Conditional Neural Generation using Sub-Aspect Functions for Extractive News Summarization

Zhengyuan Liu, Ke Shi, Nancy F. Chen
Institute for Infocomm Research, A*STAR, Singapore
{liu_zhengyuan,shi_ke,nfychen}@i2r.a-star.edu.sg

Abstract

Much progress has been made in text summarization, fueled by neural architectures using large-scale training corpora. However, in the news domain, neural models easily overfit by leveraging position-related features due to the prevalence of the inverted pyramid writing style. In addition, there is an unmet need to generate a variety of summaries for different users. In this paper, we propose a neural framework that can flexibly control summary generation by introducing a set of sub-aspect functions (i.e. importance, diversity, position). These sub-aspect functions are regulated by a set of control codes to decide which sub-aspect to focus on during summary generation. We demonstrate that extracted summaries with minimal position bias is comparable with those generated by standard models that take advantage of position preference. We also show that news summaries generated with a focus on diversity can be more preferred by human raters. These results suggest that a more flexible neural summarization framework providing more control options could be desirable in tailoring to different user preferences, which is useful since it is often impractical to articulate such preferences for different applications *a priori*.

1 Introduction

Text summarization targets to automatically generate a shorter version of the source content while retaining the most important information. As a straightforward and effective method, extractive summarization creates a summary by selecting and subsequently concatenating the most salient semantic units in a document. Recently, neural approaches, often trained in an end-to-end manner, have achieved favorable improvements on various large-scale benchmarks (Nallapati et al., 2017; Narayan et al., 2018a; Liu and Lapata, 2019).

Despite renewed interest and avid development in extractive summarization, there are still long-standing, unresolved challenges. One major problem is position bias, which is especially common in the news domain, where the majority of research in summarization is studied. In many news articles, sentences appearing earlier tend to be more important for summarization tasks (Hong and Nenkova, 2014), and this preference is reflected in reference summaries of public datasets. However, while this tendency is common due to the classic textbook writing style of the "inverted pyramid" (Scanlan, 1999), news articles can be presented in various ways. Other journalism writing styles include anecdotal lead, question-and-answer format, and chronological organization (Stovall, 1985). Therefore, salient information could also be scattered across the entire article, instead of being concentrated in the first few sentences, depending on the chosen writing style of the journalist.

As the "inverted pyramid" style is widespread in news articles (Kryscinski et al., 2019), neural models would easily overfit on position-related features in extractive summarization tasks because of the data-driven learning setup which tags on to features that correlate the most with the output. As a result, those models would select the sentences at the very beginning of a document as best candidates regardless of considering the full context, resulting in sub-optimal models with fancy neural architectures that do not generalize well to other domains (Kedzie et al., 2018).

Additionally, according to Nenkova et al. (2007): *"Content selection is not a deterministic process (Salton et al., 1997; Marcu, 1997; Mani, 2001). Different people choose different sentences to include in a summary, and even the same person can select different sentences at different times (Rath et al., 1961). Such observations lead to concerns about the advisability of using a single human model ..."*, such observations suggest that individuals differ on what she considers key information under different circumstances. This reflects the need to generate

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1453–1463
November 16 - 20, 2020.

application-specific summaries, which is challenging without establishing appropriate expectations and knowledge of targeted readers prior to model development and ground-truth construction. However, publicly available datasets only provide one associated reference summary to a document. Without any explicit instructions and targeted applications or user preferences, ground-truth construction for summarization becomes an under-constrained assignment (Kryscinski et al., 2019). Therefore, it is challenging for end-to-end models to generate alternative summaries without proper anchoring from reference summaries, making it harder for such models to reach their full potential.

In this work, we propose a flexible neural summarization framework that is able to provide more explicit control options when automatically generating summaries (see Figure 1). Since summarization has been regarded as a combination of sub-aspect functions (e.g. information, layout) (Carbonell and Goldstein, 1998; Lin and Bilmes, 2012), we follow the spirit of sub-aspect theory and adopt control codes on sub-aspects to condition summary generation. The advantages are two-fold: (1) It provides a systematic approach to investigate and analyze how one might minimize position bias in extractive news summarization in neural modeling. Most, if not all, previous work like (Jung et al., 2019; Kryscinski et al., 2019) only focus on analyzing the degree and prevalence of position bias. In this work, we take one step further to propose a research methodology direction to disentangle position bias from important and non-redundant summary content. (2) Text summarization needs are often domain or application specific, and difficult to articulate *a priori* what the user-preferences are, thus requiring potential iterations to adapt and refine. However, human ground-truth construction for summarization is time-consuming and labor-intensive. Therefore, a more flexible summary generation framework could minimize manual labor and generate useful summaries more efficiently.

An ideal set of sub-aspect control codes should characterize different aspects of summarization well in a comprehensive manner but at the same time delineate a relatively clear boundary between one another to minimize the set size (Higgins et al., 2017). To achieve this, we adopt the sub-aspects defined in (Jung et al., 2019): `IMPORTANCE`, `DIVERSITY`, and `POSITION`, and assess their characterization capability on the CNN/Daily Mail news

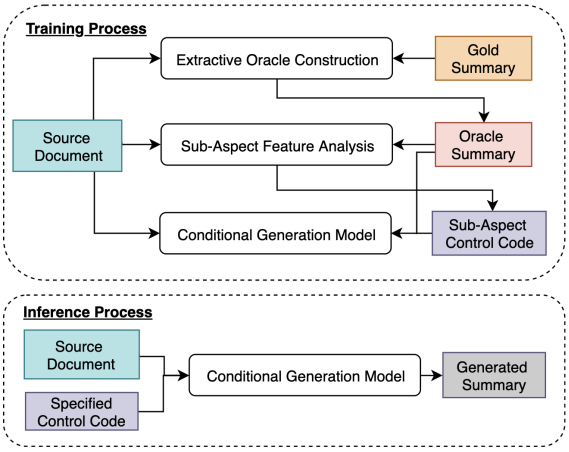

Figure 1: Proposed conditional generation framework exploiting sub-aspect functions.

corpus (Hermann et al., 2015) via quantitative analyses and unsupervised clustering. We utilize control codes based on these three sub-aspect functions to label the training data and implement our conditional generation approach with a neural selector model. Empirical results show that given different control codes, the model can generate output summaries of alternative styles while maintaining performance comparable to the state-of-the-art model; modulation with semantic sub-aspects can reduce systemic bias learned on a news corpus and improve potential generality across domains.

2 In Relation to Other Work

In text summarization, most benchmark datasets focus on the news domain, such as NYT (Sandhaus, 2008) and CNN/Daily Mail (Hermann et al., 2015), where the human-written summaries are used in both abstractive and extractive paradigms (Gehrmann et al., 2018). To improve the performance of extractive summarization, non-neural approaches explore various linguistic and statistical features such as lexical characteristics (Kupiec et al., 1995), latent topic information (Ying-Lang Chang and Chien, 2009), discourse analysis (Hirao et al., 2015; Liu and Chen, 2019), and graph-based modeling (Erkan and Radev, 2004; Mihalcea and Tarau, 2004) . In contrast, neural approaches learn the features in a data-driven manner. Based on recurrent neural networks, SummaRuNNer is one of the earliest neural models (Nallapati et al., 2017). Much development in extractive summarization has been made via reinforcement learning (Narayan et al., 2018b), jointly learning of scoring and ranking (Zhou et al., 2018), and deep contex-

tual language models (Liu and Lapata, 2019).

Despite much development in recent neural approaches, there are still challenges such as corpus bias resulting from the prevalent "inverted pyramid" journalism writing style (Lin and Hovy, 1997), and system bias (Jung et al., 2019) stemming from position preference in the ground-truth. However, to date only analysis work has been done to characterize the position-bias problem and its ramifications, such as inability to generalize across corpora or domains (Kedzie et al., 2018; Kryscinski et al., 2019). Few, if any, have attempted to resolve this long-standing problem of position bias using neural approaches. In this work, we take a first stab to introduce sub-aspect functions for conditional extractive summarization. We explore the possibility of disentangling the three sub-aspects that are commonly used to characterize summarization: `POSITION` for choosing sentences by their position, `IMPORTANCE` for choosing relevant and repeating content across the document, and `DIVERSITY` for ensuring minimal redundancy between summary sentences (Jung et al., 2019) during the summary generation process. In particular, we use these three sub-aspects as control codes for conditional training. To the best of our knowledge, this is the first work in applying auxiliary conditional codes for extractive summary generation.

In other NLP tasks, topic information is used as conditional signals and applied to dialogue response generation (Xing et al., 2017) and pre-training of large-scale language models (Keskar et al., 2019) while sentiment polarity is used in text style transfer (John et al., 2019). In image style transfer, codes specifying color or texture are used to train conditional generative models (Mirza and Osindero, 2014; Higgins et al., 2017).

3 Extractive Oracle Construction

3.1 Similarity Metric: Semantic Affinity vs. Lexical Overlap

For benchmark corpora that are widely adopted, e.g. CNN/Daily Mail (Hermann et al., 2015), there are only golden abstractive summaries written by humans with no corresponding extractive oracle summaries. To convert the human-written abstracts to extractive oracle summaries, most previous work used ROUGE score (Lin, 2004), which counts contiguous n-gram overlap, as the similarity criteria to rank and select sentences from the source content. Since ROUGE scores only conduct lexical matching using word overlapping algorithms, salient sentences from the source content paraphrased by human-editors could be overlooked as the ROUGE scores would be low, while sentences with a high count of common words could get an inflated ROUGE score (Kryscinski et al., 2019).

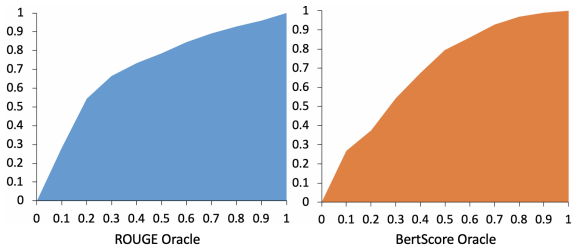

Figure 2: Cumulative position distribution of oracles built on ROGUE (Blue) and BertScore (Orange). X axis is the ratio of article length. Y axis is the cumulative percentage of summary sentences.

To tackle this drawback of ROUGE, we propose to apply the semantic similarity metric BertScore (Zhang et al., 2020) to rank the candidate sentences. BertScore has performed better than ROUGE and BLEU in sentence-level semantic similarity assessment (Zhang et al., 2020). Moreover, BertScore includes recall measures between reference and candidate sequences, a more suitable metric than distance-based similarity measures (Wieting et al., 2019; Reimers and Gurevych, 2019) for summarization related tasks, where there is an asymmetrical relationship between the reference and the generated text.

3.2 Oracle Construction and Evaluation

To build oracles with semantic similarity, we first segment sentences in source documents and human-written gold summaries[1]. Then we convert the text to a semantically rich distributed vector space. For each sentence in a gold summary, we use BertScore to calculate its semantic similarity with candidates from the source content, then the sentence with the highest recall score is chosen. Candidates with a recall score lower than 0.5 are excluded to streamline the selection process.

We observed that the oracle summaries generated through semantic similarity differ from those chosen from n-gram overlap. The positional distributions of two schemes are different, where early sentence bias is less significant for the BertScore scheme (see Figure 2). To further evaluate the effectiveness of this oracle construction approach,

[1] See details of the corpus in Appendix A.

	ROUGE-1 F1 Score	ROUGE-2 F1 Score
ROUGE Oracle	51.84	31.08
BertScore Oracle	50.56	29.41
Similarity Evaluation		**Score**
Gold Summaries		-
ROUGE Candidates		0.70
BertScore Candidates		0.84
QA Paradigm Evaluation		**Accuracy**
Entity and Event Questions:		
Gold Summaries		0.95
ROUGE Candidates		0.54
BertScore Candidates		0.72
Extended Questions:		
Gold Summaries		0.87
ROUGE Candidates		0.52
BertScore Candidates		0.70

Table 1: ROUGE and Human evaluation scores of oracle summaries built on BertScore and ROUGE.

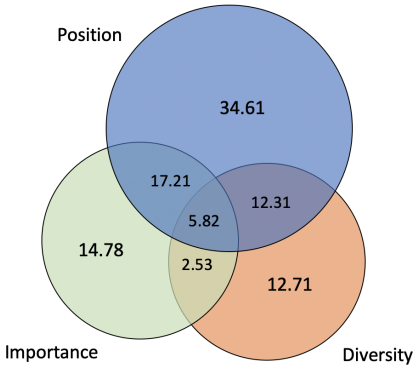

Figure 3: Sample-level distribution of sub-aspect functions of the BertScore oracle. Values are the percentage in categorized samples, which adds up to 60.03% of CNN/Daily Mail training set. The remaining 39.97% do not belong to any of these 3 sub-aspects.

we conducted two assessments. ROUGE scores were computed with the gold summaries. Table 1 shows oracle summaries derived from BertScore are comparable though slightly lower than those from ROUGE, which is not unexpected given that BertScore is mismatched with the ROUGE metric. We also conducted two human evaluations. First, we ranked the candidate summary pairs of 50 news samples based on their similarity to human-written gold summaries (Narayan et al., 2018a). Four linguistic analyzers were asked to consider two aspects: informativeness and coherence (Radev et al., 2002). The evaluation score represents the likelihood of a higher ranking, and is normalized to $[0, 1]$. Next, we adopted the question-answering paradigm (Liu and Lapata, 2019) to evaluate 30 selected samples. For each sentence in the gold summary, questions were constructed based on key information such as events and named entities. Questions where the answer can only be obtained by comprehending the full summary were also included. Human annotators were asked to answer these questions given an oracle summary. The extractive summaries constructed with BertScore are significantly higher in all human evaluations (see Table 1).

4 Sub-Aspect Control Codes

4.1 Sub-Aspect Features in News Summarization

Conditional generation often uses control codes as an auxiliary vector to adjust pre-defined style features. Classic examples include sentiment polarity in style transfer (John et al., 2019) or physical attributes (e.g. color) in image generation (Higgins et al., 2017). However, for summarization it is challenging to pinpoint such intuitive or well-defined features, as the writing style could vary according to genre, topic, or editor preference.

In this work, we adopt *position*, *importance* and *diversity* as a set of sub-function features to characterize extractive news summarization (Jung et al., 2019). Considerations include: (1) "inverted pyramid" writing style is common in news articles, thus making layout or position a salient sub-aspect for summarization; (2) Importance sub-aspect indicates the assumption that repeatedly occurring content in the source document contains more important information; (3) Diversity sub-aspect suggests that selected salient sentences should maximize the semantic volume in a distributed semantic space (Lin and Bilmes, 2012; Yogatama et al., 2015).

4.2 Summary-Level Quantitative Analysis

We apply two methods to evaluate the compatibility and effectiveness of the sub-aspects we choose for extractive news summarization. First, we conduct a quantitative analysis on the CNN/Daily Mail corpus, based on the assumption that the writing style variability of summaries can be characterized through different combinations of sub-aspects (Lin and Bilmes, 2012).

For each source document, we converted all sentences to vector representations with a pre-trained contextual language model BERT (Devlin et al., 2019)[2]. For each sentence, we averaged hidden states of all tokens as the sentence embedding. Similar to (Jung et al., 2019), to obtain the subset of sentences which correspond to *importance* sub-aspect,

[2]https://github.com/google-research/bert

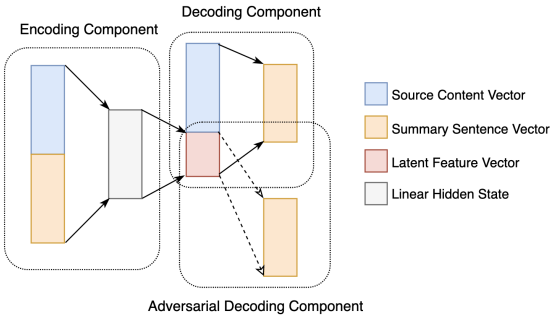

Figure 4: Autoencoder with adversarial training strategy for unsupervised clustering of sentence-level distribution of sub-aspect functions.

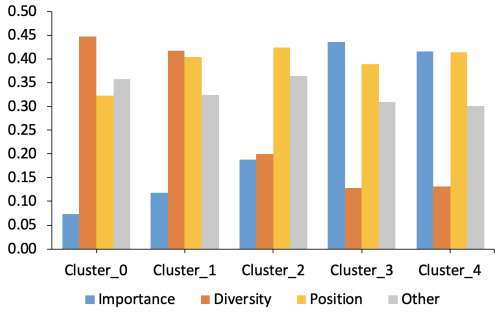

Figure 5: Sentence-level clustering result labeled with sub-aspect features. X axis is the cluster index. Y axis is the proportion of sub-aspect features in each cluster.

we adopted an N-Nearest method which calculates an averaged Pearson correlation between one sentence and the rest for all source sentence vectors, and collected the first-k candidates with the highest scores (k equals oracle summary length). To obtain the subset which corresponds to the *diversity* sub-aspect, we used one implementation[3] of the QuickHull algorithm (Barber et al., 1996) to find vertices, which can be regarded as sentences that maximize the volume size in a projected semantic space. For the subset that corresponds to the *position* sub-aspect, the first 4 sentences in the source document were chosen.

With three sets of sub-aspects, we quantified the distribution of different sub-aspects on the extractive oracle constructed in Section 3. An oracle summary will be mapped to the importance sub-aspect when at least two sentences in the summary are in the subset of *importance* sub-aspect. For those oracle summaries that are shorter than 3 sentences (occupying 19% of the oracle), only one sentence was used to determine which sub-aspect they would be mapped to. Note that the mapping is many to many; i.e. each summary can be mapped to more than one sub-aspect. Figure 3 displays the distribution of the three sub-aspect functions of the oracle summaries, where *position* occupies the largest area. This visualization shows that the three sub-aspects represent distinct linguistic attributes but could overlap with one another.

4.3 Sentence-Level Unsupervised Analysis

According to the mapping algorithm in the previous section, 39% summaries were not mapped to a sub-aspect. This finding motivated us to investigate the distribution of sub-aspect functions *at the sentence level*. Thus, we conducted unsupervised clustering, assuming that samples within one cluster are most similar to each other and they can be represented by the dominant feature.

As shown in Figure 4, we use an autoencoder architecture with adversarial training to model the correlation between document and summary sentences in the semantic space. The encoding component receives the source document representation and one summary sentence representation as input, and compresses it to a latent feature vector. Then, the latent vector and document vector are concatenated and fed to the decoding component to reconstruct the sentence vector. To obtain a compact yet effective latent vector representing the correlation between the source and summary, we adopt an adversarial training strategy as in (John et al., 2019). More specifically, the adversarial decoder we include aims to reconstruct the sentence vector directly from the latent vector. During the training process, we update parameters of the autoencoder with an adversarial penalty (see Appendix B for implementation details). After training this autoencoder, we conduct k-means clustering ($k = 5$) on the latent representation vectors. Then, we analyze the clustering output, with the sentence-level labels of sub-aspect functions as defined in Section 4.2. As shown in Figure 5, sentences with position sub-aspect is distributed relatively equally across each cluster, while importance and diversity dominate in respectively different clusters. Based on the clustering results, we assign the sub-aspect function which is dominant to unmapped sentences in the same cluster. For instance, diversity is assigned to unmapped sentences in cluster 0 and 1 while importance is assigned to those in cluster 3 and 4. By doing this, we reduce $\approx 78\%$ of unmapped sentences and further reduce 35% unmapped summaries using the same criteria in Section 4.2.

[3] http://www.qhull.org/

5 Conditional Neural Generation

In this section, we construct a set of control codes to specify the three sub-aspect features described in Section 4, and label the oracle summaries constructed in Section 3, then we propose a neural extractive model with a conditional learning strategy for a more flexible summary generation.

5.1 Control Code Specification Scheme

The control codes are constructed in the form of *[importance, diversity, position]* to specify sub-aspect features. We can flexibly indicate the *'ON'* and *'OFF'* state of each sub-aspect by switching its corresponding value to 1 or 0, thus enabling disentanglement of each sub-aspect function. For instance, the control code $[1, 0, 0]$ would tell the model to focus more on importance during sentence scoring and selection, while $[0, 1, 1]$ would focus on both diversity and position. Indeed, switching the position code to 0 would help the model obtain minimal position bias. Note that this does not mean the first few sentences would not be selected, as there is overlap between position, importance and diversity (shown in Figure 3). There are 8 control codes under this specification scheme, and we expect this code design can provide the model with sub-aspect conditions for generating summaries.

5.2 Neural Extractive Selector

Given a document D containing a number of sentences $[s_0, s_1, ..., s_n]$, the content selector assigns a score $y_i \in [0, 1]$ to each sentence i, indicating its probability of being included in the summary. A neural model can be trained as an extractive selector for text summarization tasks by contextually modeling the source content.

Here, we implemented and adapted the neural extractive selector in a sequence labeling manner (Kedzie et al., 2018). As shown in Figure 6, the model consists of three components: a contextual encoding component, a selection modeling component and an output component. First, we used BERT in the contextual encoding component to obtain feature-rich sentence-level representations. Then, in the training process, we concatenated these sentence embeddings with the pre-calculated control code vector and fed them to the next layer, which models the contextual hidden states with the conditional signals. Next, a linear layer with Sigmoid function receives the hidden states and produces scores for each segment between 0 and 1

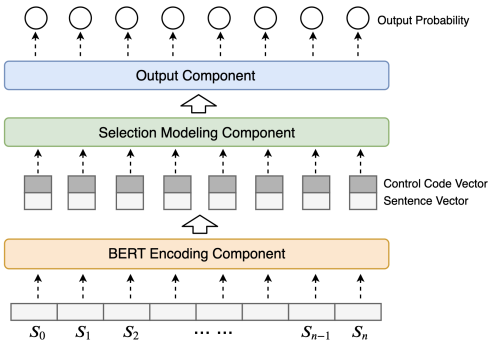

Figure 6: Overview of the neural selector architecture.

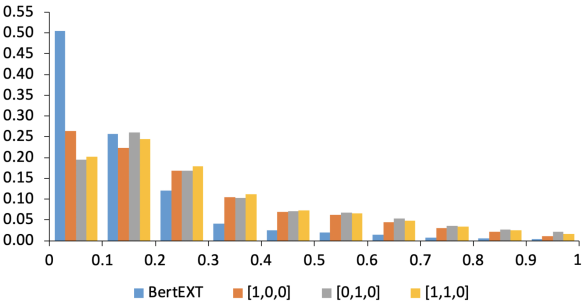

Figure 7: Position distribution of generated summaries from a strong baseline model BertEXT and our conditional summarization model with position code set to 0 (3 implementations). X axis is the position ratio. Y axis is the sentence-level proportion.

as the probability of extractive selection. While this architecture is straightforward, it has shown to be competitive when combined with state-of-the-art contextual representation (Liu and Lapata, 2019).

In our setting, sentences were processed by a subword tokenizer (Wu et al., 2016) and their embeddings were initialized with 768-dimension "base-uncased" BERT (Devlin et al., 2019) and were fixed during training. Lengthy source documents were not truncated. For the selection modeling component, we applied a multi-layer Bi-directional LSTM (Schuster and Paliwal, 1997) and a Transformer network (Vaswani et al., 2017) and it was empirically shown that a two-layer Bi-LSTM performed best (see Appendix C for more implementation details). During testing, sentences with the top-3 selection probability were extracted as output summary, and we used the Trigram Blocking strategy (Paulus et al., 2017) to reduce redundancy.

6 Experimental Results and Analysis

6.1 Quantitative Analysis

To test the possibility of reducing position bias by conditioning summary generation, we switched the position code to 0 and compared the position

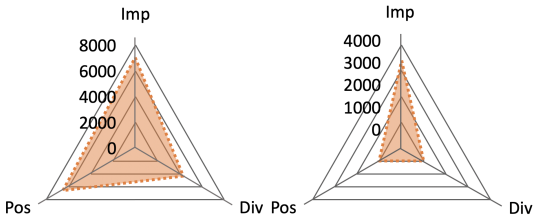

Figure 8: Sub-aspect mapping of generated summary with importance-focus code [1,0,0]. Left panel: one sentence in the summary belongs to importance sub-aspect. Right panel: two sentences in the summary belong to importance sub-aspect. Contour lines denote the number of generated summaries.

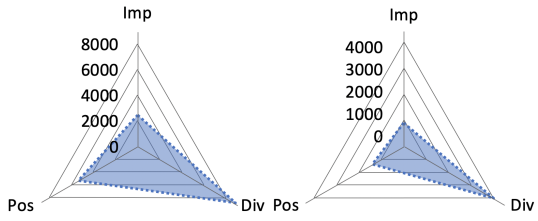

Figure 9: Sub-aspect mapping of generated summary with diversity-focus code [0,1,0]. Left panel: one sentence in the summary belongs to diversity sub-aspect. Right panel: two sentences in the summary belong to diversity sub-aspect. Contour lines denote the number of generated summaries.

of selected sentences in summaries generated by our model to the state-of-the-art baseline BertEXT, based on fine-tuning BERT (Liu and Lapata, 2019). The results show that BertEXT has a 50% chance of choosing the first 10% of sentences in the document. While the proposed framework still has a stronger tendency to choose sentences from the first 30% of the sentences, its position distribution is flattened compared to that of BertEXT.

We respectively switched importance and diversity codes to 1 and categorized the generated summaries into subset of each sub-aspect function as in Section 4.2. As shown in Figure 8 and 9, summaries in the subset of importance and diversity weigh higher when the corresponding control codes are ON. Together, these results demonstrate the feasibility of our proposed framework, which can generate output summaries of alternative styles when given different control codes.

6.2 Automatic Evaluation

We calculated F1 ROUGE scores for generated summaries under 8 control codes, and compared them with the BertScore oracle (see Section 3), the Lead-3 baseline by selecting first-3 sentences as summary, and several competitive extractive models: SummaRuNNer (Nallapati et al., 2017), TransformerEXT and BertEXT (Liu and Lapata, 2019). From Table 2 we observe that: (1) Summary generated from code [0,0,1] is similar to LEAD-3 but can dynamically learn the positional features not limited to the first 3 sentences, while isolating out diversity and importance features. (2) Only focusing on the importance sub-aspect leads to the worst performance, but performance can be improved when considering other sub-aspects. (3) Focusing on the diversity sub-aspect (i.e. Code [0,1,0]) can generate results comparable to strong baselines.

	ROUGE-1	ROUGE-2
Oracle (BertScore)	50.56	29.41
LEAD-3	40.42	17.62
SummaRuNNer*	39.60	16.20
TransformerEXT*	40.90	18.02
BertEXT*	43.23	20.24
Code [0,0,0]	39.44	17.37
Code [0,0,1]	40.21	18.25
Code [0,1,0]	39.18	17.11
Code [0,1,1]	40.70	18.42
Code [1,0,0]	36.72	14.74
Code [1,0,1]	40.33	17.90
Code [1,1,0]	37.59	15.68
Code [1,1,1]	40.87	18.50

Table 2: ROUGE F1 score evaluation with various control codes, in the form of *[importance, diversity, position]*. * denotes the results from corresponding paper.

6.3 Human Evaluation

In addition to automatic evaluation, the human evaluation was conducted by experienced linguistic analysts using Best-Worst Scaling (Louviere et al., 2015). Analysts were given 50 news articles randomly chosen from the CNN/Daily Mail test set and the corresponding summaries from 6 systems: the oracle, BertEXT, three codes disabling sub-aspect position, and one code enabling position. They were asked to decide the best and the worst summaries for each document in terms of informativeness and coherence (Radev et al., 2002; Narayan et al., 2018a). We collected judgments from 5 human evaluators for each comparison. For each evaluator, the documents were randomized differently. The order of summaries for each document was also shuffled differently for each evaluator. The score of a model was calculated as the percentage of times it was labeled as *best* minus the percentage of times it was labeled as *worst*, ranging from -1.0 to 1.0. Since these differences come in pairs, the sum of all the evaluation scores for all summary types adds up to zero. We observed that

	Evaluation Score
Oracle	0.0458
BertEXT	0.0332
Code [1,0,0]	-0.062
Code [0,1,0]	0.0198
Code [0,0,1]	-0.071
Code [1,1,0]	0.0350

Table 3: Human evaluation on samples from baselines and our model with control codes, in the form of *[importance, diversity, position]*.

	ROUGE-1	ROUGE-2
BertEXT	36.78 (-6.45)	14.95 (-5.29)
Code [1,0,0]	33.94 (-2.78)	13.04 (-1.70)
Code [0,1,0]	36.59 (-2.59)	14.33 (-2.78)
Code [0,0,1]	30.34 (-9.87)	8.90 (-9.35)

Table 4: Inference scores on samples with shuffled sentences. Control codes are in the form of *[importance, diversity, position]*. Values in brackets: absolute decrease from scores on original in-order samples.

summaries under diversity code are more favored than those under importance, and their combination can further produce better results (see Table 3). These findings resonate those from the automatic evaluation, suggesting that whether the evaluation metric is lexical overlap (ROUGE) or human judgement, the *diversity* sub-aspect plays a more salient role than *importance*. Moreover, both automatic and human evaluations show that summarizing with semantic-related sub-aspect condition codes achieves reasonable summaries. Examples in Appendix D show that generated summaries are not position-biased yet still preserve key information from the source content.

6.4 Inference on Samples of Shuffled Sentences

To further assess the decoupling between using sub-aspect signals and positional information learned by the model, we conducted an experiment on samples with shuffled sentences, similar to document shuffle in (Kedzie et al., 2018). In our setting, we only introduce the shuffle process in the model inference phase. We shuffled the sentences of all test samples we used in Section 6.2, then applied the well-trained model to generate the predicted summaries. As shown in Table 4, outputs under position sub-aspect and BertEXT suffer a significant drop in performance when we shuffle the sentence order. By comparison, there is far less decrease between the shuffled and in-order samples under diversity and importance control code, demonstrating that the latent features of these two

	R-1 F1	R-2 F1	R-2 Recall
Oracle	-	-	8.70*
Baseline	-	-	6.10*
BertEXT	26.91	3.70	2.98
Code [1,0,0]	34.81	6.23	6.34
Code [0,1,0]	31.79	5.32	4.62
Code [0,0,1]	29.67	3.98	3.47

Table 5: Inference scores on AMI corpus from baselines and our model with control codes, in the form of *[importance, diversity, position]*. * denotes results from (Kedzie et al., 2018).

semantic-related sub-aspects rely less on the position information, suggesting that applying semantic sub-aspects in the training process can reduce systemic bias learned by the model on a corpus with strong position preference.

6.5 Inference on AMI Meeting Corpus

We also conducted an inference experiment on a less position-biased corpus. The AMI corpus (Carletta et al., 2005) is a collection of meetings annotated with text transcriptions with human-written summaries. Different from news summarization, meeting summaries are abstractive with extracted keywords. Unlike the previous comparison work in (Kedzie et al., 2018), we did not train the model from scratch with the AMI training set. Instead, we only applied the pre-trained model (without any fine-tuning) in Section 6 for summarization inference on its test set (20 meeting transcript-summary pairs). Table 5 shows summaries under importance code obtain the highest ROUGE-1 and ROUGE-2 scores, better than the best-reported model in (Kedzie et al., 2018). Not surprisingly, summaries under the position code do not perform well, as there is less position bias in AMI. These findings suggest that our models with semantic-related control codes generalize across domains.

7 Conclusion

We proposed a neural framework for conditional extractive news summarization. In particular, sub-aspect functions of *importance*, *diversity* and *position* are used to condition summary generation. This framework enables us to reduce position bias, a long-standing problem in news summarization, in generated summaries while preserving comparable performance with other standard models. Moreover, our results suggest that with conditional learning, summaries can be more efficiently tailored to different user preferences and application needs.

Acknowledgments

This research was supported by funding from the Institute for Infocomm Research (I2R) under A*STAR ARES, Singapore. We thank Ai Ti Aw, Bin Chen, Shen Tat Goh, Ridong Jiang, Jung Jae Kim, Ee Ping Ong, and Zeng Zeng at I2R for insightful discussions. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

C Bradford Barber, David P Dobkin, David P Dobkin, and Hannu Huhdanpaa. 1996. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. ACM.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6.

Tsutomu Hirao, Masaaki Nishino, Yasuhisa Yoshida, Jun Suzuki, Norihito Yasuda, and Masaaki Nagata. 2015. Summarizing a document by trimming the discourse tree. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(11):2081–2092.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721, Gothenburg, Sweden. Association for Computational Linguistics.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3315–3326, Hong Kong, China. Association for Computational Linguistics.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, page 68–73, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 1997. Identifying topics by position. In *Fifth Conference on Applied Natural Language Processing*, pages 283–290, Washington, DC, USA. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2012. Learning mixtures of submodular shells with application to document summarization. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pages 479–490, Arlington, Virginia, United States. AUAI Press.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731, Hong Kong, China. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2019. Exploiting discourse-level segmentation for extractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 116–121, Hong Kong, China. Association for Computational Linguistics.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

I. Mani. 2001. Summarization evaluation: An overview. In *ACL/EACL-97 summarization workshop*.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Intelligent Scalable Text Summarization*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2).

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

GJ Rath, A Resnick, and TR Savage. 1961. The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. *American Documentation*, 12(2):139–141.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic text structuring and summarization. *Information processing & management*, 33(2):193–207.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Christopher. Scanlan. 1999. *Reporting and writing: basics for the 21st century*. Oxford University Press.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

James Glen Stovall. 1985. *Writing for the mass media*. Prentice-Hall.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4602–4608, Florence, Italy. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ying-Lang Chang and J. Chien. 2009. Latent dirichlet learning for document summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1689–1692.

Dani Yogatama, Fei Liu, and Noah A. Smith. 2015. Extractive summarization by maximizing semantic volume. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1961–1966, Lisbon, Portugal. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.