Earlier Isn't Always Better: Sub-aspect Analysis on Corpus and System Biases in Summarization

Taehee Jung[*♡] **Dongyeop Kang**[*♣] **Lucas Mentch** [♡] **Eduard Hovy** [♣]
[♡]Department of Statistics, University of Pittsburgh, Pittsburgh, PA, USA
[♣]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
{taj41,lkm31}@pitt.edu {dongyeok,hovy}@cs.cmu.edu

Abstract

Despite the recent developments on neural summarization systems, the underlying logic behind the improvements from the systems and its corpus-dependency remains largely unexplored. Position of sentences in the original text, for example, is a well known bias for news summarization. Following in the spirit of the claim that summarization is a combination of sub-functions, we define three sub-aspects of summarization: `position`, `importance`, and `diversity` and conduct an extensive analysis of the biases of each sub-aspect with respect to the domain of nine different summarization corpora (e.g., news, academic papers, meeting minutes, movie script, books, posts). We find that while `position` exhibits substantial bias in news articles, this is not the case, for example, with academic papers and meeting minutes. Furthermore, our empirical study shows that different types of summarization systems (e.g., neural-based) are composed of different degrees of the sub-aspects. Our study provides useful lessons regarding consideration of underlying sub-aspects when collecting a new summarization dataset or developing a new system.

1 Introduction

Despite numerous recent developments in neural summarization systems (Narayan et al., 2018b; Nallapati et al., 2016; See et al., 2017; Kedzie et al., 2018; Gehrmann et al., 2018; Paulus et al., 2017) the underlying rationales behind the improvements and their dependence on the training corpus remain largely unexplored. Edmundson (1969) put forth the position hypothesis: important sentences appear in preferred positions in the document. Lin and Hovy (1997) provide a method to empirically identify such positions. Later, Hong and Nenkova (2014) showed an intentional lead bias in news writing, suggesting that sentences appearing early in news articles are more important for summarization tasks. More generally, it is well known that recent state-of-the-art models (Nallapati et al., 2016; See et al., 2017) are often marginally better than the first-k baseline on single-document news summarization.

In order to address the position bias of news articles, Narayan et al. (2018a) collected a new dataset called XSum to create single sentence summaries that include material from multiple positions in the source document. Kedzie et al. (2018) showed that the position bias in news articles is not the same across other domains such as meeting minutes (Carletta et al., 2005).

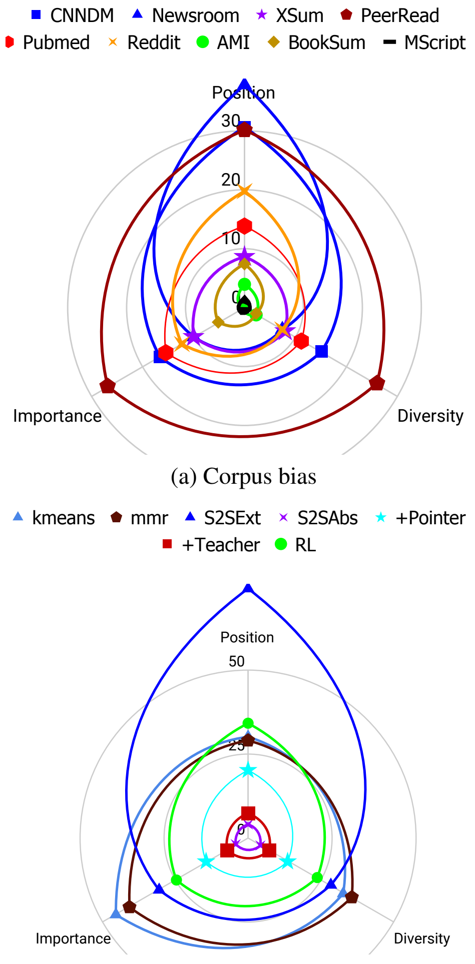
In addition to `position`, Lin and Bilmes (2012) defined other sub-aspect functions of summarization including `coverage`, `diversity`, and `information`. Lin and Bilmes (2011) claim that many existing summarization systems are instances of mixtures of such sub-aspect functions; for example, maximum marginal relevance (MMR) (Carbonell and Goldstein, 1998) can be seen as an combination of diversity and importance functions.

Following the sub-aspect theory, we explore three important aspects of summarization (§3): POSITION for choosing sentences by their position, IMPORTANCE for choosing relevant contents, and DIVERSITY for ensuring minimal redundancy between summary sentences.

We then conduct an in-depth analysis of these aspects over nine different domains of summarization corpora (§5) including news articles, meeting minutes, books, movie scripts, academic papers, and personal posts. For each corpus, we investigate which aspects are most important and develop a notion of **corpus bias** (§6). We provide an empirical result showing how current summarization systems are compounded of which sub-aspect

[*] Equal contribution, name order decided by coin flip.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3324–3335,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics



(a) Corpus bias

(b) System bias

Figure 1: Corpus and system biases with the three sub-aspects, showing what portion of aspect is used for each corpus and each system. The portion is measured by calculating ROUGE score between (a) summaries obtained from each aspect and target summaries or (b) summaries obtained from each aspect and each system.

factors called **system bias** (§7). At last, we summarize our actionable messages for future summarization researches (§8). We summarize some notable findings as follows:

- Summarization of personal post and news articles except for XSum (Narayan et al., 2018a) are biased to the position aspect, while academic papers are well balanced among the three aspects (see Figure 1 (a)). Summarizing long documents (e.g. books and movie scripts) and conversations (e.g. meeting minutes) are extremely difficult tasks that require multiples aspects together.
- Biases do exist in current summarization systems (Figure 1 (b)). Simple ensembling of multiple aspects of systems show comparable performance with simple single-aspect systems.
- Reference summaries in current corpora include less than 15% of new words that do not appear in the source document, except for abstract text of academic papers.
- Semantic volume (Yogatama et al., 2015) overlap between the reference and model summaries is not correlated with the hard evaluation metrics such as ROUGE (Lin, 2004).

2 Related Work

We provide here a brief review of prior work on summarization biases. Lin and Hovy (1997) studied the position hypothesis, especially in the news article writing (Hong and Nenkova, 2014; Narayan et al., 2018a) but not in other domains such as conversations (Kedzie et al., 2018). Narayan et al. (2018a) collected a new corpus to address the bias by compressing multiple contents of source document in the single target summary. In the bias analysis of systems, Lin and Bilmes (2012, 2011) studied the sub-aspect hypothesis of summarization systems. Our study extends the hypothesis to various corpora as well as systems. With a specific focus on `importance` aspect, a recent work (Peyrard, 2019a) divided it into three sub-categories; redundancy, relevance, and informativeness, and provided quantities of each to measure. Compared to this, ours provide broader scale of sub-aspect analysis across various corpora and systems.

We analyze the sub-aspects on different domains of summarization corpora: news articles (Nallapati et al., 2016; Grusky et al., 2018; Narayan et al., 2018a), academic papers or journals (Kang et al., 2018; Kedzie et al., 2018), movie scripts (Gorinski and Lapata, 2015), books (Mihalcea and Ceylan, 2007), personal posts (Ouyang et al., 2017), and meeting minutes (Carletta et al., 2005) as described further in §5.

Beyond the corpora themselves, a variety of summarization systems have been developed: Mihalcea and Tarau (2004); Erkan and Radev (2004) used graph-based keyword ranking algorithms. Lin and Bilmes (2010); Carbonell and Goldstein (1998) found summary sentences which are highly relevant but less redundant. Yogatama et al. (2015) used semantic volumes of bigram features for extractive summarization. Internal structures of documents have been used in summarization: syntactic parse trees (Woodsend and Lapata, 2011; Cohn and Lapata, 2008), topics (Zajic et al., 2004; Lin and Hovy, 2000), semantic word graphs (Mehdad et al., 2014; Gerani et al., 2014; Ganesan et al., 2010; Filippova, 2010; Boudin and Morin, 2013), and abstract meaning representation (Liu et al., 2015). Concept-based Integer-Linear Programming (ILP) solver (McDonald, 2007) is used for optimizing the summarization problem (Gillick and Favre, 2009; Banerjee et al., 2015; Boudin et al., 2015; Berg-Kirkpatrick et al., 2011). Durrett et al. (2016) optimized the problem with grammatical and anarphorcity constraints.

With a large scale of corpora for training, neural network based systems have recently been developed. In abstractive systems, Rush et al. (2015) proposed a local attention-based sequence-to-sequence model. On top of the seq2seq framework, many other variants have been studied using convolutional networks (Cheng and Lapata, 2016; Allamanis et al., 2016), pointer networks (See et al., 2017), scheduled sampling (Bengio et al., 2015), and reinforcement learning (Paulus et al., 2017). In extractive systems, different types of encoders (Cheng and Lapata, 2016; Nallapati et al., 2017; Kedzie et al., 2018) and optimization techniques (Narayan et al., 2018b) have been developed. Our goal is to explore which types of systems learns which sub-aspect of summarization.

3 Sub-aspects of Summarization

We focus on three crucial aspects : Position, Diversity, and Importance. For each aspect, we use different extractive algorithms to **capture how much of the aspect is used in the oracle extractive summaries**[1]. For each algorithm, the goal is to select k extractive summary sentences (equal to the number of sentences in the target summaries for each sample) out of N sentences appearing in the original source. The chosen sentences or their indices will be used to calculate the various evaluation metrics described in §4

For some algorithms below, we use vector representation of sentences. We parse a document x into a sequence of sentences $x = x_1..x_N$ where each sentence consists of a sequence of words $x_i = w_1..w_s$. Each sentence is then encoded:

$$\mathrm{E}(x_i) = \mathtt{BERT}(w_{i,1}..w_{i,s}) \tag{1}$$

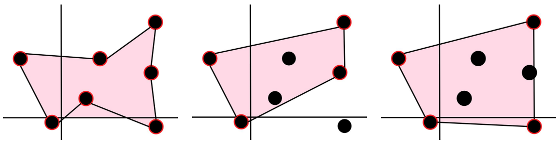
where BERT (Devlin et al., 2018) is a pre-trained bidirectional encoder from transformers (Vaswani et al., 2017)[2]. We use the last layer from BERT as a representation of each token, and then average them to get final representation of a sentence. All tokens are lower cased.

3.1 Position

Position of sentences in the source has been suggested as a good indicator for choosing summary sentences, especially in news articles (Lin and Hovy, 1997; Hong and Nenkova, 2014; See et al., 2017). We compare three position-based algorithms: **First**, **Last**, and **Middle**, by simply choosing k number of sentences in the source document from these positions.

3.2 Diversity

Yogatama et al. (2015) assume that extractive summary sentences which maximize the semantic volume in a distributed semantic space are the most diverse but least redundant sentences. Motivated by this notion, our goal is to find a set of k sentences that maximizes the volume size of them in a continuous embedding space like the BERT representations in Eq 1. Our objective is to find the optimal search function $\mathcal{S}$ that maximizes the volume size $\mathcal{V}$ of searched sentences: $\arg\max_{1..k} \mathcal{V}(\mathcal{S}_{1..c}(\mathrm{E}(x_1), \ldots, \mathrm{E}(x_N)))$.



(a) Default (b) Heuristic (c) ConvexFall

Figure 2: Volume maximization functions. Black dots are sentences in source document, and red dots are chosen summary sentences. The red-shaded polygons are volume space of the summary sentences.

If k=N, we use every sentence from the source document. (Figure 2 (a)). However, its volume space does not guarantee to maximize the volume size because of the non-convex polygonality. In order to find a convex maximum volume, we consider two different algorithms described below.

Heuristic. Yogatama et al. (2015) heuristically choose a set of summary sentences using a greedy algorithm: It first chooses a sentence which has the farthest vector representation from the centroid of whole source sentences, and then repeatedly finds sentences whose representation is farthest from

[1] See §4 for our oracle set construction.

[2] The other encoders such as averaging word embeddings (Pennington et al., 2014) show comparable performance.

the centroid of vector representations of the chosen sentences. Unlike the original algorithm in (Yogatama et al., 2015) restricting the number of words, we constrain the total number of selected sentences to k. This heuristic algorithm can fail to find the maximum volume depending on its starting point and/or the farther distance between two points detected (Figure 2 (b)).

ConvexFall. Here we first find the convex-hull[3] using Quickhull (Barber et al., 1996), implemented by Qhull library[4]. It guarantees the maximum volume size of selected points with minimum number of points (Figure 2 (c)). However, it does not reduce a redundancy between the points over the convex-hull, and usually choose larger number of sentences than k. Marcu (1999) shows an interesting study regarding an importance of sentences: given a document, if one deletes the least central sentence from the source text, then at some point the similarity with the reference text rapidly drops at sudden called the *waterfall* phenomena. Motivated by his study, we similarly prune redundant sentences from the set chosen by convex-hull search. For each turn, the sentence with the lowest volume reduction ratio is pruned until the number of remaining sentences is equivalent to k.

3.3 Importance

We assume that contents that repeatedly occur in one document contain *important* information. We find sentences that are nearest to the neighbour sentences using two distance measures: **N-Nearest** calculates an averaged Pearson correlation between one and the rest for all source sentence vector representations. k sentences having the highest averaged correlation are selected as final extractive summaries. On the other hand, **K-Nearest** chooses the K nearest sentences per each sentence, and then averages distances between each nearest sentence and the selected one. The one has the lowest averaged distance is chosen. This calculation is repeated k times and the selected sentences are removed from the remaining pool.

4 Metrics

In order to determine the aspects most crucial to the summarization task, we use three evaluation metrics:

`ROUGE` is Recall-Oriented Understudy for Gisting Evaluation (Lin and Hovy, 2000) for evaluating summarization systems. We use ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) F-measure scores which corresponds to uni-gram, bigrams and longest common subsequences, respectively, and their averaged score (R).

`Volume Overlap (VO)` ratio. Hard metrics like ROUGE often ignore semantic similarities between sentences. Based on the volume assumption in (Yogatama et al., 2015), we measure overlap ratio of two semantic volumes calculated by the model and target summaries. We obtain a set of vector representations of the reference summary sentences $\hat{Y}$ and the model summary sentences Y predicted by any algorithm *algo* in §3 for the i-th document:

$$\hat{Y}_i = (\hat{y}_{i,1} \,..\, \hat{y}_{i,k}), \quad Y_i^{algo} = (y_{i,1}^{algo} \,..\, y_{i,k}^{algo}) \tag{2}$$

Each volume `V` is then calculated using the convex-hull algorithm and their overlap ($\sqcap$) is calculated using a shapely package[5][6]. The final `VO` is then:

$$\mathrm{VO}_{algo} = \sum_{i=1}^{N} \frac{\mathtt{V}(E(Y_i^{algo})) \sqcap \mathtt{V}(E(\hat{Y}_i))}{\mathtt{V}(E(\hat{Y}_i))} \tag{3}$$

where N is the total number of input documents, E is the BERT sentence encoder in Eq 1, and $E(\hat{Y}_i)$ and $E(Y_i^{algo})$ are a set of vector representations of the reference and model summary sentences, respectively. The volume overlap indicates how two summaries are semantically overlapped in a continuous embedding space.

`Sentence Overlap (SO)` ratio. Even though ROUGE provides a recall-oriented lexical overlap, we don't know the upper-bound on performance (called `oracle`) of the extractive summarization. We extract the oracle extractive sentences (i.e. a set of input sentences) which maximizes ROUGE-L F-measure score with the reference summary. We then measure sentence overlap (`SO`) which determines how many extractive sentences from our algorithms are in the oracle summary. The `SO` is:

$$\mathrm{SO}_{algo} = \sum_{i=1}^{n} \frac{\mathtt{C}(Y_i^{algo} \cap \hat{Y}_i)}{\mathtt{C}(\hat{Y}_i)} \tag{4}$$

where `C` is a function for counting the number of elements in a set. The sentence overlap indicates

[3] Definition: a set of points is defined as the smallest convex set that includes the points.

[4] `http://www.qhull.org/`

[5] `https://pypi.org/project/Shapely/`

[6] Due to the lack of overlap calculation between two polygons of high dimensions, we reduce it to 2D PCA space.

how well the algorithm finds the oracle summaries for extractive summarization.

5 Summarization Corpora

We use various domains of summarization datasets to conduct the bias analysis across corpora and systems. Each dataset has source documents and corresponding abstractive target summaries. We provide a list of datasets used along with a brief description and our pre-processing scheme:

- CNNDM (Nallapati et al., 2016): contains 300K number of online news articles. It has multiple sentences (4.0 on average) as a summary.
- Newsroom (Grusky et al., 2018): contains 1.3M news articles and written summaries by authors and editors from 1998 to 2017. It has both extractive and abstractive summaries.
- XSum (Narayan et al., 2018a): has news articles and their single but abstractive sentence summaries mostly written by the original author.
- PeerRead (Kang et al., 2018): consists of scientific paper drafts in top-tier computer science venues as well as arxiv.org. We use full text of introduction section as source document and of abstract section as target summaries.
- PubMed (Kedzie et al., 2018): is 25,000 medical journal papers from the PubMed Open Access Subset.[7] Unlike PeerRead, full paper except for abstract is used as source documents.
- MScript (Gorinski and Lapata, 2015): is a collection of movie scripts from ScriptBase corpus and their corresponding user summaries of the movies.
- BookSum (Mihalcea and Ceylan, 2007): is a dataset of classic books paired to summaries from Grade Saver[8] and Cliffs Notes[9]. Due to a large number of sentences, we only choose the first 1K sentences for source document and the first 50 sentences for target summaries.
- Reddit (Ouyang et al., 2017): is a collection of personal posts from reddit.com. We use a single abstractive summary per post. The same data split from Kedzie et al. (2018) is used.
- AMI (Carletta et al., 2005): is documented meeting minutes from a hundred hours of recordings and their abstractive summaries.

Table 1 summarizes the characteristics of each dataset. We note that the Gigaword (Graff et al., 2003), New York Times[10], and Document Understanding Conference (DUC)[11] are also popular datasets commonly used in summarization analyses, though here we exclude them as they represent only additional collections of news articles, showing similar tendencies to the other news datasets such as CNNDM.

6 Analysis on Corpus Bias

We conduct different analyses of how each corpus is biased with respect to the sub-aspects. We highlight some key findings for each sub-section.

6.1 Multi-aspect analysis

Table 2 shows a comparison of the three aspects for each corpus where we include random selection and the oracle set. For each dataset metrics are calculated on a test set except for BookSum and AMI where we use train+test due to the smaller sample size.

Earlier isn't always better. Sentences selected early in the source show high ROUGE and SO on CNNDM, Newsroom, Reddit, and BookSum, but not in other domains such as medial journals and meeting minutes, and the condensed news summaries (XSum). For summarization of movie scripts in particular, the last sentences seem to provide more important summaries.

XSum requires much IMPORTANCE than other corpora. Interestingly, the most powerful algorithm for XSum is N-Nearest. This shows that summaries in XSum are indeed collected by abstracting multiple important contents into single sentence, avoiding the position bias.

First, ConvexFall, and N-Nearest tend to work better than the other algorithms for each aspect. First is better than Last or Middle in new articles except for XSum and personal posts, while not in academic papers (i.e., PeerRead, PubMed) and meeting minutes. ConvexFall finds the set of sentences that maximize the semantic volume overlap with the target sentences better than the heuristic one.

ROUGE and SO show similar behavior, while VO does not. In most evaluations, ROUGE scores are linear to SO ratios as expected. However, VO has high variance across algorithms and aspects.

[7] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[8] http://www.gradesaver.com

[9] http://www.cliffsnotes.com/

[10] https://catalog.ldc.upenn.edu/LDC2008T19

[11] http://duc.nist.gov

	CNNDM	Newsroom	Xsum	PeerRead	PubMed	Reddit	AMI	BookSum	MScript
Source	News	News	News	Papers	Papers	Post	Minutes	Books	Script
Multi-sents.	✓	✓	X	✓	✓	X	✓	✓	✓
Data size	287K/11K	992K/109K	203K/11K	10K/550	21K/2.5K	404/48	98/20	- /53	- /1K
Avg src sents.	40/34	24/24	33/33	45/45	97/97	19/15	767/761	- /6.7K	- /3K
Avg tgt sents.	4/4	1.4/1.4	1/1	6/6	10/10	1/1	17/17	- /336	- /5
Avg src tokens	792/779	769 /762	440/442	1K/1K	2.4K/2.3K	296/236	6.1K/6.4K	- /117K	- /23.4K
Avg tgt tokens	55/58	30/31	23/23	144/146	258/258	24/25	281/277	- /6.6K	- /104

Table 1: Data statistics on summarization corpora. Source is the domain of dataset. Multi-sents. is whether the summaries are multiple sentences or not. All statistics are divided by Train/Test except for `BookSum` and `MScript`.

		CNNDM			NewsRoom			XSum			PeerRead			PubMed			Reddit			AMI			BookSum			MScript		
		R	VO	SO	R	VO	SO	R	VO	SO	R	VO	SO	R	VO	SO	R	VO	SO	R	VO	SO	R	VO	SO	R	VO	SO
	RANDOM	19.1	18.6	14.6	10.1	2.1	9.0	9.3	-	8.4	27.9	42.5	26.2	30.1	46.9	13.0	11.8	-	11.3	12.0	39.3	2.4	29.4	85.8	4.9	8.1	25.2	0.1
	ORACLE	42.8	-	-	48.1	-	-	19.6	-	-	46.3	-	-	47.0	-	-	30.0	-	-	32.0	-	-	38.9	-	-	24.2	-	-
POSITION	First	**30.7**	13.1	**30.7**	**32.2**	**4.4**	**37.8**	9.1	-	8.7	**32.0**	40.7	**30.3**	27.6	44.3	13.8	**15.3**	-	**19.9**	11.4	48.0	**3.8**	**29.1**	85.1	**7.4**	6.9	12.4	**0.7**
	Last	16.4	18.6	8.2	7.7	1.9	4.4	8.3	-	7.0	28.9	38.5	27.0	28.9	45.2	14.0	11.2	-	10.7	7.8	42.1	2.0	26.5	85.3	3.3	**8.8**	19.5	0.2
	Middle	21.5	18.7	11.8	12.4	1.9	5.6	9.1	-	9.1	29.7	40.7	22.8	28.9	45.9	12.3	11.5	-	7.1	11.1	36.4	2.3	27.9	83.0	4.9	8.0	23.9	0.1
DIVERS.	ConvFall	21.6	**57.7**	15.0	10.6	4.2	7.3	8.4	-	8.0	29.8	**77.5**	25.9	28.2	**93.5**	11.2	11.6	-	7.5	**14.0**	**98.6**	2.4	16.9	**99.7**	2.2	8.5	**59.2**	0.2
	Heuris.	21.4	19.8	14.6	10.5	2.4	7.6	8.4	-	8.1	29.2	36.6	24.8	27.5	59.7	10.5	11.5	-	7.1	10.7	66.0	2.4	26.9	99.7	4.5	6.4	5.7	0.2
IMPORT.	NNear.	22.0	3.3	16.6	13.5	0.5	10.0	**9.8**	-	**10.1**	30.6	8.4	26.7	**31.8**	9.3	**15.5**	13.8	-	12.2	1.3	0.2	0.1	27.9	1.5	5.1	8.7	0.9	0.3
	KNear.	23.0	3.9	17.7	14.0	0.7	10.9	9.3	-	9.1	30.6	9.9	27.0	29.6	10.5	15.0	10.4	-	8.5	0.0	0.1	0.0	21.8	1.4	3.7	0.6	0.0	0.1

Table 2: Comparison of different corpora w.r.t the three sub-aspects: POSITION, DIVERSITY, and IMPORTANCE. We averaged R1, R2, and RL as R (See Appendix for full scores). Note that volume overlap (`VO`) doesn't exist when target summary has a single sentence. (i.e., `XSum`, `Reddit`)

This is mainly because the semantic volume assumption maximizes the semantic diversity, but sacrifices other aspects like importance by choosing the outlier sentences over the convex hull.

Social posts and news articles are biased to the position aspect while the other two aspects appear less relevant. (Figure 1 (a)) However, `XSum` requires all aspects equally but with relatively less relevant to any of aspects than the other news corpora.

Paper summarization is a well-balanced task. The variance of `SO` across the three aspects in `PeerRead` and `PubMed` is relatively smaller than other corpora. This indicates that abstract summary of the input paper requires the three aspects at the same time. `PeerRead` has relatively higher `SO` then `PubMed` because it only summarize text in Introduction section, while `PubMed` summarize whole paper text, which is much difficult (almost random performance).

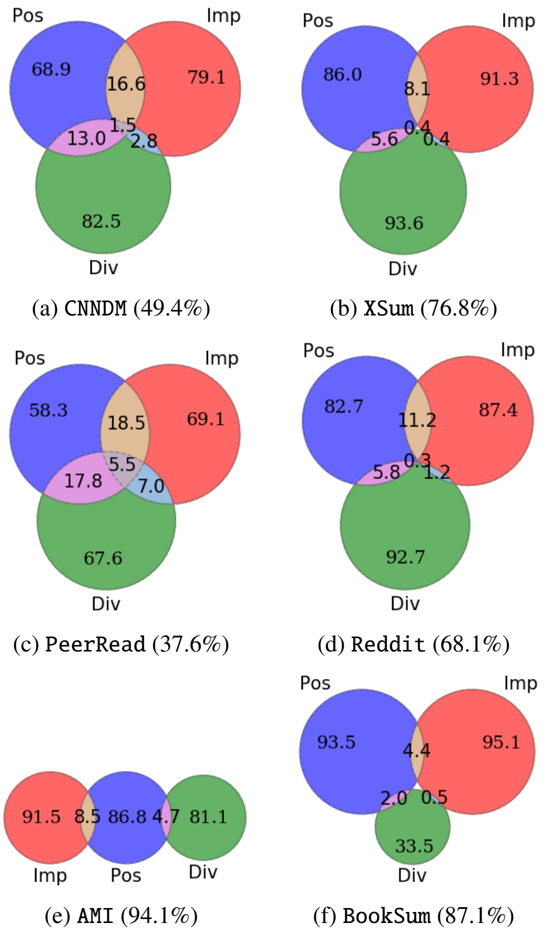
Conversation, movie script and book summarization are very challenging. Conversation of spoken meeting minutes includes a lot of witty replies repeatedly (e.g., 'okay.' , 'mm -hmm.' , 'yeah.'), causing importance and diversity measures to suffer. `MScript` and `BookSum` which include very long input document seem to be extremely difficult task, showing almost random performance.

6.2 Intersection between the sub-aspects

Averaged ratios across the sub-aspects do not capture how the actual summaries overlap with each other. Figure 3 shows Venn diagrams of how sets of summary sentences chosen by different sub-aspects are overlapped each other on average.

`XSum`, `BookSum`, and `AMI` have high Oracle Recall. If we develop a mixture model of the three aspects, the Oracle Recall means its upper bound, meaning that another sub-aspect should be considered regardless of the mixture model. This indicates that existing procedures are not enough to cover the Oracle sentences. For example, `AMI` and `BookSum` have a lot of repeated noisy sentences, some of which could likely be removed without a significant loss of pertinent information.

IMPORTANCE and DIVERSITY are less overlapped with each other. This means that important sentences are not always diverse sentences, indicating that they should be considered together.



(a) CNNDM (49.4%) (b) XSum (76.8%)

(c) PeerRead (37.6%) (d) Reddit (68.1%)

(e) AMI (94.1%) (f) BookSum (87.1%)

Figure 3: Intersection of averaged summary sentence overlaps across the sub-aspects. We use First for Position, ConvexFall for Diversity, and N-Nearest for Importance. The number in the parenthesis called *Oracle Recall* is the averaged ratio of how many the oracle sentences are **NOT** chosen by union set of the three sub-aspect algorithms. Other corpora are in Appendix with their Oracle Recalls: Newsroom(54.4%), PubMed (64.0%) and MScript (99.1%).

6.3 Summaries in a embedding space

Figure 4 shows two dimensional PCA projections of a document in CNNDM on the embedding space.

Source sentences are clustered on the convexhull border, not in the middle. We conjecture that sentences are not uniformly distributed in the embedding space but their positions gradually move over the convexhull. Target summaries reflect different sub-aspects according to the sample and corpora. For example, many target sentences in CNNDM are near by First-k sentences.

6.4 Single-aspect analysis

We calculate the frequency of source sentences overlapped with the oracle summary where the source sentences are ranked differently according to the algorithm of each aspect (See Figure 5). Heavily skewed histograms indicate that oracle sentences are positively (right-skewed) or negatively (left-skewed) related to the sub-aspect.

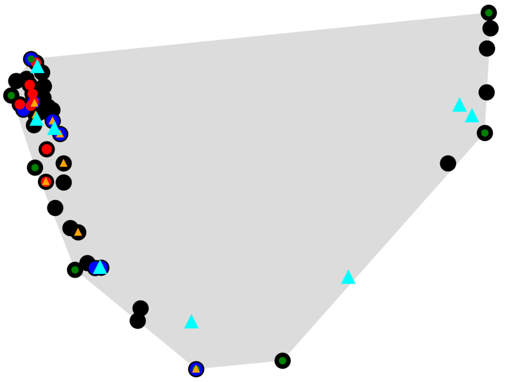


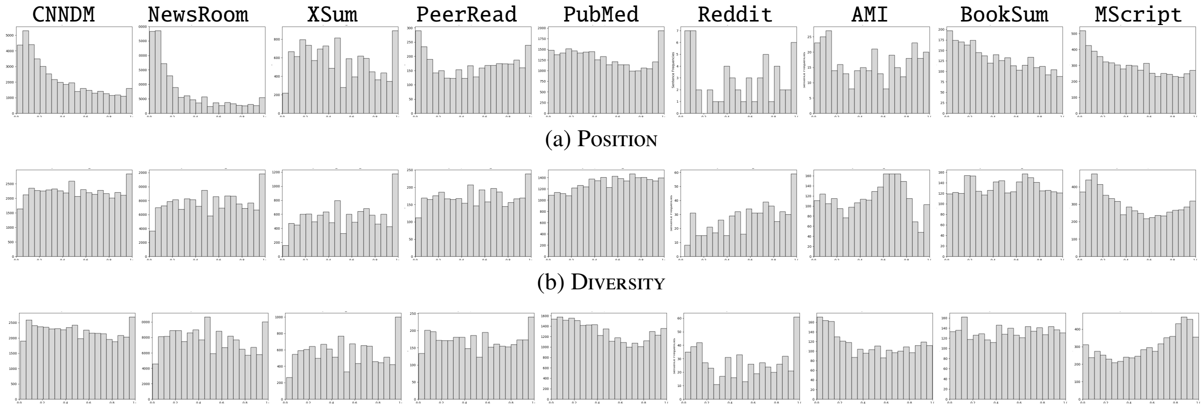
Figure 4: PCA projection of extractive summaries chosen by multiple aspects of algorithms (CNNDM). Source and target sentences are black circles (●) and cyan triangles, respectively. The blue, green, red circles are summary sentences chosen by First, ConvexFall, NN, respectively. The yellow triangles are the oracle sentences. Shaded polygon represents a ConvexHull volume of sample source document. Best viewed in color. Please find more examples in Appendix.

In most cases, some oracle sentences are overlapped to the first part of the source sentences. Even though their degrees are different, oracle summaries from many corpora (i.e, CNNDM, NewsRoom, PeerRead, BookSum, MScript) are highly related to the position. Compared to the other corpora, PubMed and AMI contain more top-ranked important sentences in their oracle summaries. News articles and papers tend to find oracle sentences without diversity (i.e., right-skewed), meaning that non-diverse sentences are frequently selected as part of the oracle.

We also measure how many *new* words occur in abstractive target summaries, by comparing overlap between oracle summaries and document sentences (Table 3). One thing to note is that XSum and AMI have less *new* words in their target summaries. On the other hand, paper datasets (i.e., PeerRead and PubMed) include a lot, indicating that abstract text in academic paper is indeed "abstract".

7 Analysis on System Bias

We study how current summarization systems are biased with respect to three sub-aspects. In addition, we show that a simple ensemble of sys-



(a) Position

(b) Diversity

(c) Importance

Figure 5: Sentence overlap proportion of each sub-aspect (row) with the oracle summary across corpora (column). y-axis is the frequency of overlapped sentences with the oracle summary. X-axis is the normalized RANK of individual sentences in the input document where size of bin is 0.05. E.g., the first / the most diverse / the most important sentence is in the first bin. If earlier bars are frequent, the aspect is positively relevant to the corpus.

	R(O,T)	**O∩T**		**T\S**	
		Unigram	Bigram	Unigram	Bigram
CNNDM	42.8	66.0	36.4	14.7	5.7
Newsroom	48.1	60.7	43.4	7.8	3.4
XSum	19.6	30.4	6.9	8.4	1.2
PeerRead	46.3	48.5	27.2	20.1	8.8
PubMed	47.0	52.1	27.7	16.7	6.7
Reddit	30.0	41.0	16.4	13.8	3.8
AMI	32.0	28.1	8.5	10.6	1.5
BookSum	38.9	25.6	8.9	6.7	1.7
MScript	38.9	13.9	4.0	0.3	0.1

Table 3: ROUGE of oracle summaries and averaged N-gram overlap ratios. O, T and S are a set of N-grams from Oracle, Target and Source document, respectively. **R(O,T)** is the averaged ROUGE between oracle and target summaries, showing how similar they are. **O∩T** shows N-gram overlap between oracle and target summaries. The higher the more overlapped words in between. **T\S** is a proportion of N-grams in target summaries not occurred in source document. The lower the more abstractive (i.e., new words) target summaries.

tems shows comparable performance to the single-aspect systems.

Existing systems. We compare various extractive and abstractive systems: For extractive systems, we use *K-Means* (Lin and Bilmes, 2010), Maximal Marginal Relevance (*MMR*) (Carbonell and Goldstein, 1998), *cILP* (Gillick and Favre, 2009; Boudin et al., 2015), *TexRank* (Mihalcea and Tarau, 2004), *LexRank* (Erkan and Radev, 2004) and three recent neural systems; *CL* (Cheng and Lapata, 2016), *SumRun* (Nallapati et al., 2017), and *S2SExt* (Kedzie et al., 2018). For abstractive systems, we use *WordILP* (Banerjee et al., 2015) and four neural systems; *S2SAbs* (Rush et al., 2015), *Pointer* (See et al., 2017), *Teacher* (Bengio et al., 2015), and *RL* (Paulus et al., 2017). The detailed description and experimental setup for each algorithm are in Appendix.

Proposed ensemble systems. Motivated by the sub-aspect theory (Lin and Bilmes, 2012, 2011), we combine different types of systems together from two different pools of extractive systems: asp from the three best algorithm from each aspect and ext from all extractive systems. For each combination, we choose the sumary sentences randomly among the union set of the predicted sentences (rand) or the most frequent unique sentences (topk).

Results. Table 4 shows a comparison of existing and proposed summarization systems on the set of corpora in §5 except for Newsroom[12]. Neural extractive systems such as *CL*, *SumRun* and *S2SExt* outperform the others in general. *LexRank* is highly biased toward the position aspect. On the other hand, *MMR* is extremely biased to the importance aspect on XSum and Reddit. Interestingly, neural extractive systems are somewhat balanced compared to the others. Ensemble systems seem to have the three sub-aspects in balance, compared to the neural extractive systems. They also outperform the others (either ROUGE or SO) on five out of eight datasets.

[12] We exclude it because of its similar behavior as CNNDM.

		CNNDM			XSum			PeerRead			PubMed			Reddit			AMI			BookSum			MScript		
		R	SO	R(P/D/I)	R	SO	R(P/D/I)	R	SO	R(P/D/I)	R	SO	R(P/D/I)	R	SO	R(P/D/I)	R	SO	R(P/D/I)	R	SO	R(P/D/I)	R	SO	R(P/D/I)
extractive	**KMeans**	22.2	16.3	14/22/34	9.8	10.0	14/8/90	30.9	28.3	24/28/38	30.6	14.2	31/40/46	14.0	12.5	10/2/82	12.3	2.5	9/6/7	27.2	4.6	5/2/14	9.1	0.3	0/0/9
	MMR	21.6	15.2	12/24/30	9.8	10.0	14/8/**97**	29.6	24.9	26/29/35	30.2	12.9	33/35/42	13.6	11.5	10/3/**88**	12.3	2.5	9/6/7	29.1	6.1	4/0/13	9.5	0.2	0/0/28
	TexRank	19.6	10.3	34/27/27	9.9	8.5	19/11/16	23.9	12.4	32/32/32	18.0	1.7	19/21/20	17.7	16.7	13/9/15	11.1	0.0	17/20/6	6.7	0.0	8/14/8	8.2	0.2	5/9/8
	LexRank	29.3	29.5	71/29/32	11.2	11.9	**61**/15/19	29.0	24.6	**66**/35/38	26.3	7.7	**56**/27/28	18.7	18.8	46/11/19	8.0	0.2	36/21/12	10.5	0.8	20/20/13	12.7	**0.5**	20/9/9
	wILP	23.1	15.6	27/28/29	11.1	2.1	28/19/21	20.2	16.0	23/27/26	15.6	6.0	14/20/18	17.4	13.5	42/16/20	5.1	0.6	17/18/17	4.3	1.3	5/12/7	6.8	0.1	6/8/6
	CL	**31.2**	30.0	**86**/29/31	**11.8**	**14.3**	25/13/19	31.3	21.8	55/35/38	26.3	9.2	41/26/26	19.4	24.0	23/14/23	23.1	10.3	19/23/5	-	-	-/-/-	14.0	0.2	6/8/7
	SumRun	30.5	27.1	68/29/31	11.6	13.1	14/13/19	**34.0**	20.5	38/36/37	29.4	10.8	27/28/27	**20.2**	19.8	23/12/21	**23.8**	11.4	21/23/6	-	-	-/-/-	14.4	0.0	5/9/9
	S2SExt	30.4	28.3	74/28/31	12.0	14.2	17/13/19	33.9	21.1	43/35/37	29.6	10.8	26/28/28	21.5	**34.4**	27/12/26	23.4	11.9	21/24/6	-	-	-/-/-	14.3	0.0	7/9/8
abstractive	**cILP**	27.8	x	43/31/32	10.9	x	49/15/18	28.2	x	35/36/38	27.8	x	23/29/30	17.7	x	**53**/15/17	12.5	x	22/33/10	7.9	x	9/19/12	10.6	x	5/7/7
	S2SAbs	16.3	x	4/4/4	10.4	x	8/7/8	9.9	x	9/9/9	10.2	x	10/10/10	11.9	x	11/7/8	20.3	x	9/12/1	-	-x	-/-/-	14.0	x	6/8/8
	+Pointer	23.9	x	20/13/14	15.6	x	12/11/12	13.6	x	13/13/13	11.2	x	11/12/11	14.3	x	14/10/12	23.0	x	11/13/1	-	-x	-/-/-	10.0	x	6/7/7
	+Teacher	29.7	x	33/21/22	17.0	x	12/10/12	8.7	x	8/8/8	11.3	x	12/12/11	15.3	x	15/10/11	20.2	x	9/13/1	-	-x	-/-/-	16.0	x	7/10/8
	+RL	30.2	x	34/23/24	18.1	x	12/11/12	30.1	x	30/29/28	12.9	x	13/14/13	16.7	x	1/1/14	23.6	x	11/13/2	-	-x	-/-/-	**16.2**	x	7/10/8
ensemble	ASP(rand)	23.3	19.5	40/**38**/**38**	9.0	9.0	40/**39**/38	29.6	25.5	54/49/52	29.5	13.5	49/**47**/**51**	12.5	5.2	21/11/22	8.9	0.9	44/**50**/**20**	**29.8**	**6.4**	57/33/55	8.4	0.4	**32**/36/37
	ASP(topk)	29.1	**30.4**	71/31/31	9.0	8.8	43/**39**/38	30.5	28.2	63/**54**/**57**	29.7	14.0	55/48/52	12.3	15.6	41/**41**/38	9.9	1.5	**99**/24/11	29.6	6.2	**58**/**34**/**56**	8.3	**0.5**	30/**37**/**38**
	EXT(rand)	24.2	20.2	39/25/27	10.2	10.9	17/13/23	29.4	23.5	42/37/39	31.7	16.0	37/34/38	14.2	17.7	22/12/13	18.7	5.1	21/28/8	28.6	5.4	37/24/42	6.7	0.0	5/9/13
	EXT(topk)	29.4	30.3	58/25/28	11.0	11.8	18/10/37	33.0	**33.0**	54/39/44	**34.1**	**20.5**	41/35/40	16.4	20.8	21/11/52	**23.8**	**13.4**	23/27/6	28.5	5.2	37/24/43	7.4	0.0	6/8/11

Table 4: Comparison of different systems using the averaged ROUGE scores (1/2/L) with target summaries (R) and averaged oracle overlap ratios (SO, only for extractive systems). We calculate R between systems and selected summary sentences from each sub-aspect (R(P/D/I)) where each aspect uses the best algorithm: First, ConvexFall and NNearest. R(P/D/I) is rounded by the decimal point. - indicates the system has too few samples to train the neural systems. x indicates SO is not applicable because abstractive systems have no sentence indices. The best score for each corpora is shown in bold with different colors.

8 Conclusion and Future Directions

We define three sub-aspects of text summarization: position, diversity, and importance. We analyze how different domains of summarization dataset are biased to these aspects. We observe that news articles strongly reflect the position aspect, while the others do not. In addition, we investigate how current summarization systems reflect these three sub-aspects in balance. Each type of approach has its own bias, while neural systems rarely do. Simple ensembling of the systems shows more balanced and comparable performance than single ones.

We summarize actionable messages for future summarization research:

- Different domains of datasets except for news articles pose new challenges to the appropriate design of summarization systems. For example, summarization of conversations (e.g., `AMI`) or dialogues (`MSCript`) need to filter out repeated, rhetorical utterances. Book summarization (e.g., `BookSum`) is very challenging due to its extremely large document size. Here current neural encoders suffer from computation limits.
- Summarization systems to be developed should clearly state their computational limits as well as effectiveness in each aspect and in each corpus domain. A good summarization system should reflect different kinds of the sub-aspects harmoniously, regardless of corpus bias. Developing such bias-free or robust models can be very important for future directions.
- Nobody has clearly defined the deeper nature of meaning abstraction yet. A more theoretical study of summarization, and the various aspects, is required. A recent notable example is Peyrard (2019a)'s attempt to theoretically define different quantities of `importance` aspect, and demonstrate the potential of the framework on an existing summarization system. Similar studies can be applied to other aspects and their combinations in various systems and different domains of corpora.
- One can repeat our bias study on evaluation metrics. Peyrard (2019b) showed that widely used evaluation metrics (e.g., ROUGE, Jensen-Shannon divergence) are strongly mismatched in scoring summary results. One can compare different measures (e.g., n-gram recall, sentence overlaps, embedding similarities, word connectedness, centrality, importance reflected by discourse structures), and study bias of each with respect to systems and corpora.

Acknowledgements

This work would not have been possible without the efforts of the authors who kindly share the summarization datasets publicly. We thank Rada Mihalcea for sharing the book summarization dataset. We also thank Diane J. Litman, Taylor Berg-Kirkpatrick, Hiroaki Hayashi, and anonymous reviewers for their helpful comments.

References

Miltiadis Allamanis, Hao Peng, and Charles Sutton. 2016. A convolutional attention network for extreme summarization of source code. *arXiv preprint arXiv:1602.03001*.

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Multi-document abstractive summarization using ilp based multi-sentence compression. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*.

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.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 481–490. Association for Computational Linguistics.

Florian Boudin and Emmanuel Morin. 2013. Keyphrase extraction for n-best reranking in multi-sentence compression. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Florian Boudin, Hugo Mougard, and Benoit Favre. 2015. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2015*.

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*, pages 335–336. 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.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 137–144. Association for Computational Linguistics.

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

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. *arXiv preprint arXiv:1603.08887*.

Harold P Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479.

Katja Filippova. 2010. Multi-sentence compression: finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 322–330. Association for Computational Linguistics.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*, pages 340–348. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of EMNLP*.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 10–18. Association for Computational Linguistics.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of* *the Association for Computational Linguistics: Human Language Technologies*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New Orleans, USA.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. *arXiv preprint arXiv:1810.12343*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8.

Chin-Yew Lin and Eduard Hovy. 1997. Identifying topics by position. In *Fifth Conference on Applied Natural Language Processing*.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 495–501. Association for Computational Linguistics.

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

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics.

Hui Lin and Jeff A Bilmes. 2012. Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871*.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, 293:123–136.

Ryan McDonald. 2007. *A study of global inference algorithms in multi-document summarization*. Springer.

Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. 2014. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proc. of ACL*, pages 1220–1230.

Rada Mihalcea and Hakan Ceylan. 2007. Explorations in automatic book summarization. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

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.

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*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

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. *arXiv preprint arXiv:1808.08745*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Jessica Ouyang, Serina Chang, and Kathy McKeown. 2017. Crowd-sourced iterative annotation for narrative summarization corpora. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 46–51.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Maxime Peyrard. 2019a. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard. 2019b. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

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

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.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the conference on empirical methods in natural language processing*, pages 409–420. Association for Computational Linguistics.

Dani Yogatama, Fei Liu, and Noah A Smith. 2015. Extractive summarization by maximizing semantic volume. In *EMNLP*, pages 1961–1966.

David Zajic, Bonnie Dorr, and Richard Schwartz. 2004. Bbn/umd at duc-2004: Topiary. In *Proceedings of the HLT-NAACL 2004 Document Understanding Workshop, Boston*, pages 112–119.