Scoring Sentence Singletons and Pairs for Abstractive Summarization

Logan Lebanoff[†] **Kaiqiang Song**[†] **Franck Dernoncourt**[§]
Doo Soon Kim[§] **Seokhwan Kim**[§] **Walter Chang**[§] **Fei Liu**[†]

[†]Computer Science Department, University of Central Florida, Orlando, FL 32816
{loganlebanoff, kqsong}@knight.ucf.edu feiliu@cs.ucf.edu

[§]Adobe Research, San Jose, CA 95110
{dernonco,dkim,seokim,wachang}@adobe.com

Abstract

When writing a summary, humans tend to choose content from one or two sentences and merge them into a single summary sentence. However, the mechanisms behind the selection of *one* or *multiple* source sentences remain poorly understood. Sentence fusion assumes multi-sentence input; yet sentence selection methods only work with single sentences and not combinations of them. There is thus a crucial gap between sentence selection and fusion to support summarizing by both compressing single sentences and fusing pairs. This paper attempts to bridge the gap by ranking sentence singletons and pairs together in a unified space. Our proposed framework attempts to model human methodology by selecting either a single sentence or a pair of sentences, then compressing or fusing the sentence(s) to produce a summary sentence. We conduct extensive experiments on both single- and multi-document summarization datasets and report findings on sentence selection and abstraction.

1 Introduction

Abstractive summarization aims at presenting the main points of an article in a succinct and coherent manner. To achieve this goal, a proficient editor can rewrite a source sentence into a more succinct form by dropping inessential sentence elements such as prepositional phrases and adjectives. She can also choose to fuse multiple source sentences into one by reorganizing the points in a coherent manner. In fact, it appears to be common practice to summarize by either compressing single sentences or fusing multiple sentences. We investigate this hypothesis by analyzing human-written abstracts contained in three large datasets: DUC-04 (Over and Yen, 2004), CNN/Daily Mail (Hermann et al., 2015), and XSum (Narayan et al., 2018). For every summary sentence, we find its *ground-truth set* containing one or more source sentences that exhibit a high degree of similarity with the summary sentence (details in §4). As shown in Figure 1, across the three datasets, 60-85% of summary sentences are generated by fusing one or two source sentences.

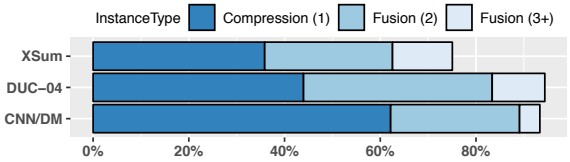

Figure 1: Portions of summary sentences generated by compression (content is drawn from one source sentence) and fusion (content is drawn from two or more source sentences). Humans often grab content from 1 or 2 document sentences when writing a summary sentence.

Selecting summary-worthy sentences has been studied in the literature, but there lacks a mechanism to weigh sentence singletons and pairs in a unified space. Extractive methods focus on selecting sentence singletons using greedy (Carbonell and Goldstein, 1998), optimization-based (Gillick and Favre, 2009; Kulesza and Taskar, 2011; Cho et al., 2019), and (non-)autoregressive methods (Cheng and Lapata, 2016; Kedzie et al., 2018). In contrast, existing sentence fusion studies tend to assume ground sets of source sentences are already provided, and the system fuses each set of sentences into a single one (Daumé III and Marcu, 2004; Filippova, 2010; Thadani and McKeown, 2013). There is thus a crucial gap between sentence selection and fusion to support summarizing by both compressing single sentences and fusing pairs. This paper attempts to bridge the gap by ranking singletons and pairs together by their likelihoods of producing summary sentences.

The selection of sentence singletons and pairs can bring benefit to neural abstractive summarization, as a number of studies seek to separate content selection from summary generation (Chen and Bansal, 2018; Hsu et al., 2018; Gehrmann

et al., 2018; Lebanoff et al., 2018). Content selection draws on domain knowledge to identify relevant content, while summary generation weaves together selected source and vocabulary words to form a coherent summary. Despite having local coherence, system summaries can sometimes contain erroneous details (See et al., 2017) and forged content (Cao et al., 2018b; Song et al., 2018). Separating the two tasks of content selection and summary generation allows us to closely examine the compressing and fusing mechanisms of an abstractive summarizer.

In this paper we propose a method to learn to select sentence singletons and pairs, which then serve as the basis for an abstractive summarizer to compose a summary sentence-by-sentence, where singletons are shortened (i.e., compressed) and pairs are merged (i.e., fused). We exploit state-of-the-art neural representations and traditional vector space models to characterize singletons and pairs; we then provide suggestions on the types of representations useful for summarization. Experiments are performed on both single- and multi-document summarization datasets, where we demonstrate the efficacy of selecting sentence singletons and pairs as well as its utility to abstractive summarization. Our research contributions can be summarized as follows:

- the present study fills an important gap by selecting sentence singletons and pairs jointly, assuming a summary sentence can be created by either shortening a singleton or merging a pair. Compared to abstractive summarizers that perform content selection implicitly, our method is flexible and can be extended to multi-document summarization where training data is limited;
- we investigate the factors involved in representing sentence singletons and pairs. We perform extensive experiments and report findings on sentence selection and abstraction.[1]

2 Related Work

Content selection is integral to any summarization system. Neural approaches to abstractive summarization often perform content selection jointly with surface realization using an encoder-decoder architecture (Rush et al., 2015; Nallapati et al., 2016; Chen et al., 2016b; Tan et al., 2017; See et al., 2017; Paulus et al., 2017; Celikyilmaz et al., 2018; Narayan et al., 2018). Training these models end-to-end means learning to perform both tasks simultaneously and can require a massive amount of data that is unavailable and unaffordable for many summarization tasks.

Recent approaches emphasize the importance of separating content selection from summary generation for abstractive summarization. Studies exploit extractive methods to identify content words and sentences that should be part of the summary and use them to guide the generation of abstracts (Chen and Bansal, 2018; Gehrmann et al., 2018; Lebanoff et al., 2018). On the other hand, surface lexical features have been shown to be effective in identifying pertinent content (Carenini et al., 2006; Wong et al., 2008; Galanis et al., 2012). Examples include sentence length, position, centrality, word frequency, whether a sentence contains topic words, and others. The surface cues can also be customized for new domains relatively easily. This paper represents a step forward in this direction, where we focus on developing lightweight models to select summary-worthy sentence singletons and pairs and use them as the basis for summary generation.

A succinct sentence can be generated by shortening or rewriting a lengthy source text. Recent studies have leveraged neural encoder-decoder models to rewrite the first sentence of an article to a title-like summary (Nallapati et al., 2016; Zhou et al., 2017; Li et al., 2017; Song et al., 2018; Guo et al., 2018; Cao et al., 2018a). Compressive summaries can be generated in a similar vein by selecting important source sentences and then dropping inessential sentence elements such as prepositional phrases. Before the era of deep neural networks it has been an active area of research, where sentence selection and compression can be accomplished using a pipeline or a joint model (Daumé III and Marcu, 2002; Zajic et al., 2007; Gillick and Favre, 2009; Wang et al., 2013; Li et al., 2013, 2014; Filippova et al., 2015). A majority of these studies focus on selecting and compressing sentence *singletons* only.

A sentence can also be generated through fusing multiple source sentences. However, many aspects of this approach are largely underinvestigated, such as determining the set of source sentences to be fused, handling its large cardinality, and identifying the sentence relationships for per-

[1] We make our code and models publicly available at https://github.com/ucfnlp/summarization-sing-pair-mix

Sentence Pair:	Merged Sentence:
(A) The bombing killed 58 people. (B) Wajid Shamsul Hasan, Pakistan's high commissioner to Britain, and Hamid Gul, former head of the ISI, firmly denied the agency's involvement in the attack.	Pakistan denies its spy agency helped plan bombing that killed 58.
Sentence Singleton:	**Compressed Sentence:**
(A) Pakistani Maj. Gen. Athar Abbas said the report "unfounded and malicious" and an "effort to malign the ISI," – Pakistan's directorate of inter-services intelligence.	Maj. Gen. Athar Abbas said the report was an "effort to malign the ISI."

Table 1: Example sentence singleton and pair, before and after compression/merging.

forming fusion. Previous studies assume a set of similar source sentences can be gathered by clustering sentences or by comparing to a reference summary sentence (Barzilay and McKeown, 2005; Filippova, 2010; Shen and Li, 2010; Chenal and Cheung, 2016; Liao et al., 2018); but these methods can be suboptimal. Joint models for sentence selection and fusion implicitly perform content planning (Martins and Smith, 2009; Berg-Kirkpatrick et al., 2011; Bing et al., 2015; Durrett et al., 2016) and there is limited control over which sentences are merged and how.

In contrast, this work attempts to teach the system to determine if a sentence singleton or a pair should be selected to produce a summary sentence. A sentence pair (A, B) is preferred over its consisting sentences if they carry complementary content. Table 1 shows an example. Sentence B contains a reference ("*the attack*") and A contains a more complete description for it ("*bombing that killed 58*"). Sentences A and B each contain certain valuable information, and an appropriate way to merge them exists. As a result, a sentence pair can be scored higher than a singleton given the content it carries and compatibility of its consisting sentences. In the following we describe methods to represent singletons and pairs in a unified framework and scoring them for summarization.

3 Our Model

We present the first attempt to transform sentence singletons and pairs to real-valued vector representations capturing semantic salience so that they can be measured against each other (§3.1). This is a nontrivial task, as it requires a direct comparison of texts of varying length—a pair of sentences is almost certainly longer than a single sentence. For sentence pairs, the representations are expected to further encode sentential semantic compatibility. In §3.2, we describe our method to utilize highest scoring singletons and pairs to a neural abstractive summarizer to generate summaries.

3.1 Scoring Sentence Singletons and Pairs

Given a document or set of documents, we create a set $\mathcal{D}$ of singletons and pairs by gathering all single sentences and arbitrary pairs of them. We refer to a singleton or pair in the set as an *instance*. The sentences in a pair are arranged in order of their appearance in the document or by date of documents. Let N be the number of single sentences in the input document(s), a complete set of singletons and pairs will contain $|\mathcal{D}|=\frac{N(N-1)}{2}+N$ instances. Our goal is to score each instance based on the amount of summary-worthy content it conveys. Despite their length difference, a singleton can be scored higher than a pair if it contains a significant amount of salient content. Conversely, a pair can outweigh a singleton if its component sentences are salient and compatible with each other.

Building effective representations for singletons and pairs is therefore of utmost importance. We attempt to build a vector representation for each instance. The representation should be invariant to the instance type, i.e., a singleton or pair. In this paper we exploit the BERT architecture (Devlin et al., 2018) to learn instance representations. The representations are fine-tuned for a classification task predicting whether a given instance contains content used in human-written summary sentences (details for ground-truth creation in §4).

BERT BERT supports our goal of encoding singletons and pairs indiscriminately. It introduces two pretraining tasks to build deep contextual representations for words and sequences. A sequence can be a single sentence (A) or pair of sentences (A+B).[2] The first task predicts *missing words* in the input sequence. The second task predicts if B is the *next sentence* following A. It requires the vector representation for (A+B) to capture the coherence of two sentences. As coherent sentences can often be fused together, we conjecture that the second task is particularly suited for our goal.

[2] In the original BERT paper (Devlin et al., 2018), a "sentence" is used in a general sense to denote an arbitrary span of contiguous text; we refer to an actual linguistic sentence.

Concretely, BERT constructs an input sequence by prepending a singleton or pair with a "[CLS]" symbol and delimiting the two sentences of a pair with "[SEP]." The representation learned for the [CLS] symbol is used as an aggregate sequence representation for the later classification task. We show an example input sequence in Eq. (1). In the case of a singleton, w_i^{B} are padding tokens.

$$\{w_i\} = \text{[CLS]}, w_1^{\mathsf{A}}, w_2^{\mathsf{A}}, \ldots, \text{[SEP]}, w_1^{\mathsf{B}}, w_2^{\mathsf{B}}, \ldots, \text{[SEP]} \quad (1)$$

$$\mathbf{e}_i = \mathbf{e}_{\text{w}}(w_i) + \mathbf{e}_{\text{sgmt}}(w_i) + \mathbf{e}_{\text{wpos}}(w_i) + \mathbf{e}_{\text{spos}}(w_i) \quad (2)$$

In Eq. (2), each token w_i is characterized by an input embedding $\mathbf{e}_i$, calculated as the element-wise sum of the following embeddings:

- $\mathbf{e}_{\text{w}}(w_i)$ is a *token* embedding;
- $\mathbf{e}_{\text{sgmt}}(w_i)$ is a *segment* embedding, signifying whether w_i comes from sentence A or B.
- $\mathbf{e}_{\text{wpos}}(w_i)$ is a *word position* embedding indicating the index of w_i in the input sequence;
- we introduce $\mathbf{e}_{\text{spos}}(w_i)$ to be a *sentence position* embedding; if w_i is from sentence A (or B), $\mathbf{e}_{\text{spos}}(w_i)$ is the embedding indicating the index of sentence A (or B) in the original document.

Intuitively, these embeddings mean that, the extent to which a word contributes to the sequence (A+B) representation depends on these factors: (i) word salience, (ii) importance of sentences A and B, (iii) word position in the sequence, and, (iv) sentence position in the document. These factors coincide with heuristics used in summarization literature (Nenkova and McKeown, 2011), where leading sentences of a document and the first few words of a sentence are more likely to be included in the summary.

The input embeddings are then fed to a multi-layer and multi-head attention architecture to build deep contextual representations for tokens. Each layer employs a Transformer block (Vaswani et al., 2017), which introduces a self-attention mechanism that allows each hidden state $\mathbf{h}_i^l$ to be compared with every other hidden state of the same layer $[\mathbf{h}_1^l, \mathbf{h}_2^l, \ldots, \mathbf{h}_{\mathsf{N}}^l]$ using a parallelizable, multi-head attention mechanism (Eq. (3-4)).

$$\mathbf{h}_i^1 = f_{\text{self-attn}}^1(\mathbf{e}_i, [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{\mathsf{N}}]) \quad (3)$$

$$\mathbf{h}_i^{l+1} = f_{\text{self-attn}}^{l+1}(\mathbf{h}_i^l, [\mathbf{h}_1^l, \mathbf{h}_2^l, \ldots, \mathbf{h}_{\mathsf{N}}^l]) \quad (4)$$

The representation at final layer L for the [CLS] symbol is used as the sequence representation $\mathbf{h}_{\text{[CLS]}}^{\mathsf{L}}$. The representations can be fine-tuned with an additional output layer to generate state-of-the-art results on a wide range of tasks including reading comprehension and natural language inference. We use the pretrained BERT base model and fine-tune it on our specific task of predicting if an instance (a singleton or pair) $p_{\text{inst}} = \sigma(\mathbf{w}^\top \mathbf{h}_{\text{[CLS]}}^{\mathsf{L}})$ is an appropriate one, i.e., belonging to the ground-truth set of summary instances for a given document. At test time, the architecture indiscriminately encodes a mixed collection of sentence singletons/pairs. We then obtain a likelihood score for each instance. This framework is thus a first effort to build semantic representations for singletons and pairs capturing informativeness and semantic compatibility of two sentences.

VSM We are interested in contrasting BERT with the traditional vector space model (Manning et al., 2008) for representing singletons and pairs. BERT learns instance representations by attending to important content words, where the importance is signaled by word and position embeddings as well as pairwise word relationships. Nonetheless, it remains an open question whether BERT can successfully weave the meaning of *topically important words* into representations. A word "border" is topically important if the input document discusses border security. A topic word is likely to be repeatedly mentioned in the input document but less frequently elsewhere. Because sentences containing topical words are often deemed summary-worthy (Hong and Nenkova, 2014), it is desirable to represent sentence singletons and pairs based on the amount of topical content they convey.

VSM represents each sentence as a sparse vector. Each dimension of the vector corresponds to an n-gram weighted by its TF-IDF score. A high TF-IDF score suggests the n-gram is important to the topic of discussion. We further strengthen the sentence vector with position and centrality information, i.e., the sentence position in the document and the cosine similarity between the sentence and document vector. We obtain a document vector by averaging over its sentence vectors, and we similarly obtain a vector for a pair of sentences. We use VSM representations as a baseline to contrast its performance with distributed representations from BERT. To score singletons and pairs, we use the LambdaMART model[3] which has demonstrated success on related NLP tasks (Chen et al., 2016a);

[3] https://sourceforge.net/p/lemur/wiki/RankLib/

it also fits our requirements of ranking singletons and pairs indiscriminately.

3.2 Generating Summaries

We proceed by performing a preliminary investigation of summary generation from singletons and pairs; they are collectively referred to as *instances*. In the previous section, a set of summary instances is selected from a document. These instances are treated as "raw materials" for a summary; they are fed to a neural abstractive summarizer which processes them into summary sentences via fusion and compression. This strategy allows us to separately evaluate the contributions from instance selection and summary composition.

We employ the MMR principle (Carbonell and Goldstein, 1998) to select a set of highest scoring and non-redundant instances. The method adds an instance $\hat{P}$ to the summary $\mathcal{S}$ iteratively per Eq. (5) until a length threshold has been reached. Each instance is weighted by a linear combination of its importance score $\mathcal{I}(P_k)$, obtained by BERT or VSM, and its redundancy score $\mathcal{R}(P_k)$, computed as the cosine similarity between the instance and partial summary. λ is a balancing factor between importance and redundancy.[4] Essentially, MMR prevents the system from selecting instances that are too similar to ones already selected.

$$\hat{P} = \underset{P_k \in \mathcal{D} \setminus \mathcal{S}}{\arg\max} \Big[\lambda \mathcal{I}(P_k) - (1-\lambda)\mathcal{R}(P_k) \Big] \quad (5)$$

Composing a summary from selected instances is a non-trivial task. As a preliminary investigation of summary composition, we make use of pointer-generator (PG) networks (See et al., 2017) to compress/fuse sentences into summary sentences. PG is a sequence-to-sequence model that has achieved state-of-the-art performance in abstractive summarization by having the ability to both copy tokens from the document or generate new tokens from the vocabulary. When trained on document-summary pairs, the model has been shown to remove unnecessary content from sentences and can merge multiple sentences together.

In this work, rather than training on document-summary pairs, we train PG exclusively on ground-truth instances. This removes most of the responsibility of content selection, and allows it to focus its efforts on merging the sentences. We use instances derived from human summaries (§4) to train the network, which includes a sentence singleton or pair along with the ground-truth compressed/merged sentence. At test time, the network receives an instance from BERT or VSM and outputs a summary sentence, then repeats this process to generate several sentences. In Figure 2 we present an illustration of the system architecture.

[4] We use a coefficient λ of 0.6.

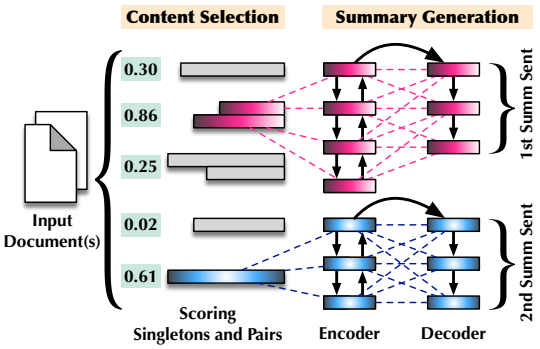

Figure 2: System architecture. In this example, a sentence pair is chosen (red) and then merged to generate the first summary sentence. Next, a sentence singleton is selected (blue) and compressed for the second summary sentence.

4 Data

Our method does not require a massive amount of annotated data. We thus report results on single- and multi-document summarization datasets.

We experiment with (i) XSum (Narayan et al., 2018), a new dataset created for extreme, abstractive summarization. The task is to reduce a news article to a short, one-sentence summary. Both source articles and reference summaries are gathered from the BBC website. The training set contains about 204k article-summary pairs and the test contains 11k pairs. (ii) CNN/DM (Hermann et al., 2015), an abstractive summarization dataset frequently exploited by recent studies. The task is to reduce a news article to a multi-sentence summary (4 sentences on average). The training set contains about 287k article-summary pairs and the test set contains 11k pairs. We use the non-anonymzied version of the dataset. (iii) DUC-04 (Over and Yen, 2004), a benchmark multi-document summarization dataset. The task is to create an abstractive summary (5 sentences on average) from a set of 10 documents discussing a given topic. The dataset contains 50 sets of documents used for testing purpose only. Each document set is associated with four human reference summaries.

We build a training set for both tasks of content selection and summary generation. This is done by creating ground-truth sets of instances based on document-summary pairs. Each document and

summary pair (D, S) is a collection of sentences $D = \{d_1, d_2, ..., d_M\}$ and $S = \{s_1, s_2, ..., s_N\}$. We wish to associate each summary sentence s_n with a subset of the document sentences $\tilde{D} \subseteq D$, which are the sentences that are merged to form s_n. Our method chooses multiple sentences that work together to capture the most overlap with summary sentence s_n, in the following way.

We use averaged ROUGE-1, -2, -L scores (Lin, 2004) to represent sentence similarity. The source sentence most similar to s_n is chosen, which we call $\tilde{d}_1$. All shared words are then removed from s_n to create s'_n, effectively removing all information already captured by $\tilde{d}_1$. A second source sentence $\tilde{d}_2$ is selected that is most similar to the remaining summary sentence s'_n, and shared words are again removed from s'_n to create s''_n. This process of sentence selection and overlap removal is repeated until no remaining sentences have at least two overlapping content words (words that are non-stopwords or punctuation) with s_n. The result is referred to as a ground-truth set $(s_n, \tilde{D})$ where $\tilde{D} = \{\tilde{d}_1, \tilde{d}_2, ..., \tilde{d}_{|\tilde{D}|}\}$. To train the models, $\tilde{D}$ is limited to one or two sentences because it captures the large majority of cases. All empty ground-truth sets are removed, and only the first two sentences are chosen for all ground-truth sets with more than two sentences. A small number of summary sentences have empty ground-truth sets, corresponding to 2.85%, 9.87%, 5.61% of summary sentences in CNN/DM, XSum, and DUC-04 datasets. A detailed plot of the ground-truth set size is illustrated in Figure 1, and samples of the ground-truth are found in the supplementary.

We use the standard train/validation/test splits for both CNN/Daily Mail and XSum. We train our models on ground-truth sets of instances created from the training sets and tune hyperparameters using instances from the validation sets. DUC-04 is a test-only dataset, so we use the models trained on CNN/Daily Mail to evaluate DUC-04. Because the input is in the form of multiple documents, we select the first 20 sentences from each document and concatenate them together into a single mega-document (Lebanoff et al., 2018). For the sentence position feature, we keep the sentence positions from the original documents. This handling of sentence position, along with other features that are invariant to the input type, allows us to effectively train on single-document inputs and transfer to the multi-document setting.

5 Results

Evaluation Setup In this section we evaluate our proposed methods on identifying summary-worthy instances including singletons and pairs. We compare this scheme with traditional methods extracting only singletons, then introduce novel evaluation strategies to compare results. We exploit several strong extractive baselines: (i) *SumBasic* (Vanderwende et al., 2007) extracts sentences by assuming words occurring frequently in a document have higher chances of being included in the summary; (ii) *KL-Sum* (Haghighi and Vanderwende, 2009) greedily adds sentences to the summary to minimize KL divergence; (iii) *LexRank* (Erkan and Radev, 2004) estimates sentence importance based on eigenvector centrality in a document graph representation. Further, we include the LEAD method that selects the first N sentences from each document. We then require all systems to extract N instances, i.e., either singletons or pairs, from the input document(s).[5]

We compare system-identified instances with ground-truth instances, and in particular, we compare against the primary, secondary, and full set of ground-truth sentences. A *primary* sentence is defined as a ground-truth singleton or a sentence in a ground-truth pair that has the highest similarity to the reference summary sentence; the other sentence in the pair is considered *secondary*, which provides complementary information to the primary sentence. E.g., let $\mathcal{S}^*$={(1, 2), 5, (8, 4), 10} be a ground-truth set of instances, where numbers are sentence indices and the first sentence of each pair is primary. Our ground-truth primary set thus contains {1, 5, 8, 10}; secondary set contains {2, 4}; and the full set of ground-truth sentences contains {1, 2, 5, 8, 4, 10}. Assume $\mathcal{S}$={(1, 2), 3, (4, 10), 15} are system-selected instances. We uncollapse all pairs to obtain a set of single sentences $\mathcal{S}$={1, 2, 3, 4, 10, 15}, then compare them against the primary, secondary, and full set of ground-truth sentences to calculate precision, recall, and F1-measure scores. This evaluation scheme allows a fair comparison of a variety of systems for instance selection, and assess their performance on identifying primary and secondary sentences respectively for summary generation.

Extraction Results In Table 2 we present in-

[5] We use N=4/1/5 respectively for the CNN/DM, XSum, and DUC-04 datasets. N is selected as the average number of sentences in reference summaries.

	System	Primary			Secondary			All		
		P	R	F	P	R	F	P	R	F
CNN/Daily Mail	LEAD-Baseline	31.9	38.4	34.9	10.7	34.3	16.3	39.9	37.3	38.6
	SumBasic (Vanderwende et al., 2007)	15.2	17.3	16.2	5.3	15.8	8.0	19.6	16.9	18.1
	KL-Summ (Haghighi et al., 2009)	15.7	17.9	16.7	5.4	15.9	8.0	20.0	17.4	18.6
	LexRank (Erkan and Radev, 2004)	22.0	25.9	23.8	7.2	21.4	10.7	27.5	24.7	26.0
	VSM-SingOnly (This work)	30.8	36.9	33.6	9.8	34.4	15.2	39.5	35.7	37.5
	VSM-SingPairMix (This work)	27.0	46.5	34.2	9.0	42.1	14.9	34.0	45.4	38.9
	BERT-SingOnly (This work)	**35.3**	41.9	38.3	9.8	32.5	15.1	44.0	38.6	41.1
	BERT-SingPairMix (This work)	33.6	**67.1**	**44.8**	**13.6**	**70.2**	**22.8**	**44.7**	**68.0**	**53.9**
XSum	LEAD-Baseline	8.5	9.4	8.9	5.3	9.5	6.8	13.8	9.4	11.2
	SumBasic (Vanderwende et al., 2007)	8.7	9.7	9.2	5.0	8.9	6.4	13.7	9.4	11.1
	KL-Summ (Haghighi et al., 2009)	9.2	10.2	9.7	5.0	8.9	6.4	14.2	9.7	11.5
	LexRank (Erkan and Radev, 2004)	9.7	10.8	10.2	5.5	9.8	7.0	15.2	10.4	12.4
	VSM-SingOnly (This work)	12.3	14.1	13.1	3.8	11.0	5.6	17.9	12.0	14.4
	VSM-SingPairMix (This work)	10.1	22.6	13.9	4.2	17.4	6.8	14.3	20.8	17.0
	BERT-SingOnly (This work)	24.2	26.1	25.1	6.6	16.7	9.5	35.3	20.8	26.2
	BERT-SingPairMix (This work)	**33.2**	**56.0**	**41.7**	**24.1**	**65.5**	**35.2**	**57.3**	**59.6**	**58.5**
DUC-04	LEAD-Baseline	6.0	4.8	5.3	2.8	3.8	3.2	8.8	4.4	5.9
	SumBasic (Vanderwende et al., 2007)	4.2	3.2	3.6	3.0	3.8	3.3	7.2	3.4	4.6
	KL-Summ (Haghighi et al., 2009)	5.6	4.5	5.0	2.8	3.8	3.2	8.0	4.2	5.5
	LexRank (Erkan and Radev, 2004)	8.5	6.7	7.5	**4.8**	6.5	5.5	12.1	6.6	8.6
	VSM-SingOnly (This work)	**18.0**	**14.7**	**16.2**	3.6	8.4	5.0	**23.6**	**11.8**	**15.7**
	VSM-SingPairMix (This work)	3.8	6.2	4.7	3.6	11.4	5.5	7.4	8.0	7.7
	BERT-SingOnly (This work)	8.4	6.5	7.4	2.8	5.3	3.7	15.6	6.6	9.2
	BERT-SingPairMix (This work)	4.8	9.1	6.3	4.2	**14.2**	**6.5**	9.0	10.9	9.9

Table 2: Instance selection results; evaluated for primary, secondary, and all ground-truth sentences. Our BERT-SingPairMix method achieves strong performance owing to its capability of building effective representations for both singletons and pairs.

stance selection results for the CNN/DM, XSum, and DUC-04 datasets. Our method builds representations for instances using either BERT or VSM (§3.1). To ensure a thorough comparison, we experiment with selecting a mixed set of singletons and pairs ("SingPairMix") as well as selecting singletons only ("SingOnly"). On the CNN/DM and XSum datasets, we observe that selecting a mixed set of singletons and pairs based on BERT representations (BERT+SingPairMix) demonstrates the most competitive results. It outperforms a number of strong baselines when evaluated on a full set of ground-truth sentences. The method also performs superiorly on identifying secondary sentences. For example, it increases recall scores for identifying secondary sentences from 33.8% to 69.8% (CNN/DM) and from 16.7% to 65.3% (XSum). Our method is able to achieve strong performance on instance selection owing to BERT's capability of building effective representations for both singletons and pairs. It learns to identify salient source content based on token and position embeddings and it encodes sentential semantic compatibility using the pretraining task of predicting the next sentence; both are valuable additions to summary instance selection.

Further, we observe that identifying summary-worthy singletons and pairs from multi-document inputs (DUC-04) appears to be more challenging than that of single-document inputs (XSum and CNN/DM). This distinction is not surprising given that for multi-document inputs, the system has a large and diverse search space where candidate singletons and pairs are gathered from a set of documents written by different authors.[6] We find that the BERT model performs consistently on identifying secondary sentences, and VSM yields considerable performance gain on selecting primary sentences. Both BERT and VSM models are trained on the CNN/DM dataset and applied to DUC-04 as the latter data are only used for testing. Our findings suggest that the TF-IDF features of the VSM model are effective for multi-document inputs, as important topic words are usually repeated across documents and TF-IDF scores can reflect topical importance of words. This analysis further reveals that extending BERT to incorporate topical salience of words can be a valuable line of research for future work.

[6]For the DUC-04 dataset, we select top K sentences from each document (K=5) and pool them as candidate singletons. Candidate pairs consist of arbitrary combinations of singletons. For all datasets we perform downsampling to balance the number of positive and negative singletons (or pairs).

System	CNN/Daily Mail R-1	R-2	R-L
SumBasic (Vanderwende et al., 2007)	34.11	11.13	31.14
KLSumm (Haghighi et al., 2009)	29.92	10.50	27.37
LexRank (Erkan and Radev, 2004)	35.34	13.31	31.93
PointerGen+Cov (See et al., 2017)	39.53	17.28	36.38
BERT-Abs w/ SS (This Work)	35.49	15.12	33.03
BERT-Abs w/ PG (This Work)	37.15	15.22	34.60
BERT-Extr (This Work)	**41.13**	**18.68**	**37.75**
GT-SingPairMix (This Work)	48.73	26.59	45.29

System	XSum R-1	R-2	R-L
SumBasic (Vanderwende et al., 2007)	18.56	2.91	14.88
KLSumm (Haghighi et al., 2009)	16.73	2.83	13.53
LexRank (Erkan and Radev, 2004)	17.95	3.00	14.30
BERT-Abs w/ PG (This Work)	**25.08**	**6.48**	**19.75**
BERT-Extr (This Work)	23.53	4.54	17.23
GT-SingPairMix (This Work)	27.90	7.31	21.04

System	DUC-04 R-1	R-2	R-SU4
SumBasic (Vanderwende et al., 2007)	29.48	4.25	8.64
KLSumm (Haghighi et al., 2009)	31.04	6.03	10.23
LexRank (Erkan and Radev, 2004)	34.44	7.11	11.19
Extract+Rewrite (Song et al., 2018)	28.90	5.33	8.76
Opinosis (Ganesan et al., 2010)	27.07	5.03	8.63
BERT-Abs w/ PG (This Work)	27.95	4.13	7.75
BERT-Extr (This Work)	30.49	5.12	9.05
GT-SingPairMix (This Work)	41.42	13.67	16.38

Table 3: Summarization results on various datasets. Whether abstractive summaries (BERT-Abst) outperform its extractive variant (BERT-Extr) appears to be related to the amount of sentence pairs selected by BERT-SingPairMix. Selecting more pairs than singletons seems to hurt the abstractor.

Summarization Results We present summarization results in Table 3, where we assess both extractive and abstractive summaries generated by BERT-SingPairMix. We omit VSM results as they are not as competitive as BERT on instance selection for the mixed set of singletons and pairs. The extractive summaries "BERT-Extr" are formed by concatenating selected singletons and pairs for each document, whereas "GT-SingPairMix" concatenates *ground-truth* singletons and pairs; it provides an upper bound for any system generating a set of singletons and pairs as the summary. To assure fair comparison, we limit all extractive summaries to contain up to 100 words (40 words for XSum) for ROUGE evaluation[7], where R-1, R-2, R-L, and R-SU4 are variants used to measure the overlap of unigrams, bigrams, longest common subsequences, and skip bigrams (with a maximum distance of 4) between system and reference summaries (Lin, 2004). The abstractive summaries are generated from the same singletons and pairs used

[7]w/ ROUGE options: -n 2 -m -2 4 -w 1.2 -c 95 -r 1000 -l 100

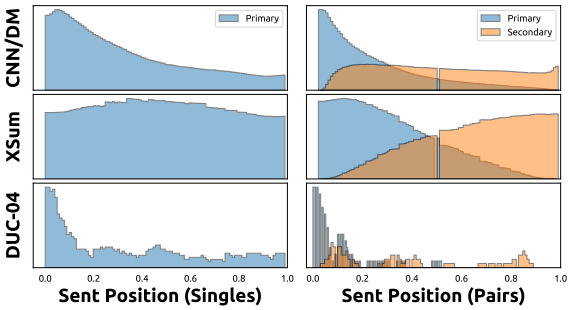

Figure 3: Position of ground-truth singletons and pairs in a document. The singletons of XSum can occur anywhere; the first and second sentence of a pair also appear far apart.

to form system extracts. "BERT-Abs-PG" generates an abstract by iteratively encoding singletons or pairs and decoding summary sentences using pointer-generator networks (§3.2).[8]

Our BERT summarization systems achieve results largely on par with those of prior work. It is interesting to observe that the extractive variant (BERT-Extr) can outperform its abstractive counterparts on DUC-04 and CNN/DM datasets, and vice versa on XSum. A close examination of the results reveals that whether abstractive summaries outperform appears to be related to the amount of sentence pairs selected by "BERT-SingPairMix." Selecting more pairs than singletons seems to hurt the abstractor. For example, BERT selects 100% and 76.90% sentence pairs for DUC-04 and CNN/DM respectively, and 28.02% for XSum. These results suggest that existing abstractors using encoder-decoder models may need to improve on sentence fusion. These models are trained to generate fluent sentences more than preserving salient source content, leading to important content words being skipped in generating summary sentences. Our work intends to separate the tasks of sentence selection and summary generation, thus holding promise for improving compression and merging in the future. We present example system summaries in the supplementary.

Further analysis In this section we perform a series of analyses to understand where summary-worthy content is located in a document and how humans order them into a summary. Figure 3 shows the position of ground-truth singletons and pairs in a document. We observe that singletons of CNN/DM and DUC-04 tend to occur at the beginning of a document, whereas singletons of XSum

[8]We include an additional in-house system "BERT-Abs-SS" for CNN/DM that takes the same input but generates summary sentences using a tree-based decoder.

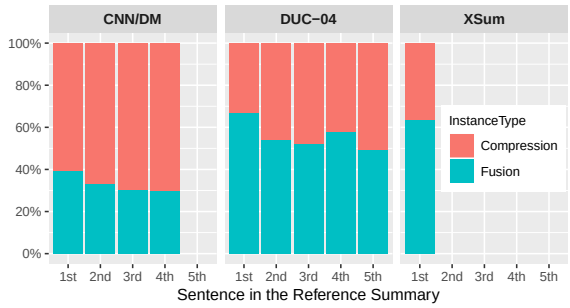

Figure 4: A sentence's *position* in a human summary can affect whether or not it is created by compression or fusion.

can occur anywhere. We also find that the first and second sentence of a pair can appear far apart for XSum, but are closer for CNN/DM. These findings suggest that selecting singletons and pairs for XSum can be more challenging than others, as indicated by the name "extreme" summarization.

Figure 4 illustrates how humans choose to organize content into a summary. Interestingly, we observe that a sentence's *position* in a human summary affects whether or not it is created by compression or fusion. The first sentence of a human-written summary is more likely than the following sentences to be a fusion of multiple source sentences. This is the case across all three datasets. We conjecture that the first sentence of a summary is expected to give an overview of the document and needs to consolidate information from different parts. Other sentences of a human summary can be generated by simply shortening singletons. Our statistics reveal that DUC-04 and XSum summaries involve more fusion operations, exhibiting a higher level of abstraction than CNN/DM.

6 Conclusion

We present an investigation into the feasibility of scoring singletons and pairs according to their likelihoods of producing summary sentences. Our framework is founded on the human process of selecting one or two sentences to merge together and it has the potential to bridge the gap between compression and fusion studies. Our method provides a promising avenue for domain-specific summarization where content selection and summary generation are only loosely connected to reduce the costs of obtaining massive annotated data.

Acknowledgments

We are grateful to the reviewers for their insightful comments that point to interesting future directions. The authors also thank students in the UCF NLP group for useful discussions.

References

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3).

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca J. Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of ACL*.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018a. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018b. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Giuseppe Carenini, Raymond Ng, and Adam Pauls. 2006. Multi-document summarization of evaluative text. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016a. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016b. Distraction-based neural networks for document summarization. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Victor Chenal and Jackie Chi Kit Cheung. 2016. Predicting sentential semantic compatibility for aggregation in text-to-text generation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of ACL*.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hal Daumé III and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hal Daumé III and Daniel Marcu. 2004. Generic sentence fusion is an ill-defined summarization task. In *Proceedings of ACL Workshop on Text Summarization Branches Out*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*.

Katja Filippova. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Katja Filippova, Enrique Alfonseca, Carlos Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the NAACL Workshop on Integer Linear Programming for Natural Langauge Processing*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft, layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of Neural Information Processing Systems (NIPS)*.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alex Kulesza and Ben Taskar. 2011. Learning determinantal point processes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chen Li, Fei Liu, Fuliang Weng, and Yang Liu. 2013. Document summarization via guided sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chen Li, Yang Liu, Fei Liu, Lin Zhao, and Fuliang Weng. 2014. Improving multi-document summarization by sentence compression based on expanded constituent parse tree. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract meaning representation for multi-document summarization. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Proceedings of ACL Workshop on Text Summarization Branches Out*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Andre F. T. Martins and Noah A. Smith. 2009. Summarization with a joint model for sentence extraction and compression. In *Proceedings of the ACL Workshop on Integer Linear Programming for Natural Language Processing*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of SIGNLL*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*.

Paul Over and James Yen. 2004. An introduction to DUC-2004. *National Institute of Standards and Technology*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for sentence summarization. In *Proceedings of EMNLP*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chao Shen and Tao Li. 2010. Multi-document summarization via the minimum dominating set. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Kaiqiang Song, Lin Zhao, and Fei Liu. 2018. Structure-infused copy mechanisms for abstractive summarization. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kapil Thadani and Kathleen McKeown. 2013. Supervised sentence fusion with single-stage inference. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. https://arxiv.org/abs/1706.03762. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*.

Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2013. A sentence compression based framework to query-focused multi-document summarization. In *Proceedings of ACL*.

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

David Zajic, Bonnie J. Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management*.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

A Ground-truth Sets of Instances

We performed a manual inspection over a subset of our ground-truth sets of singletons and pairs. Each sentence from a human-written summary is matched with one or two source sentences based on average ROUGE similarity (details in Section 4 of the paper). Tables 4, 5, and 6 present randomly selected examples from CNN/Daily Mail, XSum, and DUC-04, respectively. Colored text represents overlapping tokens between sentences. Darker colors represent content from primary sentences, while lighter colors represent content from secondary sentences. Best viewed in color.

B Example Summaries

Table 7 presents example system summaries and human-written abstracts from CNN/Daily Mail. Each Human Abstract sentence is matched with a sentence singleton or pair from the source document; these singletons/pairs make up the GT-SingPairMix summary. Similarly, each sentence from BERT-Abs is created by compressing a singleton or merging a pair selected by BERT-Extr.

Selected Source Sentence(s)	Human Summary Sentence
an inmate housed on the " forgotten floor , " where many mentally ill inmates are housed in miami before trial .	mentally ill inmates in miami are housed on the " forgotten floor "
most often , they face drug charges or charges of assaulting an officer – charges that judge steven leifman says are usually " avoidable felonies . "	judge steven leifman says most are there as a result of " avoidable felonies "
" i am the son of the president . miami , florida -lrb- cnn -rrb- – the ninth floor of the miami-dade pretrial detention facility is dubbed the " forgotten floor . "	while cnn tours facility , patient shouts : " i am the son of the president "
it 's brutally unjust , in his mind , and he has become a strong advocate for changing things in miami . so , he says , the sheer volume is overwhelming the system , and the result is what we see on the ninth floor .	leifman says the system is unjust and he 's fighting for change .

Selected Source Sentence(s)	Human Summary Sentence
the average surface temperature has warmed one degree fahrenheit -lrb- 0.6 degrees celsius -rrb- during the last century , according to the national research council .	earth has warmed one degree in past 100 years .
the reason most cited – by scientists and scientific organizations – for the current warming trend is an increase in the concentrations of greenhouse gases , which are in the atmosphere naturally and help keep the planet 's temperature at a comfortable level . in the worst-case scenario , experts say oceans could rise to overwhelming and catastrophic levels , flooding cities and altering seashores .	majority of scientists say greenhouse gases are causing temperatures to rise .
a change in the earth 's orbit or the intensity of the sun 's radiation could change , triggering warming or cooling . other scientists and observers , a minority compared to those who believe the warming trend is something ominous , say it is simply the latest shift in the cyclical patterns of a planet 's life .	some critics say planets often in periods of warming or cooling .

Table 4: Sample of our ground-truth labels for singleton/pair instances from CNN/Daily Mail. Large chunks of text are copied straight out of the source sentences.

Selected Source Sentence(s)	Human Summary Sentence
the premises , used by east belfast mp naomi long , have been targeted a number of times . army explosives experts were called out to deal with a suspect package at the offices on the newtownards road on friday night .	a suspicious package left outside an alliance party office in east belfast has been declared a hoax .

Selected Source Sentence(s)	Human Summary Sentence
nev edwards scored an early try for sale , before castres ' florian vialelle went over , but julien dumora 's penalty put the hosts 10-7 ahead at the break .	a late penalty try gave sale victory over castres at stade pierre-antoine in their european challenge cup clash .

Selected Source Sentence(s)	Human Summary Sentence
speaking in the dil , sinn fin leader gerry adams also called for a commission of investigation and said his party had " little confidence the government is protecting the public interest " . last year , nama sold its entire 850-property loan portfolio in northern ireland to the new york investment firm cerberus for more than # 1bn .	the irish government has rejected calls to set up a commission of investigation into the sale of nama 's portfolio of loans in northern ireland .

Table 5: Sample of our ground-truth labels for singleton/pair instances from XSum. Each article has only one summary sentences, and thus only one singleton or pair matched with it.

Selected Source Sentence(s)	Human Summary Sentence
hun sen 's cambodian people 's party won 64 of the 122 parliamentary seats in july 's elections , short of the two-thirds majority needed to form a government on its own .	cambodian elections , fraudulent according to opposition parties , gave the cpp of hun sen a scant majority but not enough to form its own government .
opposition leaders prince norodom ranariddh and sam rainsy , citing hun sen 's threats to arrest opposition figures after two alleged attempts on his life , said they could not negotiate freely in cambodia and called for talks at sihanouk 's residence in beijing . cambodian leader hun sen has guaranteed the safety and political freedom of all politicians , trying to ease the fears of his rivals that they will be arrested or killed if they return to the country .	opposition leaders fearing arrest , or worse , fled and asked for talks outside the country .
the cambodian people 's party criticized a non-binding resolution passed earlier this month by the u.s. house of representatives calling for an investigation into violations of international humanitarian law allegedly committed by hun sen .	the un found evidence of rights violations by hun sen prompting the us house to call for an investigation .
cambodian politicians expressed hope monday that a new partnership between the parties of strongman hun sen and his rival , prince norodom ranariddh , in a coalition government would not end in more violence .	the three-month governmental deadlock ended with han sen and his chief rival , prince norodom ranariddh sharing power .
citing hun sen 's threats to arrest opposition politicians following two alleged attempts on his life , ranariddh and sam rainsy have said they do not feel safe negotiating inside the country and asked the king to chair the summit at gis residence in beijing . after a meeting between hun sen and the new french ambassador to cambodia , hun sen aide prak sokhonn said the cambodian leader had repeated calls for the opposition to return , but expressed concern that the international community may be asked for security guarantees .	han sen guaranteed safe return to cambodia for all opponents but his strongest critic , sam rainsy , remained wary .
diplomatic efforts to revive the stalled talks appeared to bear fruit monday as japanese foreign affairs secretary of state nobutaka machimura said king norodom sihanouk has called on ranariddh and sam rainsy to return to cambodia . king norodom sihanouk on tuesday praised agreements by cambodia 's top two political parties – previously bitter rivals – to form a coalition government led by strongman hun sen .	chief of state king norodom sihanouk praised the agreement .

Table 6: Sample of our ground-truth labels for singleton/pair instances from DUC-04, a multi-document dataset. Ground-truth sentences are widely dispersed among all ten documents.

Extractive Upper Bound	**Human Abstract**
• She's a high school freshman with Down syndrome. • Trey – a star on Eastern High School's basketball team in Louisville, Kentucky, who's headed to play college ball next year at Ball State – was originally going to take his girlfriend to Eastern's prom.	• College-bound basketball star asks girl with Down syndrome to high school prom.
• Trina Helson, a teacher at Eastern, alerted the school's newspaper staff to the prom-posal and posted photos of Trey and Ellie on Twitter that have gone viral.	• Pictures of the two during the "prom-posal" have gone viral.
BERT-Extractive	**BERT-Abstractive**
• But all that changed Thursday when Trey asked Ellie to be his prom date. • Trey – a star on Eastern High School's basketball team in Louisville, Kentucky, who's headed to play college ball next year at Ball State – was originally going to take his girlfriend to Eastern's prom.	• Trey asked Ellie to be his prom date.
• Trina Helson, a teacher at Eastern, alerted the school's newspaper staff to the prom-posal and posted photos of Trey and Ellie on Twitter that have gone viral.	• Trina Helson, a teacher at Eastern, alerted the school's newspaper staff.
• (CNN) He's a blue chip college basketball recruit. • She's a high school freshman with Down syndrome.	• He's a high school student with Down syndrome.

Extractive Upper Bound	**Human Abstract**
• Marseille prosecutor Brice Robin told CNN that "so far no videos were used in the crash investigation."	• Marseille prosecutor says "so far no videos were used in the crash investigation" despite media reports.
• Reichelt told "Erin Burnett: outfront" that he had watched the video and stood by the report, saying Bild and Paris Match are "very confident" that the clip is real.	• Journalists at Bild and Paris Match are "very confident" the video clip is real, an editor says.
• Lubitz told his Lufthansa flight training school in 2009 that he had a "previous episode of severe depression," the airline said Tuesday.	• Andreas Lubitz had informed his Lufthansa training school of an episode of severe depression, airline says.
BERT-Extractive	**BERT-Abstractive**
• Marseille, France (CNN) - the French prosecutor leading an investigation into the crash of Germanwings flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. • Marseille prosecutor Brice Robin told CNN that "so far no videos were used in the crash investigation."	• New : French prosecutor says he was not aware of video footage from on board the plane.
• Robin's comments follow claims by two magazines, German Daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings flight 9525 as it crashed into the French Alps. • The two publications described the supposed video, but did not post it on their websites.	• Two magazines, including German Daily Bild, have been described as the video.

Table 7: Example system summaries and human-written abstracts. Each Human Abstract sentence is lined up horizontally with its corresponding ground-truth instance, which is found in Extractive Upper Bound summary. Similarly, each sentence from BERT-Abstractive is lined up horizontally with its corresponding instance selected by BERT-Extractive. The sentences are manually de-tokenized for readability.