Inducing Document Structure for Aspect-based Summarization

Lea Frermann
Amazon
uedi@frermann.de

Alexandre Klementiev
Amazon
klementa@amazon.com

Abstract

Automatic summarization is typically treated as a 1-to-1 mapping from document to summary. Documents such as news articles, however, are structured and often cover multiple topics or aspects; and readers may be interested in only some of them. We tackle the task of aspect-based summarization, where, given a document and a target aspect, our models generate a summary centered around the aspect. We induce latent document structure jointly with an abstractive summarization objective, and train our models in a scalable synthetic setup. In addition to improvements in summarization over topic-agnostic baselines, we demonstrate the benefit of the learnt document structure: we show that our models (a) learn to accurately segment documents by aspect; (b) can leverage the structure to produce both abstractive and extractive aspect-based summaries; and (c) that structure is particularly advantageous for summarizing long documents. All results transfer from synthetic training documents to natural news articles from CNN/Daily Mail and RCV1.

1 Introduction

Abstractive summarization systems typically treat documents as unstructured, and generate a single generic summary per document (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017). In this work we argue that incorporating document structure into abstractive summarization systems is beneficial for at least three reasons. First, the induced structure increases model interpretability, and can be leveraged for other purposes such as document segmentation. Second, structure-aware models help alleviate performance bottlenecks associated with summarization of long documents by learning to focus only on the segments relevant to the topic of interest. Third, they can adapt more flexibly to demands of a user who, faced with a long document or a document collection, might be interested only in some of its topics.

For example given a set of reviews of a smartphone, one user might be interested in a summary of opinions on `battery life` while another may care more about its `camera quality`; or, given a news article about a body builder running for governour, a reader might care about the effect on his `sports` career, or on the `political` consequences (cf., Figure 1 (bottom) for another example). Throughout this paper, we will refer to such topics or perspectives collectively as *aspects*. We develop models for aspect-based summarization: given a document and a target aspect, our systems generate a summary specific to the aspect.

We extend recent neural models (See et al., 2017) for abstractive summarization making the following contributions:

- We propose and compare models for aspect-based summarization incorporating different aspect-driven attention mechanisms in both the encoder and the decoder.
- We propose a scalable synthetic training setup and show that our models generalize from synthetic to natural documents, sidestepping the data sparsity problem and outperforming recent aspect-agnostic summarization models in both cases.
- We show that our models induce meaningful latent structure, which allows them to generate abstractive and extractive aspect-driven summaries, segment documents by aspect, and generalize to long documents.[1] We argue that associating model attention with aspects also improves model interpretability.

[1] A well-known weakness of encoder-decoder summarization models (Vaswani et al., 2017; Cohan et al., 2018)

Our models are trained on documents paired with aspect-specific summaries. A sizable data set does not exist, and we adopt a scalable, synthetic training setup (Choi, 2000; Krishna and Srinivasan, 2018). We leverage aspect labels (such as `news` or `health`) associated with each article in the CNN/Daily Mail dataset (Hermann et al., 2015), and construct synthetic multi-aspect documents by interleaving paragraphs of articles pertaining to different aspects, and pairing them with the original summary of one of the included articles. Although assuming one aspect per source article may seem crude, we demonstrate that our model trained on this data picks up subtle aspect changes within natural news articles. Importantly, our setup requires no supervision such as pre-trained topics (Krishna and Srinivasan, 2018) or aspect-segmentation of documents. A script to reproduce the synthetic data set presented in this paper can be found at `www.TODO.com`.

Our evaluation shows that the generated summaries are more aspect-relevant and meaningful compared to aspect agnostic baselines, as well as a variety of advantages of the inferred latent aspect representations such as accurate document segmentation, that our models produce both extractive and abstractive summaries of high quality, and that they do so for long documents. We also show that our models, trained on synthetic documents, generalize to natural documents from the Reuters and the CNN/Daily Mail corpus, through both automatic and human evaluation.

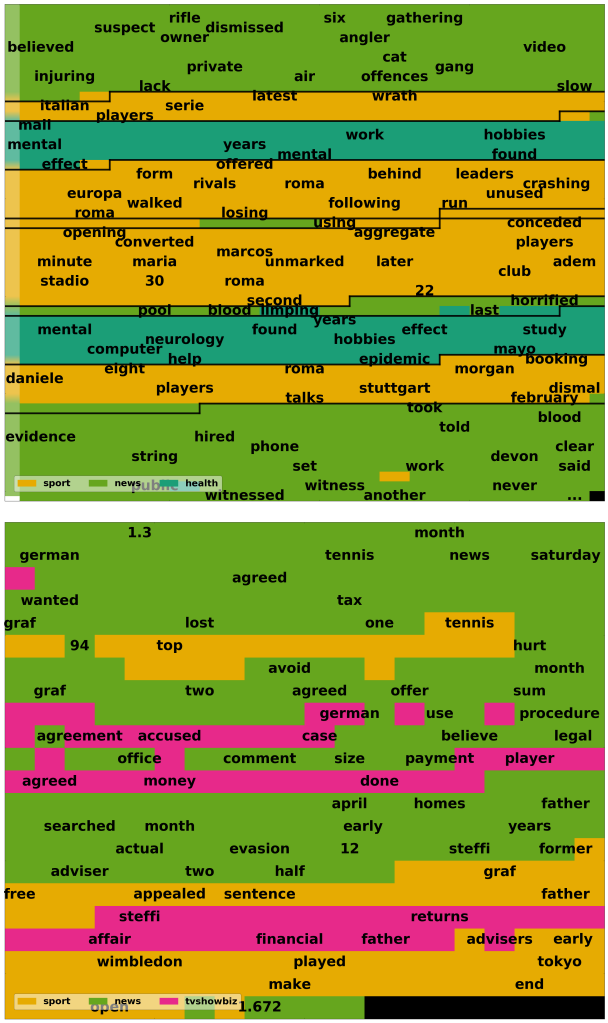
2 Related Work

Aspect-based summarization has previously been considered in the customer feedback domain (Hu and Liu, 2004; Zhuang et al., 2006; Titov and McDonald, 2008; Lu et al., 2009; Zhu et al., 2009), where a typical system discovers a set of relevant aspects (product properties), and extracts sentiment and information along those aspects. In contrast, we induce latent aspect representations under an abstractive summarization objective. Gerani et al. (2016) consider discourse and topical structure to abstractively summarize product reviews using a micro planning pipeline for text generation rather than building on recent advances in end-to-end modeling. Yang et al. (2018) propose an aspect- and sentiment-aware neural summarization model in a multi-task learning setup. Their model is geared towards the product domain and requires document-level category labels, and sentiment- and aspect lexica.

In *query*-based summarization sets of documents are summarized with respect to a natural language input query (Dang, 2005; Daumé III and Marcu, 2006; Mohamed and Rajasekaran, 2006; Liu et al., 2012; Wang et al., 2014; Baumel et al., 2018). Our systems generate summaries with respect to abstract input aspects (akin to topics in a topic model), whose representations are learnt jointly with the summarization task.

We build on neural encoder-decoder architectures with attention (Nallapati et al., 2016; Cheng and Lapata, 2016; Chopra et al., 2016; See et al., 2017; Narayan et al., 2017), and extend the pointer-generator architecture of See et al. (2017) to our task of aspect-specific summarization. Narayan et al. (2018) use topic information from a pre-trained LDA topic model to generate ultra-short (single-topic) summaries, by scoring words in their relevance to the overall document. We learn topics jointly within the summarization system, and use them to directly drive summary content selection.

Our work is most related to Krishna and Srinivasan (2018) (KS), who concurrently developed models for topic-oriented summarization in the context of artificial documents from the CNN/Daily Mail data. Our work differs from theirs in several important ways. KS use pointer-generator networks directly, whereas we develop novel architectures involving aspect-driven attention mechanisms (Section 3). As such, we can analyze the representations learnt by different attention mechanisms, whereas KS re-purpose attention which was designed with a different objective (coverage). KS use pre-trained topics to pre-select articles from CNN/Daily Mail whose summaries are highly separable in topic space, whereas we do not require such resources nor do we pre-select our data, resulting in a simpler and more realistic setup (Section 4). In addition, our synthetic data set is more complex (ours: 1-4 aspects per document, selected from a set of 6 global aspects; KS: 2 aspects per document, unknown total number of aspects). We extensively evaluate the benefit of latent document structure (Sections 5.1–5.3), and apply our method to human-labeled multi-aspect news documents from the Reuters corpus (Section 5.4).



Synthetic multi-aspect news article from the MT-news corpus

a father spent #10,000 on private detectives after police failed to track down the thug who killed his daughter's kitten with an air rifle. neil tregarthen devoted six weeks to gathering information [...] ‖roma 's players were the latest to face the wrath of angry fans following thursday 's capitulation against italian rivals fiorentina. serie a pundit matteo bonetti tweeted that [...] ‖having a demanding job can help stave off dementia later in life, a study has found. keeping your brain active throughout your lifetime, both at work and by enjoying stimulating hobbies, can delay mental decline [...] ‖the europa league had offered the last realistic chance [...] ‖the most beneficial hobbies included reading, having an active social life and using a computer regularly. ‖aylish was horrified to find farah lying in a pool of blood after limping home wounded and crippled with pain last september. [...]

news neil tregarthen spent #10,000 on private detectives after police failed to track down thug who killed the thug who killed his daughter's kitten with an air rifle. the kitten was shot near her owner's home in exeter.

health having a demanding job can help stave off dementia later in life. doctors have long said training your brain in later years can prevent dementia but but this is the first time mental activity earlier in life.

sport mats francesco totti also spoke to fans despite being an unused substitute. roma captain francesco totti also spoke to fans despite being an unused substitute.

News article from the Reuters RCV1 corpus

steffi graf reluctantly paid 1.3 million marks to charity last month as part of a settlement with german prosecutors who dropped their tax evasion investigation [...] spiegel magazine said graf had 'agreed with a heavy heart' to the bargain with prosecutors because she wanted to put the 'media circus' about her tax affairs behind her and concentrate on tennis. [...] prosecutors dropped their investigation last month after probing graf's finances for nearly two years when she agreed to their offer to pay a sum to charity. [...] german prosecutors often use the charity donation procedure , with the agreement of the accused, to end a case which they do not believe merits a lengthy legal process. [...] the seven-times wimbledon champion, who has not played since the semifinals [...]

sport seven-times wimbledon champion could make a return to the court at the end of april in the german open . former family tax adviser joachim eckardt received two and a half years for complicity .

news prosecutors dropped their investigation last month after probing graf 's finances for nearly two years when she agreed to their offer to pay a sum to charity last month as part of a settlement with german prosecutors who dropped their tax evasion investigation of the tennis player , a news magazine

tvshowbiz steffi graf reluctantly paid 1.3 million marks $ 777,000) to charity last month as part of a settlement with german prosecutors who dropped their tax evasion investigation of the tennis player . the player said she had entrusted financial matters to her father and his advisers from an early age .

Figure 1: Two news articles with color-coded encoder attention-based document segmentations, and selected words for illustration (left), the abridged news article (top right) and associated aspect-specific model summaries (bottom right). **Top:** Article from our synthetic corpus with aspects `sport`, `tvshowbiz` and `health`. The true boundaries are known, and indicated by black lines in the plot and ‖in the article. **Bottom:** Article from the RCV1 corpus with document-level human-labeled aspects `sports`, `news` and `tvshowbiz` (gold segmentation unknown).

3 Aspect-specific Summarization

In this section we formalize the task of aspect-specific document summarization, and present our models. Given an input document x and a target aspect a, our model produces a summary of x with respect to a such that the summary (i) contains *only* information relevant to a; and (ii) states this information in a concise way (cf., examples in Figure 1).

Our model builds on the pointer-generator networks (PG-net; See et al. (2017)), an encoder-decoder architecture for abstractive summarization. Unlike traditional document summarization, a model for aspect-based summarization needs to include aspects in its input document representation in order to select and compress relevant information. We propose three extensions to PG-net which allow the resulting model to learn to detect aspects. We begin by describing PG-net before we describe our extensions. Our models are trained on documents paired with aspect-specific summaries (cf., Section 4). Importantly, all proposed extensions treat aspect segmentation as latent, and as such learn to segment documents by aspects without exposure to word- or sentence-level aspect labels at train time. Figure 2 visualizes our models.

PG-net. PG-net (See et al., 2017) is an encoder-decoder abstractive summarization model, consisting of two recurrent neural networks. The encoder network is a bi-directional LSTM which reads in the article $x = \{w_i\}_1^N$, token by token, and produces a sequence of hidden states

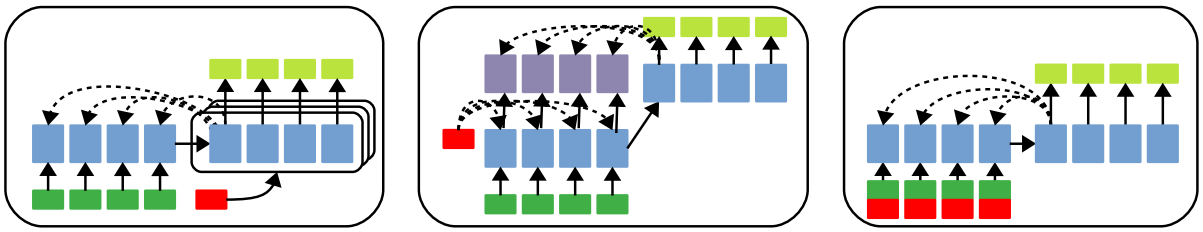


Figure 2: Visualization of our three aspect-aware summarization models, showing the embedded input aspect (red), word embeddings (green), latent encoder and decoder states (blue) and attention mechanisms (dotted arrows). **Left**: the decoder aspect attention model; **Center**: the encoder attention model; **Right**: the source-factors model.

$h = \{h_i\}_1^N$. This sequence is accessed by the decoder network, also an LSTM, which incrementally produces a summary, by sequentially emitting words. At each step t the decoder produces word y_t conditioned on the previously produced word y_{t-1}, its own latent LSTM state s_t and a time-specific representation of the encoder states h_t^*. This time-specific representation is computed through Bahdanau attention (Bahdanau et al., 2015) over the encoder states,

$$e_i^t = v^T tanh(W_h h_i + W_s s_t + b) \quad (1)$$

$$a^t = softmax(e^t) \quad (2)$$

$$h_t^* = \sum_i a_i^t h_i, \quad (3)$$

where v, W_h, W_s and b are model parameters. Given this information, the decoder learns to either generate a word from a fixed vocabulary or copy a word from the input. This procedure is repeated until either the maximum output sequence length is reached, or a special $< STOP >$ symbol is produced.[2]

Loss. The loss of PG-net, and all proposed extensions, is the average negative log-likelihood of all words in the summary

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} -logP(w_t) \quad (4)$$

3.1 Aspect-aware summarization models

Our proposed models embed all words $\{w\} \in x$ into a latent space, shared between the encoder and the decoder. We also embed the input aspect a (a 1-hot indicator) into the same latent space, treating aspects as additional items of the vocabulary. The embedding space is randomly initialized and updated during training.

[2]A coverage mechanism was proposed with PG-net to avoid repetition in the summary. However, in order to minimize interaction with the aspect-attention mechanisms we propose, we do not include it in our models.

Decoder aspect attention. As a first extension, we modify the decoder attention mechanism to depend on the target summary aspect a (Figure 2, left). To this end, we learn separate attention weights and biases for each possible input aspect, and use the parameters specific to target-aspect a during decoding, replacing equation (1) with

$$e_i^t = v^T tanh(W_h^a h_i + W_s^a s_t + b^a). \quad (5)$$

Intuitively, the model can now focus on parts of the input not only conditioned on its current decoder state, but also depending on the aspect the summary should reflect.

Encoder attention. Intuitively, all information about aspects is present in the input, independently of the summarization mechanism, and as such should be accurately reflected in the latent document representation. We formalize this intuition by adding an attention mechanism to the encoder (Figure 2, center). After LSTM encoding, we attend over the LSTM states $h = \{h_i\}_1^N$ conditioned on the target aspect as follows

$$\tilde{a}_i = tanh(W_{\tilde{a}} h_i + b_{\tilde{a}}) \quad (6)$$

$$a = sigmoid(e_a^T \tilde{a}) \quad (7)$$

$$h_i' = a h_i, \quad (8)$$

where $W_{\tilde{a}}$ and $b_{\tilde{a}}$ are parameters, and e_a is the embedded target aspect. The decoder will now attend over h' instead of h in equations (1)-(3). Intuitively, we calculate a weight for each token-specific latent representation, and scale each latent representation independently by passing the weight through a sigmoid function. Words irrelevant to aspect a should be scaled down by the sigmoid transformation.

Source-factors. Our final extension uses the original PG-net, and modifies its input by treating the target aspect as additional information (factor),

which gets appended to our input document (Figure 2, right).[3] We concatenate the aspect embedding e^a to the embedding of each word $w_i \in x$. The *target summary* aspect, not the word's true aspect (which is latent and unknown), is utilized. Through the lexical signal from the target summary, we expect the model to learn to up- or downscale the latent token representations, depending on whether they are relevant to target aspect a. Note that this model does not provide us with aspect-driven attention, and as such cannot be used for document segmentation.

4 A Multi-Aspect News Dataset

To train and evaluate our models, we require a data set of documents paired with aspect-specific summaries. Several summarization datasets consisting of long and multifaceted documents have been proposed recently (Cohan et al., 2018; Liu et al., 2018). These datasets do not include aspect-specific summaries, however, and as such are not applicable to our problem setting.

We synthesize a dataset fulfilling our requirements from the CNN/Daily Mail (CNN/DM) dataset (Hermann et al., 2015). Our dataset, MA-News, is a set $\mathcal{D}$ of data points $d = (x, y, a)$, where x is a multi-aspect document, a is an aspect in d, and y is a summary of x wrt. aspect a. We assemble synthetic multi-aspect documents, leveraging the article-summary pairs from the CNN/DM corpus, as well as the URL associated with each article, which indicates its topic category. We select six categories as our target aspects, optimizing for diversity and sufficient coverage in the CNN/DM corpus: $\mathcal{A} = \{$ `tvshowbiz`, `travel`, `health`, `sciencetech`, `sports`, `news`$\}$.

We then create multi-aspect documents by interleaving paragraphs of documents belonging to different aspects. For each document d, we first sample its number of aspects $n_d \sim U(1, 4)$. Then, we sample n_d aspects from $\mathcal{A}$ without replacement, and randomly draw a document for each aspect from the CNN/DM corpus.[4] We randomly interleave paragraphs of the documents, maintaining each input document's chronological order. Since paragraphs are not marked in the input data, we draw paragraph length between 1 and 5 sentences. The six aspects are roughly uniformly distributed in the resulting dataset, and the distribution of number of aspects per document is slightly skewed towards more aspects.[5]

Finally, we create n_d data points from the resulting document, by pairing the document once with each of its n_d components' reference summaries. We construct 284,701 documents for training and use 1,000 documents each for validation and test.

In order to keep training and evaluation fast, we only consider CNN/DM documents of length 1000 words or less, and restrict the length of assembled MA-News documents to up to 1500 words. Note that the average MA-News article (1350 words) is longer than CNN/DM (770 words), increasing the difficulty of the summarization task, and emphasizing the importance of learning a good segmentation model, which allows the summarizer to focus on relevant parts of the input. We present evidence for this in Section 5.3.

5 Evaluation

This section evaluates whether our models generate concise, aspect-relevant summaries for synthetic multi-aspect documents (Section 5.1), as well as natural documents (Sections 5.3, 5.4). We additionally explore the quality of the induced latent aspect structure, by (a) evaluating our models on document segmentation (Section 5.2), and (b) demonstrating the benefit of structure for summarizing *long* natural documents (Section 5.3).

Model parameters. We extend the implementation of pointer-generator networks[6], and use their training parameters. We set the maximum encoder steps to 2000 because our interleaved training and test documents are longer on average than the original CNN/DM articles. We use the development set for early stopping. We do not use coverage (See et al., 2017) in any of our models to minimize interaction with the aspect-attention mechanisms. We also evaluated systems trained with all combinations of our three aspect-awareness mechanisms, but we did not observe systematic improvements over the single-mechanism systems. Hence, we will only report results on those.

5.1 Summarization

This section evaluates the quality of produced summaries using the ROUGE metric (Lin, 2004).

[3] This model most closely resembles the model presented in (Krishna and Srinivasan, 2018), who append 1-hot topic indicators to each word in the input.

[4] Train, validation and test documents are assembled from non-overlapping sets of articles.

[5] # aspects/proportion: 1/0.107, 2/0.203, 3/0.297, 4/0.393

[6] https://github.com/abisee/pointer-generator

Model Comparison. We compare the aspect-aware models with decoder aspect attention (dec-attn), encoder attention (enc-attn), and source factors (sf) we introduced in Section 3.1 against a baseline which extracts a summary as the first three sentences in the article (lead-3). We expect any lead-n baseline to be weaker for aspect-specific summarization than for classical summarization, where the first n sentences typically provide a good *generic* summary. We also apply the original pointer-generator network (PG-net), which is aspect-agnostic. In addition to the abstractive summarization setup, we also derive extractive summaries from the aspect-based attention distributions of two of our models (enc-attn-extract and dec-attn-extract). We iteratively extract sentences in the input which received the highest attention until a maximum length of 100 words (same threshold as for abstractive) is reached. Sentence attention a_s is computed as average word attention a_w for words in s: $a_s = \frac{1}{|s|} \sum_{w \in s} a_w$. Finally, as an upper bound, we train our models on the subset of the original CNN/DM documents from which the MA-News documents were created (prefixed with ub-).

Table 1 (top) presents results of models trained and tested on the synthetic multi-aspect dataset. All aspect-aware models beat both baselines by a large margin. For classical summarization, the lead-3 baseline remains a challenge to beat even by state-of-the-art systems, and also on multi-aspect documents we observe that, unlike our systems, PG-net performs worse than lead-3. Unsurprisingly, the extractive aspect-aware models outperform their abstractive counterparts in terms of ROUGE, and the decoder attention distributions are more amenable to extraction than encoder attention scores. Overall, our structured models enable both abstractive and extractive aspect-aware summarization at a quality clearly exceeding structure-agnostic baselines.

To assess the impact of the synthetic multi-aspect setup, we apply all models to the original CNN/DM documents from which MA-news was assembled (Table 1, bottom). Both baselines show a substantial performance boost, suggesting that they are well-suited for general summarization but do not generalize well to aspect-based summarization. The performance of our own models degrades more gracefully. Note that some of our aspect-aware methods outperform the PG-net on natural documents, showig that our models can pick up and leverage their less pronounced structure (compared to synthetic documents) as well. Aspect-based summarization requires models to leverage topical document structure to produce relevant summaries, and as such a baseline focusing on the beginning of the article, which typically summarizes its main content, is no longer viable.

	rouge 1	rouge 2	rouge L
lead-3	0.2150	0.0690	0.1410
PG-net	0.1757	0.0472	0.1594
enc-attn	0.2750	0.1027	0.2502
dec-attn	0.2734	0.1005	0.2509
sf	0.2802	0.1046	0.2536
enc-attn-extract	0.3033	0.1092	0.2732
dec-attn-extract	0.3326	0.1379	0.3026
ub-lead-3	0.3836	0.1765	0.2468
ub-PG-net	0.3446	0.1495	0.3159
ub-enc-attn	0.3603	0.1592	0.3282
ub-dec-attn	0.3337	0.1427	0.3039
ub-sf	0.3547	0.1570	0.3262

Table 1: Quantitative comparison (ROUGE 1, 2 and L) of models on aspect-specific summarization.

5.2 Segmentation

The model attention distribution over the input document, conditioned on a target aspect, allows us to qualitatively inspect the model's aspect representation, and to derive a document segmentation. Since we know the true aspect segmentations for documents in our synthetic dataset, we can evaluate our models on this task, using all test documents with > 1 aspect (896 in total). We decode each test document multiple times conditioned on each of its aspects, and use the attention distributions over the input document under different target aspects to derive a document segmentation. Figure 1 visualizes induced segmentations of two documents. We omit the source-factor model in this evaluation, because it does not provide us with a latent document representation.

For the encoder attention model, we obtain n_d attention distributions (one per input aspect), and assign each word the aspect under which it received highest attention. For the decoder aspect attention model, we obtain $n_d \times T$ attention distributions, one for each decoder step t and input aspect. For each aspect we assign each word the maximum attention it received over the T decoder

model	P_k	WD	acc w	acc s	ratio
global-max	0.694	0.694	0.138	0.142	10.8
sent-max	0.694	0.694	0.474	0.503	10.8
word-max	0.694	0.694	0.487	0.488	10.8
Considering only aspects $\in$ input x					
LDA	0.375	0.789	0.294	0.282	0.722
MNB	**0.223**	0.594	0.753	0.732	0.553
enc-attn	0.270	**0.348**	**0.793**	**0.784**	**0.784**
dec-attn	0.285	0.385	0.727	0.780	0.697
Considering all global aspects $\in \mathcal{A}$					
LDA	0.590	0.697	0.250	0.204	3.725
MNB	0.268	0.784	0.591	0.564	0.398
enc-attn	0.337	0.482	0.667	0.663	0.580
dec-attn	0.454	0.708	0.385	0.424	0.374

Table 2: Text segmentation results: Segmentation metrics P_k and windiff (WD; lower is better), aspect label accuracies (acc w, acc s), and the ratio of system to summary segments (ratio). Three majority baselines (global-max, word-max, sent-max), and a topic model (LDA) and classification baseline (MNB). The majority baselines assign the same aspect to all words (sentences) in a doc, so that P_k and WD scores are identical.

steps.[7] Since our gold standard provides us with sentence-level aspect labels, we derive sentence-level aspect labels as the most prevalent word-level aspect in the sentence.

Baselines. global-max assigns each word to the globally most prevalent aspect in the corpus. A second baseline assigns each word to the document's most prevalent aspect on word- (word-max) or sentence level (sent-max). An unsupervised topic model baseline (LDA) is trained on the training portion of our synthetic data set ($K = 6$; topics were mapped manually to aspects). At decode time, we assign each word its most likely topic and derive sentence labels as the topic assigned to most of its words. Finally, a supervised classification baseline (multinomial naive bayes; MNB) is trained to classify sentences into aspects.

Metrics. We either consider the set of aspects present in a document (Table 2 center) or all possible aspects in the data set (Table 2 bottom). We measure traditional segmentation metrics P_k (Beeferman et al., 1999) and windiff (WD; Pevzner and Hearst (2002)) (lower is better) which estimate the accuracy of segmentation *boundaries*, but do not evaluate whether a *correct* aspect has been assigned to any segment. Hence, we also include aspect label accuracy on the word level (acc w) and sentence level (acc s) (higher is better). We also compute the ratio of the true number of segments to the predicted number of segments (ratio).

[7] We also experimented with mean instead of max, but observed very similar results.

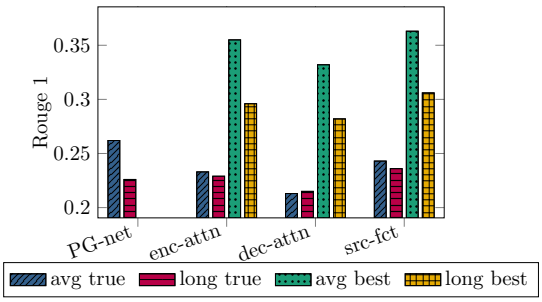


Figure 3: Models trained on synthetic data evaluated on original CNN/DM documents, of either <1000 words (short) or >2000 words (long). True uses the summary under the document's true aspect. 'Best' takes the best-scoring summary under all possible input aspects.

The attention-aware summarization models outperform all baselines across the board (Table 2). LDA outperforms the most basic global-max baseline, but not the more informed per-document majority baselines. Unsurprisingly, MNB as a supervised model trained specifically to classify sentences performs competitively. Overall, the performance drops when considering the larger set of all six aspects (bottom) compared to only aspects present in the document (between 2 and 4; center).

5.3 Long Documents

Accurately encoding long documents is a known challenge for encoder-decoder models. We hypothesize that access to a structured intermediate document representation would help alleviate this issue. To this end, we compare our models against the aspect-agnostic PG-net on *natural* average and long documents from CNN/DM. All models are trained on the multi-aspect data set. We construct two test datasets: (i) the CNN/DM documents underlying our test set (up to 1000 words; `avg`), and (ii) CNN/DM documents which are at least 2000 words long (`long`) and are tagged with one of our target aspects. The total number of average and long documents is 527 and 4560, respectively.

Results (Figure 3) confirm that our aspect-aware models indeed degrade more gracefully in performance when applied to long documents, and that the source-factor model (R1=0.236) outperforms the PG (R1=0.226) model by one ROUGE point on long documents (red bars).

We finally explore our aspect-aware models on the task of aspect-agnostic summarization, decoding test documents under all possible aspects, and

rand	max	LDA	MNB	enc-attn	dec-attn
0.34	0.71	0.40	0.53	**0.75**	0.37

Table 3: Sentence labelling accuracy of aspects present in the input article.

selected the aspect with the highest-scoring summary in terms of ROUGE (`avg best` and `long best`, respectively). In this setup, all our models outperform the PG-baseline by a large margin, both on long and average documents.

5.4 Evaluation on Reuters News

Finally, we evaluate our models on documents with multiple gold-annotated aspects, using the Reuters RCV1 dataset (Lewis et al., 2004). Our target aspects `sport`, `health`, `sciencetech` and `travel` are identically annotated in the Reuters data set. We map the remaining tags `tvshowbiz` and `news` to their most relevant Reuters counterparts.[8] We obtain 792 document (with average length of 12.2 sentences), which were labeled with two or more aspects. Figure 1 (bottom) shows an example of generated summaries for a multi-aspect Reuters document.

Automatic evaluation. We evaluate how well our models recover aspects actually present in the documents. We use the approach described in Section 5.2 to assign aspects to sentences in a document, then collect all of the aspects we discover in each document. We compare aspect to document assignment accuracy against two baselines, one assigning random aspects to sentences (rand), and one always assigning the globally most prominent aspect in the corpus (max). Note that we do not include PG-net or the source-factor model because neither can assign aspects to input tokens.

Table 3 shows that the encoder attention model outperforms all other systems and both baselines. The global majority baseline shows that the gold aspect distribution in the RCV1 corpus is peaked (the most frequent aspect, `news`, occurs in about 70% of the test documents), and majority class assignment leads to a strong baseline.

Human evaluation. We measure the quality and aspect diversity in aspect-specific summaries of RCV1 articles through human evaluation, using Amazon Mechanical Turk. We randomly select a subset of 50 articles with at least two aspects from the Reuters RCV1 data, and present Turkers with a news article and two summaries. We ask the Turkers to (1) select a topic for each summary from the set of six target topics;[9]; (2) rate the summary with respect to its fluency (0=not fluent, 1=somewhat fluent, 2=very fluent); and (3) analogously rate its informativeness.

We evaluate the extractive and abstractive versions of our three aspect aware models. We do not include the original PG-net, because it is incapable of producing distinct, aspect-conditioned summaries for the same document. Like in our automatic summarization evaluation we inlcude a *lead* baseline. Since the annotators are presented with two summaries for each article, we adopt a lead-2 baseline, and present the first two sentences of a document as a summary each (lead-2). This baseline has two advantages over our systems: first, it extracts summaries as single, complete sentences which are typically semantically coherent units; second, the two sentences (i.e., summaries) do not naturally map to a gold aspect each. We consider both mappings, and score the best.

Results are displayed in Table 4. As expected, the extractive models score higher on fluency, and consequently on aspect-agnostic informativeness. Our abstractive models, however, outperform all other systems in terms of aspect-labeling accuracy (acc), and annotators more frequently assign *distinct* aspects to two summaries of an article (diversity). The results corroborate our conclusion that the proposed aspect-aware summarization models produce summaries aspect-focussed summaries with and distinguishable and human

model	acc	diversity	fluency	info
lead-2	0.540	0.127	**1.930**	**1.647**
enc-attn	0.543	0.177*	1.567	1.317
enc-attn ex	0.436	0.129	1.924	1.367
dec-attn	**0.553***	**0.197***	1.447	1.277
dec-attn ex	0.440	0.151	1.889	1.448
sf	**0.553**	0.133	1.667	1.433

Table 4: Human evaluation: aspect label accuracy (acc), aspect label diversity for two summaries (diversity), and fluency and informativeness (info) scores. Systems performing significantly better than the lead-2 baseline are marked with a * ($p < 0.05$, paired t-test; Dror et al. (2018)).

[8] `tvshowbiz` → `fashion`, `biographies_personalities_people`, `art_culture_entertainment` `news` → `disasters_accidents`, `crime_lawenforcement`, `international_relations`

[9] A random baseline would achieve acc=0.17.

interpretable focus.

6 Conclusions

This paper presented the task of aspect-based summarization, where a system summarizes a document with respect to a given input aspect of interest. We introduced neural models for abstractive, aspect-driven document summarization. Our models induce latent document structure, to identify aspect-relevant sements of the input document. Treating document structure as latent allows for efficient training with no need for subdocument level topic annotations. The latent document structure is induced jointly with the summarization objective.

Sizable datasets of documents paired with aspect-spedific summaries do not exist and are expensive to create. We proposed a scalable synthetic training setup, adapting an existing summarization data set to our task. We demonstrated the benefit of document structure aware models for summarization through a diverse set of evaluations. Document structure was shown to be particularly useful for long documents. Evaluation further showed that models trained on synthetic data generalize to natural test documents.

An interesting challenge, and open research question, concerns the extent to which synthetic training impacts the overall model generalizability. This work's range of considered aspects, and creation of synthetic data by interleaving documents which are maximally distinct with respect to the target aspects leaves room for refinement. Ideas for incorporaing more realistic topic structure in artificial documents include leveraging more fine-grained (or hierarchical) topics in the source data; or adopt a more sophisticated selection of article segments to interleave by controlling for confounding factors like author, time period, or general theme.[10] We believe that training models on heuristic, but inexpensive data sets is a valuable approach which opens up exciting opportunities for future research.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning - Special issue on natural language learning*, 34(1-3):177–210.

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

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 26–33, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Hoa T. Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. Association for Computational Linguistics.

[10] E.g., constructing articles about a fixed theme (Barack Obama) from different aspects (`politics` and `showbiz`).

Shima Gerani, Giuseppe Carenini, and Raymond T Ng. 2016. Modeling content and structure for abstractive review summarization. *Computer Speech & Language*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Kundan Krishna and Balaji Vasan Srinivasan. 2018. Generating topic-oriented summaries using neural attention. 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 1697–1705, New Orleans, Louisiana. Association for Computational Linguistics.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

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.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Yan Liu, Sheng-hua Zhong, and Wenjie Li. 2012. Query-oriented multi-document summarization via unsupervised deep learning. In *AAAI*.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 131–140, New York, NY, USA. ACM.

Ahmed A. S. Mohamed and Sanguthevar Rajasekaran. 2006. Query-based summarization based on document graphs.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

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.

Shashi Narayan, Nikos Papasarantopoulos, Mirella Lapata, and Shay B. Cohen. 2017. Neural extractive summarization with side information. *CoRR*, abs/1704.04530.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Lu Wang, Hema Raghavan, Claire Cardie, and Vittorio Castelli. 2014. Query-focused opinion summarization for user-generated content. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1660–1669.

Min Yang, Qiang Qu, Ying Shen, Qiao Liu, Wei Zhao, and Jia Zhu. 2018. Aspect and sentiment aware abstractive review summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1110–1120, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jingbo Zhu, Muhua Zhu, Huizhen Wang, and Benjamin K. Tsou. 2009. Aspect-based sentence segmentation for sentiment summarization. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 65–72, New York, NY, USA. ACM.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50. ACM.