-MMR are fully copied, compared to about a quarter in our method. While the intensified abstractiveness of our summaries could have
potentially hindered readability, our system was nevertheless preferred along this aspect as well.

Our approach leaves fertile ground for further improving readability by fusing several clusters together to generate sentences containing multiple propositions, and by developing sentence planning and ordering models. Compatible training datasets for these models can be derived out of the gold reference summaries, as was done in this work for the salience (§3.2) and fusion (§3.5) models.

<table>
<thead>
<tr>
<th>method</th>
<th>Content</th>
<th>Readability</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PG-MMR</td>
<td>18%</td>
<td>27%</td>
</tr>
<tr>
<td>ProCluster</td>
<td>82%</td>
<td>73%</td>
</tr>
<tr>
<td>DUC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PG-MMR</td>
<td>35%</td>
<td>41%</td>
</tr>
<tr>
<td>ProCluster</td>
<td>65%</td>
<td>59%</td>
</tr>
</tbody>
</table>

Table 6: 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.

5 Paraphrastic Clusters as Summary Evidence

A unique advantage of a cluster-based summary is that each summary sentence is linked explicitly to a group of propositions from which the sentence was generated, in so providing an “explanation”, or support evidence, for the output. These cluster explanations can expand the reader’s knowledge and provide complementary facts from the nearby source context regarding the information from the generated sentence. Such a feature may 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.

To assess the reliability of such feature, we verified that clusters indeed “explain” their generated sentences. To that end, we conducted a crowdsourced annotation, where a worker marked whether a cluster proposition mentions the main idea of its corresponding generated sentence. Each pair was examined by three workers, with the majority vote used for the final decision. On a random selection of 25% of the clusters, we found that, on average, 89% and 84% of a cluster’s propositions in DUC 2004 and TAC 2011 support their corresponding generated sentence, with an average cluster size of 3.4 and 4.8 propositions, respectively.

Furthermore, given this 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 “explaining” cluster instead of against the full document set. An example of an unfaithful abstraction is marked in red in Table 1. To the best of our knowledge, this is the first attempt for efficient manual 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 the sentence was faithful to its origin cluster or not. Overall, this task cost a reasonable price of 240$ for both the DUC 2004 and TAC 2011 datasets together. 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.

6 Conclusion

We advocate the potential of proposition-level units as a cleaner and more accurate unit for summarization. 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 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 pipeline steps, as well as the potential for future improvement. Moreover, we demonstrate the utility of the clustering-based approach for providing source documents explanations and for manually validating summary faithfulness.
References


Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In NIPS.


A Ethical Considerations

Computation. We ran on 3 GPUs for 20 minutes to finetune each of the salience model and the fusion model. The summarization model runs 10 minutes on 4 GPUs to generate a summary. Most of the time is spent on the clustering step, in which we calculate the SuperPAL similarity score between all salient proposition pairs.

Dataset. The DUC 2003 and 2004 and TAC 2008-2011 datasets were acquired according to the required NIST guidelines (duc.nist.gov).

Crowdsourcing. All human annotations and evaluations conducted with crowdsourcing were compensated as a 12S per hour wage. We estimated the task payment by completing sample assignments and obtaining the average assignment time.

B Implementation Details

B.1 Proposition Salience Model

Datasets. For many previous summarization systems these benchmarks were insufficiently large enough for training their models. Consequently, they pretrained on a large scale summarization dataset, such as CNN/DailyMail (Hermann et al., 2015), and then finetuned on DUC/TAC datasets (e.g., Lebanoff et al., 2018; Mao et al., 2020). In our case, we avoid external sources. However, as DUC training data is much smaller than TAC’s (30 topics vs. 138), and it was apparently too small for the salience model training, we adopted the trained salience model for TAC benchmark (that was trained with TAC 2008-2010) as a pretrained model and then finetuned it with DUC 2003. Accordingly, validating the TAC benchmark using DUC 2004 during the salience model training causes data leakage since this model is later finetuned to test on the same DUC 2004. To avoid that, during the salience model training we used part of TAC 2010 that was omitted from training data, as a validation set (instead of DUC 2004).

Training Parameters. We trained the model for 10 epochs with learning rate of 1e-5 and batch size of 6 instances on 3 DGX GPUs (meaning effective batch size was 18).

Training. The 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 infeasible to feed the full document set as a long sequence. 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).

Each instance contains a proposition marked with start and end special tokens, within its multiple document context. A discontinuous proposition is marked with special tokens before and after each of its parts. In addition, sentence special token separators and document special token separators are used, as required for CDLM.

In order to reduce computation complexity, CDLM uses “local attention” (of 512 tokens) for all tokens, while specific tokens are attended to all 4096 tokens (“global attention”). In our case, we assigned global attention to the CLS token and to the candidate proposition tokens, including their special tokens.

For classification, we have added a binary classi-
Albert Einstein published the theory of relativity in 1915 and received the Physics Nobel Prize in 1921.

B.2 SuperPAL Usage

In this work we used the SuperPAL model (Ernst et al., 2021) as the similarity metric between propositions for the clustering step (§3.3), and to create training data for the fusion model (§3.5). Originally, SuperPAL was tuned with a validation set that contains three topics from DUC 2004 (taken from the full validation set which also contains 7 additional topics, not from DUC 2004). In our setting, it may cause leakage since DUC 2004 is used as the test data. To avoid such leakage, we tuned SuperPAL again without using DUC 2004 topics at all (using the other 7 topics as a validation set).

B.3 Fusion Model

Training Parameters. We trained the model for 3 epochs with learning rate of 3e-5 and batch size of 10 instances on 3 DGX GPUs (meaning effective batch size was 30).

C Compared Methods

We compare our method to several strong abstractive baselines: Opinosis (Ganesan et al., 2010) generates abstracts from salient paths in a word co-occurrence graph; Extract+Rewrite (Song et al., 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 (Carbonell and Goldstein, 1998) and fuses them with a PG-based model; Hi-MAP (Fabbri et al., 2019) is a hierarchical version of the PG model that allows calculating sentence-level MMR scores; 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.

We additionally compare 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 neural 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.

### D Annotation Guidelines

We used Amazon Mechanical Turk\(^\text{10}\) for all three crowdsourced tasks with a list of 90 pre-selected workers from English speaking countries. These workers accomplished high quality work in other NLP-related tasks we have conducted in the past.

The crowdsourcing instructions of the tasks mentioned in §5 are as follows: (the crowdsourcing instructions for the manual summarization evaluation are already specified in §4.4)

**Supporting cluster evaluation.** Read the following two text spans, and answer the question below.

**Span Text A:**

*The generated sentence*

**Span Text B:**

*A proposition from the cluster*

Is the main fact of Span Text A mentioned in Span Text B? (ignoring additional details)

Yes/No

**Faithfulness Evaluation.** Read the following group of text spans A and text span B, and answer the questions below. You can assume that all text spans in group A describe the same event, and therefore can be consolidated together to imply Text Span B.

**Examples:**

1. **Group of Text Spans A:**

   - *They arrested John.*
   - *John was arrested.*

   **Text Span B:**

   *The FBI arrested John*

   Is the Group of Text Spans A implies the fact in Text Span B?

2. **Group of Text Spans A:**

   - *there were 10-12 girls and 15 boys in the schoolhouse*
   - *there were boys and girls in the schoolhouse*

   **Text Span B:**

   *there were 1012 girls and 15 boys in the schoolhouse*

   Is the Group of Text Spans A implies the fact in Text Span B?

   Text Span B contradicts Group A (instead of 10-12 girls it says 1012 girls). Therefore the answer is No.

\(^{10}\)https://www.mturk.com