Intrinsic Evaluation of Summarization Datasets

RISHI BOMMASANI

Stanford | NLP

CLAIRE CARDIE

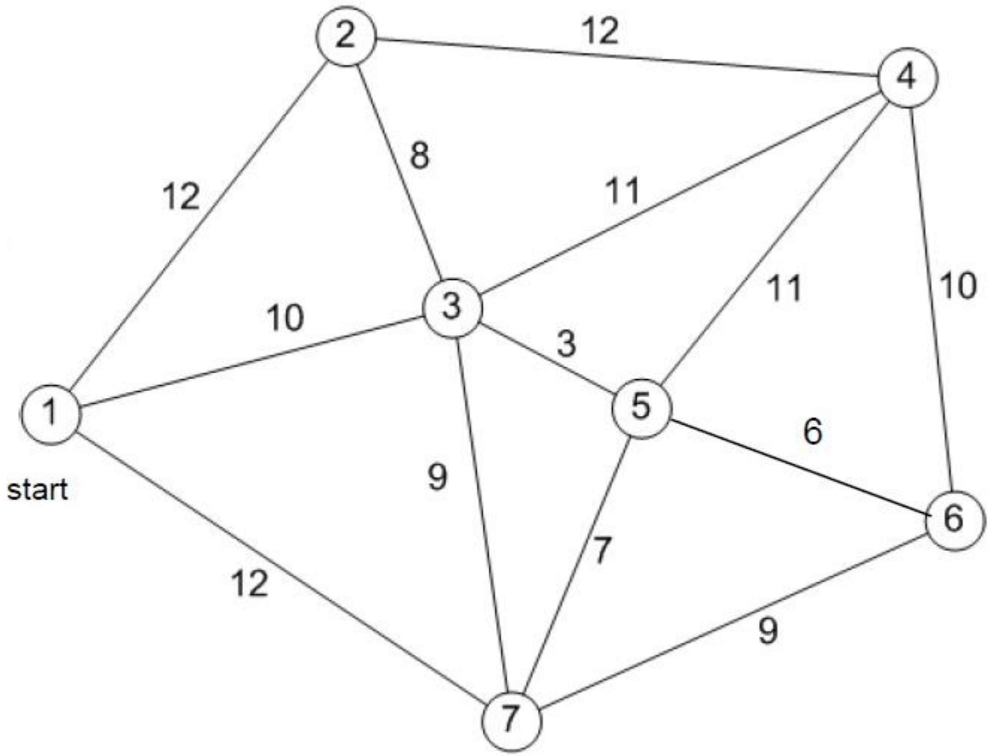
EMNLP 2020
KEYNOTE!!!

Revisit
Summarization Data

Ensuring data quality

- Crowdsourcing practices
  - NLVR, NLVR2 from Yoav Artzi's group at Cornell
  - Several commonsense datasets from Yejin Choi's group at UW/AI2
  - Adversarial filtering, human-in-the-loop generation
- WMT
- Penn Treebank

Is evaluating data quality easy?

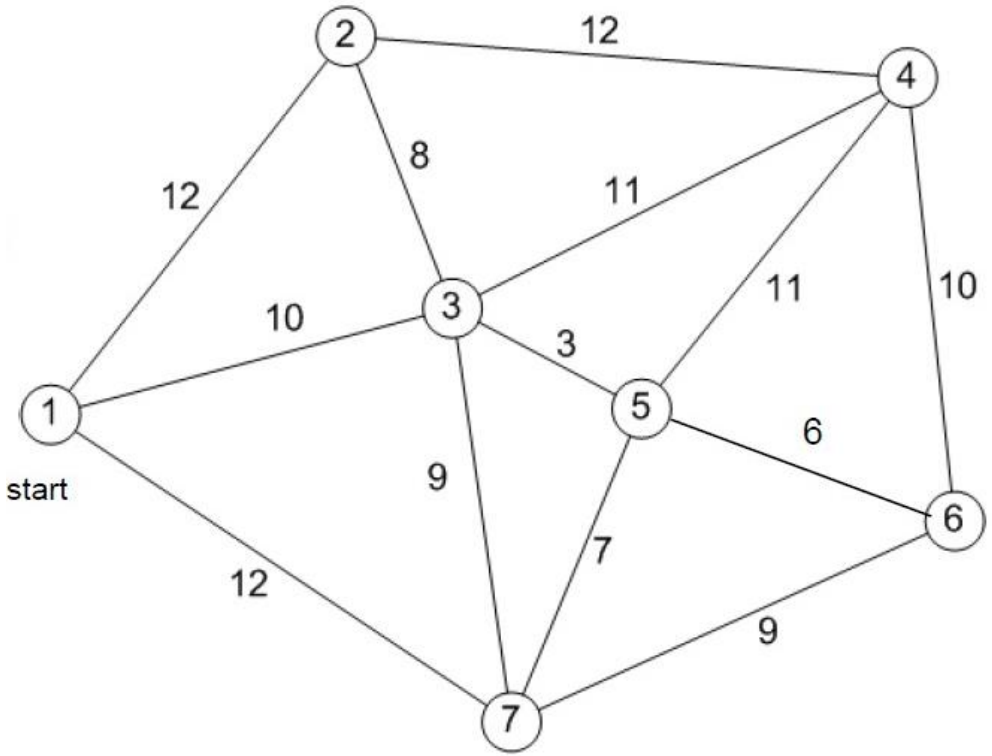


Travelling Salesman Problem

Find shortest cycle that starts and ends at 1, such that you reach every vertex.

NP-Hard

Is evaluating data quality easy?

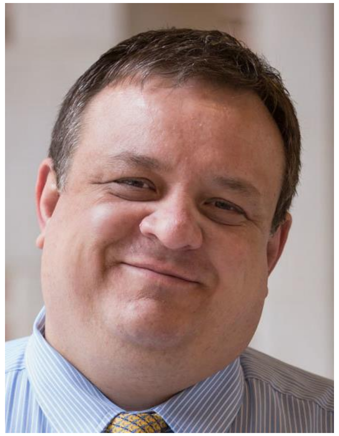


Trav. Sale. Decision Problem

Does there exist a solution to TSP of cost at most L?

NP-Hard






What do we do?

- Identify important aspects of summarization data

- Measure each of these aspects

Properties we measure

1. Compression **(CMP)**
2. Abstractivity **(ABS)**
3. Topic Similarity **(TS)**
4. Redundancy **(RED)**
5. Semantic Coherence **(SC)**

A taste of the measures: Topic Similarity

$$\mathbf{TS}\left(D_i, S_i\right) = 1 - \mathrm{JS}(\theta_{D_i|\mathcal{M}}, \theta_{S_i|\mathcal{M}})$$

M – LDA topic model learned on reference documents in dataset

JS – Jensen-Shannon distance

θ – Inferred topic distributions under *M*

A taste of the measures: Semantic Coherence

$$\mathbf{SC}\left(S_i\right) = \frac{\sum_{j=2}^{||S||} \mathbb{1}_{\mathrm{BERT}\left(S_i^j \mid S_i^{j-1}\right)}}{||S_i|| - 1}$$

S_i^j – Sentence *j* in summary *i*

$\boldsymbol{BERT}\left(S_i^j | S_i^{j-1}\right)$ – Next-sentence prediction under $\boldsymbol{BERT}$ of S_i^j given S_i^{j-1}

$||S_i||$ – Length of summary S_i in sentences

Compression captures dataset type

	News					Scientific		Social Media	Meeting	Script
	CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
# ex.	287K	655K	995K	3804K	203K	9963	21K	3084K	97	1061
avg. $\lvert D_i \rvert$	717	822	677	34	438	1203	2394	238	6020	28K
avg. $\lvert S_i \rvert$	50	46	40	9.6	24	160	270	27	314	122
avg. $\lVert D_i \rVert$	31	34	26	1	19	54	95	11	568	3156
avg. $\lVert S_i \rVert$	3.52	1.00	1.75	1.00	1.00	6.10	10.0	1.71	17.1	5.14
CMP$_w$	0.909	0.869	0.910	0.714	0.904	0.763	0.870	0.876	0.941	0.994
CMP$_s$	0.838	0.915	0.890	0.001	0.902	0.765	0.874	0.811	0.964	0.998
TS	0.634	0.586	0.539	0.478	0.578	0.702	0.774	0.438	0.573	0.547
ABS$_1$	0.135	0.249	0.191	0.334	0.346	0.201	0.122	0.384	0.184	0.147
RED	0.157	-	0.037	-	-	0.168	0.17	0.056	0.215	0.152
SC	0.964	-	0.981	-	-	0.994	0.990	0.961	0.968	0.983

Table 1: **Upper half**: Standard dataset statistics. **Lower half**: Aspect-level scores for each dataset (0 is minimal value, 1 is maximal value). Corresponding standard deviations appear in Table 9. Redundancy and semantic coherence are not reported for datasets with $>95\%$ single-sentence summaries.

Mismatch between dataset and modelling (abstractive vs. extractive)

	News					Scientific		Social Media	Meeting	Script
	CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
# ex.	287K	655K	995K	3804K	203K	9963	21K	3084K	97	1061
avg. $\lvert D_i \rvert$	717	822	677	34	438	1203	2394	238	6020	28K
avg. $\lvert S_i \rvert$	50	46	40	9.6	24	160	270	27	314	122
avg. $\lVert D_i \rVert$	31	34	26	1	19	54	95	11	568	3156
avg. $\lVert S_i \rVert$	3.52	1.00	1.75	1.00	1.00	6.10	10.0	1.71	17.1	5.14
CMP$_w$	0.909	0.869	0.910	0.714	0.904	0.763	0.870	0.876	0.941	0.994
CMP$_s$	0.838	0.915	0.890	0.001	0.902	0.765	0.874	0.811	0.964	0.998
TS	0.634	0.586	0.539	0.478	0.578	0.702	0.774	0.438	0.573	0.547
ABS$_1$	0.135	0.249	0.191	0.334	0.346	0.201	0.122	0.384	0.184	0.147
RED	0.157	-	0.037	-	-	0.168	0.17	0.056	0.215	0.152
SC	0.964	-	0.981	-	-	0.994	0.990	0.961	0.968	0.983

Table 1: **Upper half**: Standard dataset statistics. **Lower half**: Aspect-level scores for each dataset (0 is minimal value, 1 is maximal value). Corresponding standard deviations appear in Table 9. Redundancy and semantic coherence are not reported for datasets with $> 95\%$ single-sentence summaries.

Since datasets feature significant redundancy, we may need to post-process to deploy systems

	News					Scientific		Social Media	Meeting	Script
	CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
# ex.	287K	655K	995K	3804K	203K	9963	21K	3084K	97	1061
avg. $\|D_i\|$	717	822	677	34	438	1203	2394	238	6020	28K
avg. $\|S_i\|$	50	46	40	9.6	24	160	270	27	314	122
avg. $\|\|D_i\|\|$	31	34	26	1	19	54	95	11	568	3156
avg. $\|\|S_i\|\|$	3.52	1.00	1.75	1.00	1.00	6.10	10.0	1.71	17.1	5.14
CMP$_w$	0.909	0.869	0.910	0.714	0.904	0.763	0.870	0.876	0.941	0.994
CMP$_s$	0.838	0.915	0.890	0.001	0.902	0.765	0.874	0.811	0.964	0.998
TS	0.634	0.586	0.539	0.478	0.578	0.702	0.774	0.438	0.573	0.547
ABS$_1$	0.135	0.249	0.191	0.334	0.346	0.201	0.122	0.384	0.184	0.147
RED	0.157	-	0.037	-	-	0.168	0.17	0.056	0.215	0.152
SC	0.964	-	0.981	-	-	0.994	0.990	0.961	0.968	0.983

Table 1: **Upper half**: Standard dataset statistics. **Lower half**: Aspect-level scores for each dataset (0 is minimal value, 1 is maximal value). Corresponding standard deviations appear in Table 9. Redundancy and semantic coherence are not reported for datasets with $> 95\%$ single-sentence summaries.

Extra Results in Paper

1. Pairwise correlations between metrics
2. Standard deviations for metrics
3. Other metrics we tried for measuring same properties

What happens for “extreme” examples

Extreme – Top or bottom 10% of examples in a given dataset for a given metric

Manually examine several hundred such examples

Noticed that extremes often correlate with being generically low quality (for a conservative notion of low quality)

Fraction of low quality examples

	News					Scientific		Social Media	Meeting	Script
	CNN-DM	**NYT**	**NWS**	**GW**	**XSum**	**PeerRead**	**PubMed**	**TL;DR**	**AMI**	**MovieScript**
CMP_w ↑	50	50	70	60	30	10	10	80	0	10
ABS_1 ↑	40	30	70	50	50	70	50	80	0	10
CMP_w ↓	20	50	40	10	40	70	20	30	0	10
ABS_1 ↓	30	10	30	0	50	10	0	50	0	10

Table 3: **Upper half**: Percent of examples sampled from the top (↑) 10% for the given metric that were low quality. **Lower half**: Percent of examples sampled from the bottom (↓) 10% for the given metric that were low quality.

Original Text (truncated): Brodie (the dog) was neglected, and ended up with serious anger and health issues concerning his skin and allergies. My boyfriend adopted him ...

Summary: Onions.

Detector: Extremely High Compression

Figure 8: **Dataset: TL;DR**. We observe this trend quite frequently in **TL;DR**. Specifically, since authors on the social discussion platform Reddit choose to provide these summaries at their discretion, we often find the "summaries" are attention-grabbing and serve a starkly different rhetorical purpose from how summaries are generally conceived.

Original Text (truncated): these are external links and will open in a new window1908 - king carlos and eldest son assassinated in lisbon. second son manuel becomes king. 1910 - king manuel ii abdicates amid revolution ...

Summary: a chronology of key events :

Detector: Extremely High Compression

Figure 9: **Dataset: XSum**. We observe this trend quite frequently in **XSum**. For articles that are essentially timelines or other types of chronologies discussing historic events diachronically (which forms a small but distinctive section of the writing style of BBC from our analysis), the summary extracted to accompany it is generally this string or a slightly altered version. We argue this summary is fairly unhelpful (and is likely fairly uninteresting to test models on; simple rule-based filtering made be preferable to avoid overestimating performance on this dataset because of these examples).

Original Text (truncated): a lógica é o estudo dos princípios e critéiros de inferências e demonstrações válidas. um sistema lógico é composto por três partes: a sintaxe (ou notação), ...

Summary (truncated): logic is the science of correct inferences and a logical system is a tool to prove assertions in a certain logic in a correct way ...

Detector: Extremely High Abstraction

Figure 3: **Dataset: PeerRead**. This summary simply is not in the same language and hence achieves a very high abstractivity.

Original Text: BASEBALL American League BALTIMORE ORIOLES – Agreed to terms with INF-OF Mark McLemore on a minor league contract. BOSTON RED SOX – Named Dale Sveum third base coach.

Summary: Sports transactions

Detector: Extremely High Abstraction

Figure 5: **Dataset: NYT**. This summary is unlikely to be informative to someone who has not read the reference document and is more of a categorization/label than a summary. This is similar to the previous **NYT** example given.

Discussion

- How do we handle low quality examples
  - Prune, provide new reference summaries, keep as is, …
- Same types of ideas applied beyond summarization
- Only English language, single doc., single ref. summary

- **Higher scrutiny for future summarization datasets**
- **More deliberate in choosing datasets in modelling**