How well do you know your summarization datasets?

Priyam Tejaswin * **Dhruv Naik** **Pengfei Liu** †
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA
{ptejaswi, drn, pliu3}@andrew.cmu.edu

Abstract

State-of-the-art summarization systems are trained and evaluated on massive datasets scraped from the web. Despite their prevalence, we know very little about the underlying characteristics (data noise, summarization complexity, etc.) of these datasets, and how these affect system performance and the reliability of automatic metrics like ROUGE. In this study, we manually analyse 600 samples from three popular summarization datasets. Our study is driven by a six-class typology which captures different noise types (missing facts, entities) and degrees of summarization difficulty (extractive, abstractive). We follow with a thorough analysis of 27 state-of-the-art summarization models and 5 popular metrics, and report our key insights: (1) Datasets have distinct data quality and complexity distributions, which can be traced back to their collection process. (2) The performance of models and reliability of metrics is dependent on sample complexity. (3) Faithful summaries often receive low scores because of the poor diversity of references. We release the code, annotated data and model outputs.[1]

1 Introduction

The past few years have witnessed major breakthroughs and improvements in automatic summarization (See et al., 2017; Celikyilmaz et al., 2018; Jadhav and Rajan, 2018; Liu and Lapata, 2019; Liu, 2019; Dou et al., 2020; Yuan et al., 2021; Liu et al., 2021). Apart from the improvements in the summarization model architectures (Zhang et al., 2019; Zhong et al., 2020), this growth has been aided by large-scale datasets (Nallapati et al., 2016; Narayan et al., 2018a; Sharma et al., 2019) and automatic evaluation metrics (Lin, 2004; Zhao et al., 2019; Kryscinski et al., 2020) which are used for tuning hyperparameters and comparing models. While the reliability of these metrics has been explored extensively (Peyrard, 2019; Bhandari et al., 2020; Fabbri et al., 2020), few studies have focused on the underlying characteristics of different datasets, and how these impact model performance and metric reliability.

Datasets like CNN/DailyMail (Nallapati et al., 2016), Gigaword (Rush et al., 2015), XSum (Narayan et al., 2018a), and many more (Wang and Ling, 2016; Koupaee and Wang, 2018; Kim et al., 2019; Ganesan et al., 2010) were collected by scraping a large collection of web-pages. And for all the benefits this approach offers (seemingly infinite samples, diverse subjects, etc) there are some caveats:

Data Noise We have no idea about the **noise** in the dataset. In the context of text summarization, noise could be an incomplete or irrelevant reference. At the moment, its quantity and impact on the performance is unknown.

Summarization Complexity What do we really know about the nature of samples in the dataset? Gigaword is a headline generation dataset with short sources and references. Does this imply a higher volume of simpler (i.e. more *extractive*) samples? The degree of summarization complexity, and its impact on model performance is unknown.

Exploring these open questions is critical for two reasons: (1) Information about the noise could lead to more informed data collection and preprocessing methods: in a recent study, Kryscinski et al. (2019) quantified HTML artefacts in popular summarization datasets, and proposed ways to detect and remove them. (2) Awareness about the complexity could better explain model performance, metrics, and even lead to new model architectures. In the tasks of machine comprehension

*This author was the primary contributor.
†Corresponding author.
[1]https://github.com/priyamtejaswin/howwelldoyouknow

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3436–3449

and question answering, Chen et al. (2016) and Yatskar (2019) manually inspected random samples and drew insights which led to new state-of-the-art models. Such analysis could also help researchers choose datasets and metrics more carefully.

In this study, we perform intrinsic and model-centric evaluation of three popular summarization datasets (Gigaword, CNN/DM and XSum). We are interested in answering the following questions:

Q1. What are the underlying intrinsic properties of summarization datasets? We are interested in (1) Identifying and quantifying the different types of "noise" that could occur and could penalize models. (2) Whether samples have different levels of difficulty. Armed with this, we ask the following questions.

Q2 a. How do these properties impact model performance? Specifically, we'd like to know (1) If, and how, the performance varies across the different types of samples discovered from Q1. (2) If the performance is consistent across metrics.

Q2 b. If the reliability of metrics changes with these properties? This is motivated (in part) from prior metric-analysis studies, where researchers have explored inter-metric agreement and alignment with human-judgement under different conditions (Peyrard, 2019; Bhandari et al., 2020). Here we are more interested in knowing if the metrics are more correlated with human judgement for simpler samples, than complex ones.

Large-scale automatic intrinsic dataset evaluation has been explored with some promising results (Bommasani and Cardie, 2020). However, these methods rely on heuristics like content-value, density and compression (Grusky et al., 2018). We are interested in a more fine-grained, interpretable analysis that can only come from manual inspection, much like the analysis by Chen et al. (2016) and by Yatskar (2019). To that end, we first define a six-class typology: the first three classes cover types of data-noise and the last three cover varying degrees of summarization difficulty. We then proceed to answer the aforementioned research questions, and discuss our key observations which are summarized below:

Key Observations: (1) Datasets have distinct modalities – a mix of simpler samples (which we call *Extractive*) and complex ones (which we call *Paraphrase* and *Inference*. (2) Gigaword is *majorly* Extractive but suffers from data noise (45% of the targets have some key entity, or fact that is absent from the source). (3) CNN/DM is relatively cleaner, and the authors' attempts to create a more abstractive dataset seems to be successful compared with Gigaword (only 18% of samples are Extractive). (4) XSum has no Extractive samples, but also has the greatest fraction of noise: 54% of the test samples have key entities or facts missing from the source. (5) Within the datasets, the broad performance trends between the typology classes are *consistent across all metrics*: **simpler samples score higher than complex ones**. (6) Metric reliability is also complexity dependent: On CNN/DM the agreement with human judgement decreases as summarization complexity increases.

The remainder of the paper is organised as follows: in Section 2 we answer **Q1**, describe the three datasets, define the typology, and present results from the annotation. In Section 3 we explore **Q2 a.** and evaluate different models on a variety of metrics (automatic and human-judgement). In Section 4 we explore **Q2 b.** and investigate metric reliability. In Section 5 we share some learnings from our experience. We conclude with Section 7.

2 Evaluating the intrinsic properties of summarization datasets (Q1)

	Length(Doc)		Length(Ref)		Sample	
	train	test	train	test	train	test
Gigawords	31	29	8	8	3.8M	1.9K
CNNDM	691	682	51	54	287K	11K
XSum	374	376	21	21	204K	11.3K

Table 1: Statistics of the three datasets. Length refers to the average number of words per Document/Reference.

2.1 Datasets for Annotation

Among many summarization datasets, we choose the following:

Gigaword is a summarizaiton dataset extracted from news articles (Rush et al., 2015)[2].

CNN/DailyMail or "CNN/DM" question answering dataset (Hermann et al., 2015; Nallapati et al., 2016) is commonly used for summarization. The dataset consists of online news articles paired with human-generated summaries.[3]

[2]We use the version most commonly used by summarization systems: https://github.com/harvardnlp/sent-summary

[3]We use the non-anonymized data as See et al. (2017).

Label Source Dataset	Source Target State-of-the-art Model Output
Incomplete / Irrelevant Gigaword	Andre Blom and Mark Scharrenberg scored tries and some tactical kicks in the final 10 minutes sent the United States to the Rugby World Cup with a 21-16 victory over Uruguay on Saturday . London testing , please ignore . United States beats Uruguay 21-16 in Rugby World Cup .
Entity Missing Gigaword	The United States claimed credit Tuesday for a ceasefire that ended fighting between Israel and Lebanese guerrillas , and rejected suggestions that it was forced to model the agreement after a French draft . US takes the credit for Israel-Hezbollah ceasefire by Carole Landry . Us claims credit for lebanon ceasefire .
Extractive CNN-DM	Ed Miliband's US adviser pays no tax in Britain on his reported £300,000 salary, he has admitted. David Axelrod masterminded two presidential election victories for Barack Obama and was hired by the Labour leader amid great fanfare last year. He has helped refine Mr Miliband's message ...*(truncated)* ... have been aware of Labour's eye-catching crackdown on non-doms last week. But speaking in the US where he is promoting his autobiography, Mr Axelrod revealed he is not resident for tax purposes in the UK. Asked whether he pays tax in Britain, he told the Daily Telegraph: 'I don't do my accounting so I don't know but I'm not in residence there.' Labour confirmed it pays Mr Axelrod in dollars through his consultancy firm and that he 'lives in the US, works in the US and pays taxes in the US'. ... (*truncated*) David Axelrod masterminded two Obama presidential election victories . He was hired by Labour leader Ed Miliband amid great fanfare last year . Revealed at a book launch that he is not resident for tax purposes in UK . Labour confirms it pays Mr Axelrod in dollars through consultancy firm . David Axelrod masterminded two presidential election victories for Barack Obama . He was hired by the Labour leader amid great fanfare last year . Has helped refine Mr Miliband 's message about tackling the cost of living and making sure the wealthy pay their fair share . Mr Axelrod makes infrequent visits to the UK to meet Mr Miliband and offers advice by phone .
Paraphrase CNN-DM	The number of women in Britain becoming nuns is at a 25-year high. Figures from the Catholic Church show the number of women taking Holy Vows has trebled from 15 in 2009 to 45 last year. From a low of only seven in 2004, the figure has been rising for the past decade. Theodora Hawksley, 29, was until recently a post-doctoral researcher in theology at the University of Edinburgh. But at the beginning of the year she decided to become a nun. *(truncated)*. Far from being trapped in traditional habits, Miss Hawksley said her order tends to dress down in T-shirts and jeans. Father Christopher Jamison, director of the National Office for Vocation of England and Wales, said: 'There is a gap in the market for meaning in our culture. One of the ways women may find that meaning is through religious life.' Sister Cathy Jones, religious life vocations promoter at the office, said: *(truncated)* . Figures from the Catholic Church show more and more becoming nuns . The number of women taking Holy Vows stood at just seven back in 2004 . But that figure had risen to 15 in 2009 and increased further to 45 last year . One father said a ' gap in the market for meaning ' led people toward religion . Figures from Catholic Church show number of women taking Holy Vows has trebled from 15 in 2009 to 45 last year . From a low of seven in 2004 , the figure has been rising for the past decade . Theodora Hawksley , 29 , was until recently a post - doctoral researcher in theology at the University of Edinburgh . But at the beginning of the year she decided to become a nun .
Inference Gigaword	Three Malaysian and Indonesian seamen kidnapped by Philippine Abu Sayyaf kidnap-for-ransom group allegedly had been executed and the skeletons discovered in the southern Philippines are believed to be their remains , a local television reported Wednesday . Abu Sayyaf hostages allegedly executed : report . 3 filipino , Indonesian seamen executed in southern Philippines .

Table 2: Examples for each of the six categories. Text spans with the same colors correspond to the same fact in the source and target. Target spans in RED are missing or unsupported in the source. The last sample is "Inference" because the writer will have to understand the concept of hostages, and then generalise from the group to an individual.

XSum or "Extreme Summarization" (Narayan et al., 2018a) was constructed from online news articles for highly abstractive summarization.

We consider these datasets because of their popularity, and the difference in the nature of samples. The latter enables a more comprehensive analysis; Table 1 captures the size of source and target documents along with the number of samples.

2.2 Typology Definition

The classes are defined below in order of priority. Some examples are in Table 2. Readers may refer to the Appendix B, C, D for more examples.

- **Incomplete/Irrelevant**: The target summary ends abruptly. Or the source and target are unrelated.
- **Entity Missing**: The target summary contains entities (names, dates, events, etc) that are absent from the source.
- **Evidence Missing**: The target summary is based on concepts which are absent from the source. However, the target is not **Incomplete** and all **Entities** are present.
- **Extractive**: The target is constructed by copying tokens from the source, mostly in-order of their appearance. Minor modifications, like stemming and abbreviating, are permitted. Word substitutions, and additions, are limited to a few. No reasoning, conclusion or co-ref resolution is performed as part of the summarization. The complete context of the target should be present in the source.
- **Paraphrase**: The majority of tokens in the target are substituted, or appear out of order, or both. There is no reasoning, conclusion or co-ref resolution. The complete context of the

target should be present in the source.

- **Inference**: A non-trivial "inference" activity has to be completed to construct the target: some reasoning, conclusion, or complex co-reference resolution. The complete context of the target should be present in the source.

We annotate 200 samples from each dataset, on par with similar studies on intrinsic evaluation (Chen et al., 2016; Cao et al., 2017). Two authors annotate samples independently. Annotations matched for 70%, 68% and 73% of Gigaword, CNN-DM and XSum samples, respectively. Disagreements were discussed between all authors before arriving at a consensus for the final label.

2.2.1 Motivation and Advantages

To the best of our knowledge, summarization datasets have not been manually analysed in this manner. A review of the most relevant summarization dataset analysis research shows that the most common form of intrinsic evaluation is to use surface-level heuristics. Most studies only cover a part of our typology, while almost all studies ignore the noise present in datasets.

Coverage , Density, Redundancy Grusky et al. (2018); Bommasani and Cardie (2020); Zhong et al. (2019b) use similar forms of token-level coverage between the source and the reference to measure the *extractiveness* of the summary. In it's simplest form, this is a ratio of the number of overlapping tokens and reference length. In our definition of **Extractive**, we first set a meaninful, well-defined criterion, and then manually check for *extractive* references, while allowing for some relaxations.

Content Compression In most papers (Grusky et al., 2018; Zhong et al., 2019b; Bommasani and Cardie, 2020), the summarization complexity is defined by a compression ratio (usually the normalized word-count ratio of the source and reference). As a standalone metric, this does indeed capture the difficulty in replication. However, token rearrangement, substitution, reformulation is ignored in this measure of "complexity". To combat this, we distinctly defined **Paraphrase** and **Inference**. By manually analysing samples, we are able to differentiate between the obviously simple Extractive samples, the relatively tougher Paraphrase samples and the most difficult Inference samples. Together these three offer a highly intuitive classification of samples. Part of the reason that the Machine Comprehension analysis by Chen et al. (2016) was so effective was the interpretability of their classes. We hope our analysis will also enable researchers to improve summarization models.

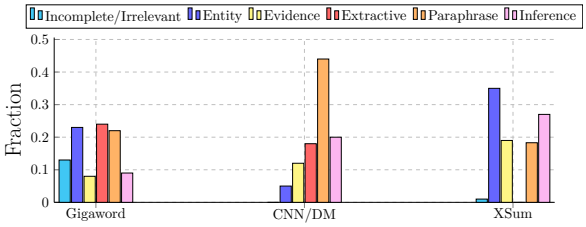


Figure 1: Distribution of the different class of samples in all datasets.

Noise Prior works have not focused on quantify the noise in popular datasets. Moreover, none of these metrics are designed to account for noise or factual inconsistencies. A high value for content compression might imply a high-degree of summarization complexity. But this ignores the possibility that the source-reference pair is unrelated (like row 1 in Table 2). In addition, the manual analysis allows us to identify factual errors and co-ref errors.

This is not to say the typology is perfect and exhaustive. Limitations and possible extensions to our typology are discussed in Section 5.

2.3 Dataset Analysis

The distribution of classes in the datasets is in Figure 1. We have made the following **key observations** in our analysis of the labels.

Gigawords is Extractive, but very noisy. 24.5% of summaries are Extractive, but 44.5% of samples belong to Entity Missing, Evidence Missing, or Incomplete. Not unexpected considering the "headline" nature of the samples.

XSum is Abstractive, but also very noisy. The authors (Narayan et al., 2018a) designed the dataset to be highly abstractive. This is reflected in the distribution: there were no Extractive samples in our analysis, suggesting a significantly higher level of difficulty. However, 55% of samples belong to Entity Missing, Evidence Missing, or Incomplete classes. The remaining 45% belongs to Paraphrase and Inference categories. Since we found only two incomplete samples, this class is ignored in all further XSum analysis.

CNN/DM is cleaner, and lives up to the design goals. The authors (Hermann et al., 2015) de-

signed CNN/DM to be *abstractive* in nature, and this is reflected in the distribution: 64% of samples belong to Paraphrase and Inference categories. Of the three, CNN/DM has the lowest fraction of factual and data noise: there are no Incomplete/Irrelavant samples, and only 18% of samples belong to Entity Missing and Evidence Missing.

The degree with which missing facts affects automatic evaluation varies. In some samples, one or two entities are missing (like Row 2 in Table 2), but in others multiple facts are missing. Empirical analysis of model performance for each class of samples is discussed in Section 3.

3 Performance on different classes (Q2 a)

In this section, we list the different models and metrics considered for analysis, and then describe how model performance varies across class labels.

3.1 Models for evaluation

We collect outputs from 7 systems for Gigaword: (1) PEGASUS (Zhang et al., 2019), (2) PROPHET (Qi et al., 2020) (Lewis et al., 2020), (3) UNILM (Dong et al., 2019) , (4) BISET (Song et al., 2020), (5) CONCOPY (Wang et al., 2019) , (6) POINTERGENERATOR (See et al., 2017), (7) POINTERGENERATORCOPYING (See et al., 2017)

For CNN/DM, we use the outputs of 11 top-performing summarization systems collected by Bhandari et al. (2020)[4]: (1) HETERGRAPH (Wang et al., 2020), (2) MATCHSUMM (Lewis et al., 2020), (3) REFRESH (Narayan et al., 2018b) , (4) TWOSTAGERL (Song et al., 2020), (5) NEUSUMM (Wang et al., 2019) , (6) BOTTOMUP (Gehrmann et al., 2018) (7) SEMSIM (Yoon et al., 2020) (8) UNILM (Dong et al., 2019) (9) BARTABSTRACTIVE (Lewis et al., 2020) (10) BANDITSUMM (Dong et al., 2018) (11) BARTEXTRACTIVE (Lewis et al., 2020)

For XSum, we use the outputs of 9 different summarization systems: (1) CONVSEQ2SEQ (Gehring et al., 2017), (2) TCONVS2S (Narayan et al., 2018a) (3) POINTERGENERATOR (See et al., 2017), (4) BART (Lewis et al., 2020), (5) PRESUMMEXTRACTIVE (Liu and Lapata, 2019), (6) PRESUMMABSTRACCTIVE (Liu and Lapata, 2019), (7) PRESUMMTRANSFORMER (Liu and Lapata, 2019), (8) LEAD (Nenkova, 2005), (9) EXTORACLE (Nallapati et al., 2017)

[4]https://github.com/neulab/REALSumm

3.2 Metrics for evaluation

Existing summarization systems are usually evaluated using automated metrics or manually using human judgments. We list popular automatic metrics explored in this work. Except for the last two, all outputs from every model is scored on the following metrics.

ROUGE-1/2/L measure overlap of unigrams, bigrams and longest common subsequence. respectively[5] (Lin, 2004).

BERTScore (BS) measures soft overlap between contextual BERT embeddings of tokens between the two texts[6] (Zhang et al., 2020).

MoverScore (MS) applies a distance measure to contextualized BERT and ELMo word embeddings[7] (Zhao et al., 2019).

FactCC is introduced to measure the fact consistency between the generated summaries and source documents (Kryscinski et al., 2020). Due to issues with the setup and training procedure, this metric was only used in the CNN/DM analysis.

Human Pyramid (HP) provides a robust technique for evaluating content selection by exhaustively obtaining a set of Semantic Content Units (SCUs) from a set of references, and then scoring system summaries on the number of SCUs that can be inferred (Nenkova and Passonneau, 2004). We use the scores shared by Bhandari et al. (2020) for the first 100 samples of CNN/DM subset.

3.3 Model Performance

For each dataset, we group the samples by their labels. For all samples in a subset, the model response is scored using a metric. The mean of these sample scores returns a single subset-model-metric score, which is then averaged across all models in the subset, leaving us with a single subset-metric score. This is repeated for all (subset $\times$ metric) pairs. The results are captured in Figures 2, 3 and 4 for Gigaword, CNN/DM and XSum respectively. The last column in each group is the average score across all samples.

3.3.1 Impact of Data Quality and Noise

Incomplete and Irrelevant Of the three datasets, only Gigaword contains Incomplete (or Irrelevant) samples. Across all metrics, the

[5]For ROUGE-1,2, and L, we used the Python implementation: https://github.com/sebastianGehrmann/rouge-baselines

[6]Used code at github.com/Tiiiger/bert_score

[7]We used a faster version of the code provided by the author at github.com/AIPHES/emnlp19-moverscore

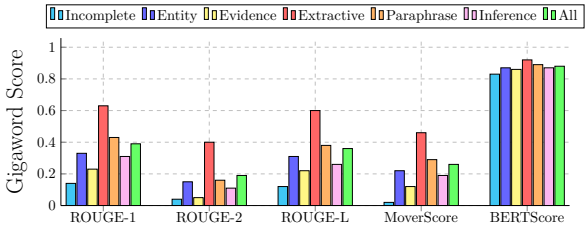


Figure 2: Gigaword class-level performance, averaged across all models.

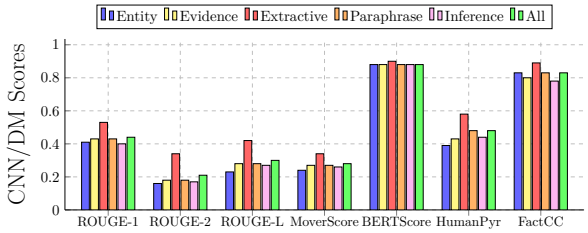


Figure 3: CNN/DM class-level performance, averaged across all models.

performance on this label is lowest, which is to be expected – high overlap will be rare if the source and target are unrelated or incomplete (like Row 1, Table 2). What's alarming is the volume of such samples in Gigaword – if the distribution is the same for the training set, then the model is being trained on extremely noisy data (almost 14%). In addition, such samples needlessly penalise the model performance during evaluation.

Entity scores more than Evidence in Gigaword! The results for these subsets are a bit surprising. In Gigaword, the Entity Missing subset receives relatively higher scores than the Evidence Missing category. We attribute this to a combination of factors. Consider Row 2 in Table 2. Entities are missing, but token overlap is high (more than 50%), which explains the high R1 scores, but low R2 scores. In our observations, the impact of missing facts and entities varies by the length of the target, as well as the number of entities.

Are Evidence Missing and Paraphrase are all the same for CNN/DM and XSum? When compared with Gigaword, samples with data quality issues (i.e. Incomplete/Irrelevant, Entity Missing and Evidence Missing samples) in CNN/DM and XSum get relatively higher scores. The reasons are similar to the Gigaword phenomenon discussed before. The average summary length of CNN/DM (54 tokens) is about 7 times that of Gigaword (8 tokens). As a result, with respect to the complete reference, one or two missing facts amounts to a

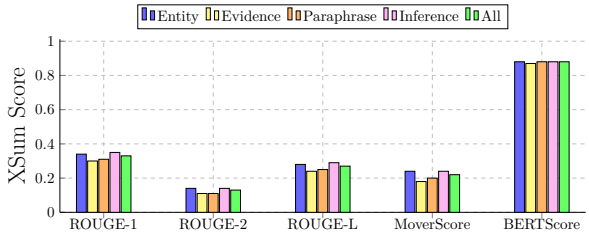


Figure 4: XSum class-level performance, averaged across all models.

much smaller fraction of the reference in CNN/DM. The high overlap with the remainder leads to higher scores.

Factual Correctness in CNN/DM Automatic metrics only consider the token overlap (or "semantic distance") between the target and the model output. While such metrics exhibit high correlation with human-judgement, a low score does not necessarily imply an incorrect generation, as demonstrated by Freitag et al. (2020) for machine translation. Hence we check for factual correctness of model outputs using FactCC. The competitive scores on the first three categories for FactCC in Fig .3 suggests the outputs generated by the model are factually faithful, which points to issues with the metric reliability. We discuss this in Section 4.

3.3.2 Impact of Summarization Complexity

For the last three categories (Extractive, Paraphrase and Inference) Gigaword and CNN/DM exhibit a common trend: the highest performance, across all metrics is on the Extractive subset, followed by Paraphrase samples which are more difficult to reproduce. The lowest performance is on the Inference samples. However, concluding models perform poorly would be incorrect. The last three samples in Table 2 suggest that model outputs are coherent, logical and factually faithful. FactCC scores in Figure 3 also suggest the outputs are factually consistent.

Some metrics are biased towards simpler samples? For the Extractive, Paraphrase and Inference samples, the samples we manually observed (some of which are captured in Table 2) and the FactCC scores indicates a gap in the token-based metrics. However, we cannot fault the metrics entirely. If we had diverse target references for the same sources, some outputs would have found better matches, and thus, higher scores! In fact, we see that BERTScore (a more "semantically" oriented metric) is extremely competitive across all

categories in all three datasets (Figures 2, 3, 4), suggesting the generations are similar to the references. These results lead us to believe that token-based summarization metrics might also suffer from a "summarization-ese" effect: **the metrics could be biased towards simpler, more "extractive" references**. Recently, Freitag et al. (2020) also arrived at the same conclusion for machine translation and BLEU (Papineni et al., 2002).

In the next section, we continue to explore the reliability of these metrics.

4 Does the reliability of metrics change with data properties? (Q2 b)

For each document $d_i, i \in \{1 \dots n\}$ in a dataset $\mathcal{D}$, we have J system outputs, where the outputs can come from different systems. Let $s_{ij}, j \in \{1 \dots J\}$ be the j^{th} summary of the i^{th} document, m_i be a specific metric (including human judgment).

$$K^{sum}_{m_1 m_2} = \frac{1}{n} \sum_{i=1}^{n} \Big(K\big([m_1(s_{i1}) \dots m_1(s_{iJ})], [m_2(s_{i1}) \dots m_2(s_{iJ})]\big) \Big). \quad (1)$$

Correlation is calculated for each document, among the different system outputs of that document, and the mean value is reported. Like other meta-evaluation studies, we consider the Pearson correlation and Spearman correlation as measures for K. Due to space constraints we only show the Pearson plots for some critical results. More plots are available in Appendix A.1.

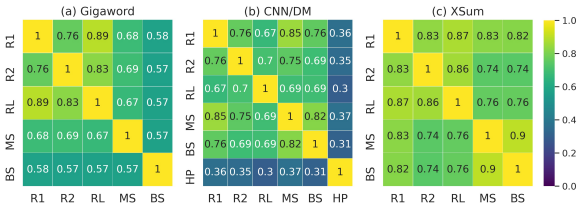
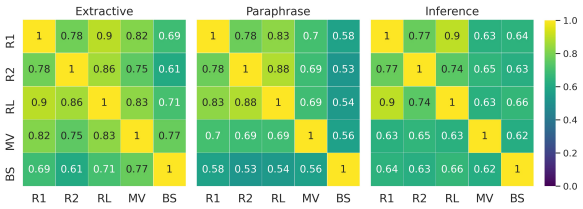
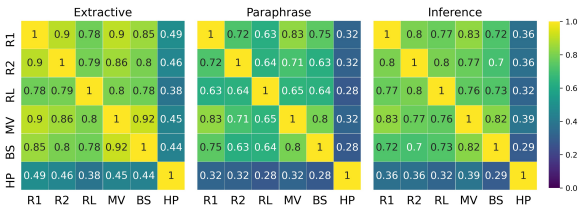


Figure 5: Pearson correlation between different metrics for all three datasets.

Inter-metric Correlation We present a pairwise correlation analysis of the automatic metrics to understand metric agreement in Figure 5. We conjecture that a strong correlation between two vastly different metrics (say ROUGE and MoverScore) might show that the metric is more reliable. Overall, we can see in Figure 5 that correlations between token-based metrics (ROUGE) and embedding-distance metrics (BERTScore, MoverScore) is lower in Gigaword, compared to CNN/DM and XSum. It is possible that the short length summaries of Gigaword is leading to this; perhaps there isn't enough context for BERTScore. Although, we could not find any results in the original papers to support this claim.



(a) Gigaword



(b) CNN/DM

Figure 6: Pearson correlations for Extractive, Paraphrase and Evidence samples in Gigaword and CNN/DM.

Correlation variation with complexity We observe that the correlation is heavily sample dependent. In Figure 5, averaged across all samples, R1 and MoverScore have a Pearson correlation of about 0.68 in Gigaword. This increases to 0.82 for the Extractive samples in Figure 6-(a), which are the simplest to reproduce. As the complexity increases, the correlation scores decrease (in Paraphrase, and then in Inference). The trends for R2 and MoverScore are similar. This is also observed for CNN/DM: in Figure 6-(b), correlations for R1-MoverScore and R1-BERTScore drop from 0.9, 0.85 for Extractive samples to about 0.83, 0.72 for Paraphrase and Inference samples. **This suggests that the inter-metric correlation is heavily sample dependent**. We cannot comment on XSum, because we did not encounter any Extractive samples in that dataset.

Correlation with Human Judgement For CNN/DM, we also compute the metric correlations with the human pyramid score (HP) in Figure 5 and

Figure 6-(b). We observe the highest agreement with the human-judgement for the Extractive subset, and it is significantly lower in Paraphrase and Inference. **This suggests that automatic metrics are more reliable when evaluating simpler examples, than complex ones**.

5 Discussion

Limitations of the typology. Forcing samples to have a single label did limit our analysis. In retrospect, the typology could have allowed for two labels: one for quality, one for complexity. In XSum for instance most samples which were labelled Entity Missing could also be labelled Paraphrase and Inference. We also realise that the impact of positional-bias could be important. This has been explored by Zhong et al. (2019a,b), and we plan to include similar metrics in our future work.

Collecting better datasets. Our results suggest that current metrics are not equally reliable across all categories of samples. If the quality of the references cannot be controlled, then having a diverse set of references for the source is also advised. This will allow for multi-reference evaluation and could offset the "summarization-ese" issues.

Limits of the Pyramid Scores. At the moment, the Pyramid Scores (and judgement criteria in general) only compare the output to the gold-reference, assuming the latter is *true*. As we see from our analysis, ignoring the source is not the right approach, for references from the web could have quality issues. A modified judgement procedure, that also accounts for the faithfulness of the gold-reference (perhaps by using automatic factuality metrics FactCC) might be better.

Architecture specific performance. In this study, we were interested in measuring the broader, *averaged* trends that summarization models exhibit. However, it would be interesting to see how specific architectural decisions impact individual model performance across different classes. We plan to explore this in the future.

"But what's the best metric for my data?" Specifically for metrics, our objective was to empirically demonstrate that (a) datasets have different modalities, and (b) metrics are not equally reliable across these modalities. In this process, we also observed some results suggesting possible biases in certain token-based metrics, and a need for diverse reference sets. We'll continue to explore this question.

6 Related Work

For the task of text-summarization, the data analysis heuristics presented in Zhong et al. (2019a,b); Bommasani and Cardie (2020); Grusky et al. (2018) are most relevant to our work. Their analysis is focused on surface level heuristics which ignores all noise present in the data. This has been discussed in Sections 2.2.1, 5. Researchers have also explored other dataset biases (Jung et al., 2019; Zhong et al., 2019b; Chen et al., 2020). As discussed in Section 5, we plan to include this in our future work.

For metric reliability and meta-analysis, we build on correlation analysis presented in earlier works (Peyrard, 2019; Bhandari et al., 2020; Fabbri et al., 2020). The key difference and novelty is the introduction of our typology and measuring the impact of sample complexity on model performance and metric reliability. To the best of our knowledge, metrics and models have not been evaluated on such a typology. As results in Section 3 and 4 show, sample complexity is indeed very critical for metric reliability.

7 Conclusion

In this study, we manually analysed 600 samples from three popular datasets, using a typology that captures data quality issues and varying degrees of sample-complexity. Our analysis of 27 summarization models reveals that the metric performance is heavily dependent on samples. On closer inspection, we found that the agreement of popular metrics also changes with the complexity, thus the scores might not reflect true model performance. This analysis also led to some suggestions for creating better summarization datasets and highlights some limitations of the current human-judgement procedures.

Acknowledgements

We thank Professor Graham Neubig, Yiran Chen and anonymous reviewers for valuable feedback and helpful suggestions. Thanks Kaiqiang Song for providing system outputs. This work was supported in part by a grant under the Northrop Grumman SOTERIA project and the Air Force Research Laboratory under agreement number FA8750-19-2-0200.

References

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the original: Fact aware neural abstractive summarization.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1662–1675.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Yiran Chen, Pengfei Liu, Ming Zhong, Zi-Yi Dou, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. CDEvalSumm: An empirical study of cross-dataset evaluation for neural summarization systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3679–3691, Online. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent.

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 (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada. Association for Computational Linguistics.

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

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, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.

Aishwarya Jadhav and Vaibhav Rajan. 2018. Extractive summarization with swap-net: Sentences and words from alternating pointer networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 142–151.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3324–3335, Hong Kong, China. Association for Computational Linguistics.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive Summarization of Reddit Posts with Multi-level Memory Networks. In *NAACL-HLT*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset.

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

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

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

Yixin Liu, Dou Ziyi, and Pengfei Liu. 2021. Refsum: Refactoring neural summarization. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, page 280.

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

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

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, AAAI'05, page 1436–1441. AAAI Press.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maxime Peyrard. 2019. 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.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

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. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741*.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Ren Liu, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8902–8909. AAAI Press.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*.

Kai Wang, Xiaojun Quan, and Rui Wang. 2019. BiSET: Bi-directional selective encoding with template for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2153–2162, Florence, Italy. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Mark Yatskar. 2019. A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2318–2323, Minneapolis, Minnesota. Association for Computational Linguistics.

Wonjin Yoon, Yoon Sun Yeo, Minbyul Jeong, Bong-Jun Yi, and Jaewoo Kang. 2020. Learning by semantic similarity makes abstractive summarization better. *arXiv preprint arXiv:2002.07767*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019a. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058, Florence, Italy. Association for Computational Linguistics.

Ming Zhong, Danqing Wang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2019b. A closer look at data bias in neural extractive summarization models. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 80–89, Hong Kong, China. Association for Computational Linguistics.

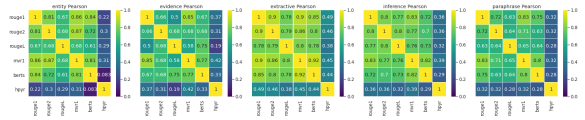
Appendix A Figures and Annotation Details

A.1 Correlation plots

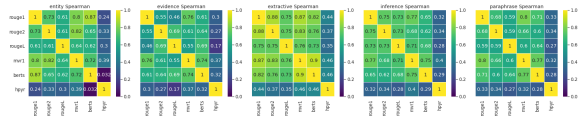
(a) Pearson

(b) Spearman

Figure 7: Gigaword correlations.

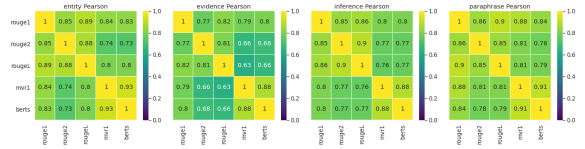


(a) Pearson

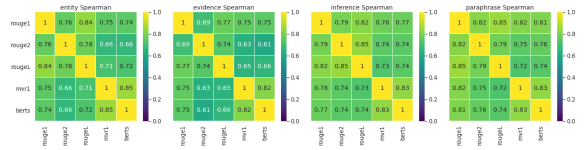


(b) Spearman

Figure 8: CNN/DM correlations.



(a) Pearson



(b) Spearman

Figure 9: XSum correlations.

A.2 Annotation Details

Each sample is annotated by 2-3 annotators independently. Given the limited number of samples, and the laborious nature of the exercise, we chose not to select final labels based on majority vote. For all disagreements, annotators discussed their reasoning and came to an consensus for final label. For 70% of Gigaword samples, 68% of CNN-DM samples, and 73% of XSum samples, the initial annotations were in agreement.

Appendix B Gigaword

B.1 Gigaword: Paraphrase and Inference samples

Label	Source
	SoTA Output
	Gold Reference
Paraphrase	A woman street cleaner and her three young daughters were killed Saturday when a bomb in a metal container exploded in Bangladesh , police said .
	Mother , three daughters die in in Bangladesh blast .
	Mother , three daughters killed in Bangladesh blast .
Paraphrase	The UN chief of Eastern Slavonia , the last Serb-held part of Croatia , confirmed Tuesday that key elections would be held here on April 13 as part of local ballots throughout Croatia .
	UN chief confirms key elections in Eastern Slavonia .
	UN confirms elections to be on April 13 in Eastern Slavonia .
Paraphrase	Business at Taiwan 's theme parks and resorts grew significantly in the first quarter of this year compared to Q1 last year , the Tourism Bureau said Thursday , attributing the growth to the government 's shopping voucher program and other promotion efforts .
	Business at Taiwan 's theme parks and resorts grows .
	Shopping vouchers help boost theme parks business : tourism bureau .
Inference	Col. Robert E. Lee skirted the unleaded gasoline pit , negotiated a thicket of telephone cords stretched as tight as trip wires and took the center of the New York Mercantile Exchange 's main trading floor just before 3 p.m. last Monday .
	New York Mercantile Exchange 's trading floor .
	MILITARY STRATEGISTS PRACTICE IN REAL BATTLE ON WALL STREET .
Inference	Finland scored three goals in a 40-second span of the first period Tuesday night for a 7-3 victory over the Czech Republic in their World Cup of Hockey opener .
	Finland 7 , Czech Republic 3 .
	Finland Routs Czech Republic at World Cup .
Inference	Q. I 've heard that cow manure can be used for energy production , but not human waste .
	Cow manure can be used for energy production .
	ON NOT WASTING WASTE .

Table 3: Source, outputs and targets, from Gigaword.

Appendix C CNN/DM

C.1 CNN/DM: Paraphrase and Inference samples

Label	Source SoTA Output Gold Reference
Paraphrase	Her neighbour's leylandii hedge stands 40ft tall and, says Audrey Alexander, has left parts of her garden in deep shade. What's more, it now seems likely to remain that way. (truncated). A row between neighbours over a 40ft high leylandii hedge (pictured) has finally come to and end after 35 years . The battle between the neighbouring properties started in 1980 when the owner planted a vegetable patch which withered and died in the shade of her neighbour's massive hedge . Then, 23 years ago, single mother Mrs Alexander bought the house and asked her neighbour Jeanette Robinson to trim the hedge. She claims Mrs Robinson refused and declared: 'I would rather move than touch these trees.' (truncated) . Audrey Alexander (pictured) also claims other neighbours have had to move their children from their bedrooms at night for fear of the falling branches . But her council has ruled that Mrs Robinson can keep the hedge, although it has to be cut to 20ft. Mrs Alexander said the ruling made 'no difference'. (truncated) Audrey Alexander 's vegetable patch withered and died in the shade of hedge . She asked neighbour Jeanette Robinson to trim it but she refused . Mrs Alexander claims hedge knocked £ 20,000 off the value of her house . Stirling Council has ruled that Mrs Robinson can keep the hedge . But it has to be cut to 20 ft , a height which she claims will still block most of her sunlight . Audrey Alexander wanted her neighbours to chop down their huge hedge . She claims the 40 ft leylandii was blocking sunlight from reaching her home . Feud started in 1980 when it blocked light from reaching a vegetable patch . Council finally rules that the hedge can stay - but must be cut back to 20 ft .
Paraphrase	The number of women in Britain becoming nuns is at a 25-year high. Figures from the Catholic Church show the number of women taking Holy Vows has trebled from 15 in 2009 to 45 last year. From a low of only seven in 2004, the figure has been rising for the past decade. Theodora Hawksley, 29, was until recently a post-doctoral researcher in theology at the University of Edinburgh. (truncate). Far from being trapped in traditional habits, Miss Hawksley said her order tends to dress down in T-shirts and jeans. Father Christopher Jamison, director of the National Office for Vocation of England and Wales, said: 'There is a gap in the market for meaning in our culture. One of the ways women may find that meaning is through religious life.' Sister Cathy Jones, religious life vocations promoter at the office, said: (truncated) . Figures from Catholic Church show number of women taking Holy Vows has trebled from 15 in 2009 to 45 last year . From a low of seven in 2004 , the figure has been rising for the past decade . Theodora Hawksley , 29 , was until recently a post - doctoral researcher in theology at the University of Edinburgh . But at the beginning of the year she decided to become a nun . Figures from the Catholic Church show more and more becoming nuns . The number of women taking Holy Vows stood at just seven back in 2004 . But that figure had risen to 15 in 2009 and increased further to 45 last year . One father said a ' gap in the market for meaning ' led people toward religion .
Inference	Following all his inspired charity work, Didier Drogba has been awarded with a Barclays Spirit of the Game trophy. The Chelsea forward set up the 'Didier Drogba Foundation in Africa,' as he hopes to inspire the next generation of footballers in Africa to fall in love with the game. (truncated) He said 'I come from a poor family where I played football in the streets with my friends with no shoes, there was no grass but we still enjoyed it. The 'Didier Drogba Foundation,' contribute financial and material support in education and health including school bags for the school children, as well as a medical clinic in his hometown of Abidjan, Ivory Coast, which will be opening its doors later this year. Chelsea's stars such as Eden Hazard, Petr Cech and Branislav Ivanovic were out in force earlier this month as they raises £400,000 for the foundation at a charity ball. The money raised will be used to complete the medical clinic in Abidjan and help finance mobile clinics that will travel outside of the capital to those who are either to sick or poor to make the journey to the medical centre. Didier Drogba has been awarded with a Barclays Spirit of the Game trophy . The Chelsea forward set up the ' DidierDrogba Foundation in Africa ' He hopes to inspire the next generation of footballers in Africa to fall in love with the game . The 37-year - old scored the equaliser against Leicester on Wednesday . Didier Drogba given the Barclays Spirit of the Game award . The 37-year - old 's foundation has done impressive work in Africa . Some of Chelsea 's stars attended a charity ball which raised £ 400,000 . CLICK HERE for all the latest Chelsea news .
Inference	(truncated) Resorts on its Black Sea coast offer the best value in terms of a meal out, buying a cup of coffee and essentials such as sun cream and a cold drink, according to a study. Scroll down for video . Affordable: Bulgaria has been named Europe's cheapest destination, with Black Sea resorts like Sunny Beach (pictured) offering the best value in terms of a meal out and other holiday activities . Hotspot: Bulgaria's most popular resort of Sunny Beach is a carbon copy of those of Spain and Greece . It is one of 13 European hotspots out of 14 where your cash will go far further this summer, largely thanks to rock-bottom exchange rates and higher inflation in some countries. Research into an imaginary shopping basket of ten typical holiday purchases showed a total price of £37.39 for Bulgaria, which is down by 13.6 per cent from last summer. There was a bigger fall of 22 per cent for the Algarve in Portugal, taking the total cost to £44.02, helping it beat Spain's Costa del Sol to become the second cheapest destination. Only in Turkey, where inflation is 7.6 per cent – compared to virtually zero in Britain and the eurozone – will Britons find the cost of a day out much more expensive. The figures, compiled for the annual Post Office Holiday Costs Barometer, show the spending basket in Turkey is up by 21.4 per cent on last year, at £65.70. Bulgaria's most popular resort of (truncated) . Former Soviet state has gained the most from the strong pound . Resorts on its Black Sea coast offer the best value in terms of a meal out , buying a cup of coffee and essentials such as sun cream and a cold drink . It is one of 13 European hotspots out of 14 where your cash will go far further this summer . Bulgaria's Black Sea resorts cheaper than hotspots in Italy, Spain and Turkey . Researchers found cheapest destination using 'imaginary shopping basket' Cheap prices are driven by low exchange rates and country's high inflation . Its most popular resort of Sunny Beach copies those of Spain and Greece .

Table 4: Source, outputs and targets, from CNN/DM.

Appendix D XSum

D.1 XSum: Paraphrase and Inference samples

Label	Source / SoTA Output / Gold Reference
Paraphrase	**Source:** More than 700,000 employees face unpaid leave due to the shutdown which was triggered after the two houses of Congress did not agree on a new budget. Hyundai said affected employees who currently own its vehicles will be given a payment relief "for as long as they are out of work". Employees looking to buy a new car will be given a 90-day payment deferral. "We recognize the impact on family budgets that the furlough will drive," John Krafcik, chief executive of Hyundai Motor America, said in a statement. Hyundai had offered a similar scheme, the Hyundai Assurance programme, during the peak of the global financial crisis four years ago to help consumers who had lost their jobs. Many analysts have said that the move had helped the South Korean firm win customer loyalty and boosted its sales in recent years. The company said that its latest offer to help the federal employees was an addition to that programme and aimed at "helping workers at a time when they most need it". "Like we did almost four years ago when we launched Hyundai Assurance, this is our way of saying 'We've got your back' during this uncertain time," Mr Krafcik said. Under the latest offer, Hyundai will extend all auto loan and lease payments during the shutdown for current Hyundai owners who are put on unpaid leave. The programme is available to all customers who have financed their purchase or lease through Hyundai Finance America. **SoTA Output:** US carmaker Hyundai Motor has offered financial help to federal employees who have been affected by the government shutdown . **Gold Reference:** Hyundai Motor will defer payments due from US federal employees affected by the partial government shutdown .
Paraphrase	**Source:** Gary Price was suspended from all council duties for five months in November after Powys council's Standards Committee ruled he had breached the code of conduct. His appeal has been dismissed by the Adjudication Panel for Wales following a two-day hearing in Llandrindod Wells. Mr Price has been contacted for comment. He was found to have sent information which the council said "incorrectly and unfairly" portrayed what happened at a grievance appeal hearing, in which he was a panel member. The Adjudication Panel for Wales unanimously agreed to refer the matter back to the Standards Committee with a recommendation that Mr Price be suspended for three months. Council leader Barry Thomas said the decision "sends out a clear message that those who enter public office have to operate within the members' code of conduct and maintain the highest possible standards". **SoTA Output:** A Powys council chief executive has lost his appeal against a decision to suspend him . **Gold Reference:** A decision to suspend a Powys county councillor has been upheld .
Inference	**Source:** Derby City Council wanted to shut Moorways Pool from April in a bid to save about Â£350,000 a year. The Labour-led authority, which needs to save Â£79m over the next three years, said it had found the savings by making cuts in other areas. Campaigners who gathered more than 4,000 signatures on a petition said they were delighted at the news. Ranjit Banwait, leader of the authority, said the council had committed to keep it open for a year. He said the council had identified savings "in back-office areas" and a restructuring of management jobs, which had been "untouched" since 2010. However, he stressed if the authority failed to get a "fair deal" from central government in the future, the pool would still have to close. Campaigners had accepted the pool, which is 33m in length, was in need of repair. There are plans for a new 50m pool to be built by 2018 to replace it. However, closing it would have left only one other public pool in the city - the Queen's Leisure Centre, they said. Doug Whitlam, of the Derbyshire Amateur Swimming Association, said: "One of the main things for me would have been the loss of teaching. "Twelve hundred young people use this facility every week and that would be lost forever." **SoTA Output:** A council has backed down over plans to close a public swimming pool in a bid to save money . **Gold Reference:** A Derby swimming pool threatened with closure is to remain open for another year , council bosses have confirmed .
Inference	**Source:** It is likely to include a scrappage scheme for older diesel cars in areas with high levels of dirty air. Speed bumps could be removed in some cities to cut pollution from cars slowing down and speeding up. Environmental lawyers ClientEarth said they would "thoroughly analyse" the proposals. According to the Royal College of Physicians, air pollution across the UK is linked to around 40,000 premature deaths every year. The UK has struggled to keep within EU limits on some pollutants, particularly nitrogen dioxide (NO2), which is produced by diesel engines and is linked to a range of respiratory diseases including asthma. Some 37 of the 43 regions of the UK are in breach of NO2 limits. Under earlier government plans, some parts of the UK would not have met EU NO2 standards until 2030. The original deadline to achieve these limits was 2010. Exasperated by what they believed was government foot-dragging on the question of cleaner air, ClientEarth mounted a legal challenge to force faster action. In April 2015, the UK Supreme Court ruled the government had to take immediate steps on the issue. Unhappy with the timescales in the plan that was then produced, ClientEarth went to the High Court last November for a judicial review. Once again the court supported the lawyers, telling the government that its scheme was "woefully inadequate" and giving ministers until 24 April this year to produce a new draft. With a general election in the offing, the government last week asked the judge for permission to delay the draft plan. But Mr Justice Garnham disagreed and ordered publication by 9 May. "These steps are necessary in order to safeguard public health," he said. Earlier this week, the government said it would not appeal against the ruling and would publish. In their previous plans, ministers wanted to create "clean air zones" in five cities outside London with high levels of NO2. Only the most polluting vehicles would have to pay a charge to enter the zone under that scheme. The new draft plan is expected to create many more such zones. Councils will be given the power to impose fines or restrictions on all polluting vehicles in these areas. In the worst cities, so called "toxin taxes" could range up to Â£20 a day but the government is said to be keen not to punish drivers who bought diesels as a result of incentives brought in by a previous Labour administration. This is something that the lawyers at ClientEarth support. "Successive governments have encouraged people to buy diesel. We don't want to see diesel drivers vilified, and we think the plans should also include properly funded incentives to help people move to cleaner forms of transport," said ClientEarth CEO James Thornton. "We will thoroughly analyse the government's draft plans when they are produced. If we do not think they are in line with the court order, to deal with illegal levels of pollution as soon as possible, then we will consider our next steps." According to newspaper reports, the government has agreed to back a "targeted" scrappage scheme for older diesel cars, but limited to vehicles in areas of high pollution. There may also be funding for a retrofitting scheme to help existing diesel car and van owners cut their emissions of NO2. The government is also said to be pushing for councils to use alternatives to charging, including the removal of speed bumps in some places and the better sequencing of traffic lights in others. Both of these measures could limit cars having to slow down and speed up repeatedly, actions that can almost double the amount of NO2 produced. However, the idea that speed bumps which slow down traffic would be sacrificed to help clean up the air we breathe is not a welcome concept according to road safety charity Brake. "We ought not to be made to choose between having cleaner air and safer roads," a spokesman said. "The evidence shows that air pollution is contributing to the early deaths of thousands of people. It's now clear that there's more than one way a car can kill you." The new proposals will be out for consultation for six weeks before the government produces a final plan at the end of July. Follow Matt on Twitter and on Facebook. **SoTA Output:** The government is expected to publish a new draft plan to tackle air pollution in the UK later this week . **Gold Reference:** The UK government is set to publish a draft air pollution plan after a protracted legal battle with environmental campaigners .

Table 5: Source, outputs and targets, from XSum.