Towards Trustworthy Evaluation and Interpretation: Summarization & Dialogue

Chien-Sheng (Jason) Wu

2021.11.10 Eval4NLP @ EMNLP

Salesforce AI Research Team
Palo Alto, CA

Outline

- Re-evaluating metrics and models for text summarization
  - SummEval: Re-evaluating Summarization Evaluation
- Visualize and open the black box for model prediction
  - SUMMVIS: Interactive Visual Analysis of Models, Data, and Evaluation for Text Summarization
- Factual Consistency in Summarization and Dialogue

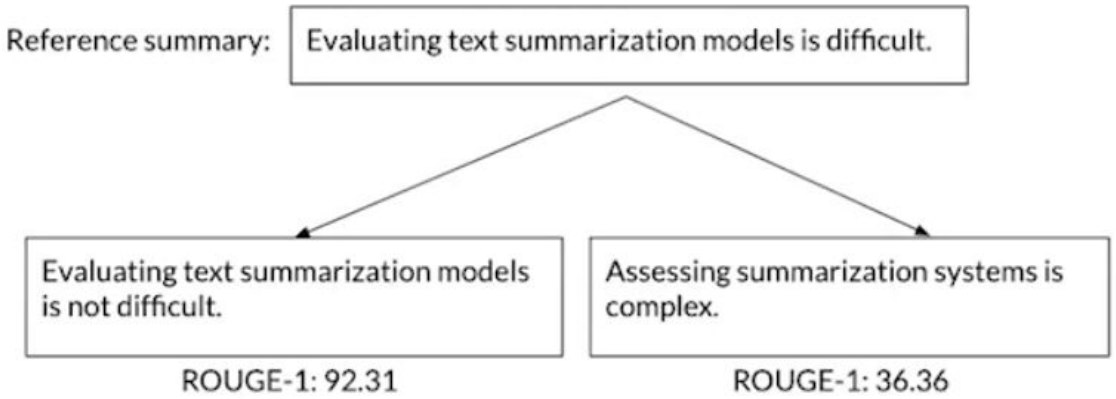
SummEval: Re-evaluating Summarization Evaluation

Fabbri et al., 2020

Summarization Evaluation

- ROUGE-N (Lin, 2004): Lexical overlap with a reference summary (or summaries)

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$



Need for Consistent Summarization Evaluation

- Model evaluation
    - Recent papers vastly differ in their evaluation protocol (e.g., different ROUGE packages) as noted in Hardy et al. (2019)
    - Most papers compare to only several other models
- Metric evaluation
    - New Metrics proposed but not widely adopted.
    - Metrics are evaluated on DUC and TAC shared tasks, not representative of modern systems (Peyrard, 2019).

Re-evaluating Metrics and Models

- Re-evaluating metrics
    - 14 automatic evaluation metrics
    - Toolkit with extensible and unified API
    - Largest and most diverse, in terms of model types, collection of human judgments of model-generated summaries on the CNN/DM dataset
- Re-evaluating models
    - Consistently benchmark 23 recent summarization models (2017 to 2019)
    - Largest collection of summaries on the CNN/DM news dataset for easier comparison.

Evaluation Metrics

- **ROUGE-based:** ROUGE (Lin, 2004b); ROUGE-WE (Ng and Abrecht, 2015); S3 (Peyrard et al., 2017)
- **Contextual Embedding-Based:** BertScore (Zhang* et al., 2020), MoverScore (Zhao et al., 2019), Sentence Mover's Similarity (Clark et al., 2019); SummaQA (Scialom et al., 2019)
    - **Reference-less:** BLANC (Vasilyev et al., 2020); SUPERT (Gao et al., 2020)
- **Machine translation, text generation metrics:** BLEU (Papineni et al., 2002); CHRF (Popović, 2015); METEOR (Lavie and Agarwal, 2007); CIDEr (Vedantam et al., 2015)
- **Data Statistics:** Grusky et al. (2018)

Summarization Models

Extractive Models	
NEUSUM (Zhou et al., 2018)	BanditSum (Dong et al., 2018)
LATENT (Zhang et al., 2018b)	REFRESH (Narayan et al., 2018)
RNES (Wu and Hu, 2018)	JECS (Xu and Durrett, 2019)
STRASS (Bouscarrat et al., 2019)	

Summarization Models

Non-pretrained Abstractive Models	
Pointer Generator (See et al., 2017)	ROUGESal (Pasunuru and Bansal, 2018)
Fast-abs-rl (Chen and Bansal, 2018)	Multi-task (Ent + QG) (Guo et al., 2018)
Bottom-Up (Gehrmann et al., 2018)	Closed book decoder (Jiang and Bansal, 2018)
Improve-abs (Kryściński et al., 2018)	SENECA (Sharma et al., 2019)
Unified-ext-abs (Hsu et al., 2018)	NeuralTD (Böhm et al., 2019)

Summarization Models

Pretrained Abstractive Models	
T5 (Raffel et al., 2019)	UniLM (Dong et al., 2019)
BertSum-abs (Liu and Lapata, 2019)	BART (Lewis et al., 2019)
GPT-2 (Ziegler et al., 2019)	Pegasus (Zhang et al., 2019a)

Human Judgments

- 100 articles from CNN/DM dataset; 16 models; 3 expert and 5 crowdsourced judgments

- 4 quality dimensions (rated from 1 to 5, higher better)
    - **Coherence** - the structure and organization of all summary sentences
    - **Consistency** - the factual alignment between summary and input
    - **Fluency** - the grammatical quality of individual sentences
    - **Relevance** - selection of important content from the source.

- Two rounds of expert annotations for better agreement (0.71 Krippendorf's alpha), problems with crowdsourced judgments

Problems with Crowdsourced Judgements

	Expert	Crowdworker
holidaymaker david meadwell recorded the unscheduled manouevre outside buckingham palace . he lost his footing and slid sideways , knocking bearskin on the side of the box . queen 's guard was left red-faced after he slipped on manhole cover . the entire incident was caught on a manhole cover . the embarrassed soldier quickly scrambled to his feet as his colleagues marched past .	Coh: 2.7 Con: 2.0 Flu: 4.7 Rel: 3.7	Coh: 3.2 Con: 3.4 Flu: 3.4 Rel: 4.0
buckingham palace guard slipped on manhole cover in front of hundreds of horrified tourists . the queen 's guard was left red-faced after he slipped on a manhole cover . he lost his footing and dropped his rifle on the side of the box and dropping his rifle . the incident was caught on camera camera camera . the guard is thought to have slipped because of metal shutters nailed to the soles of his boots .	Coh: 3.3 Con: 5.0 Flu: 1.7 Rel: 4.3	Coh: 3.0 Con: 3.2 Flu: 2.8 Rel: 3.2

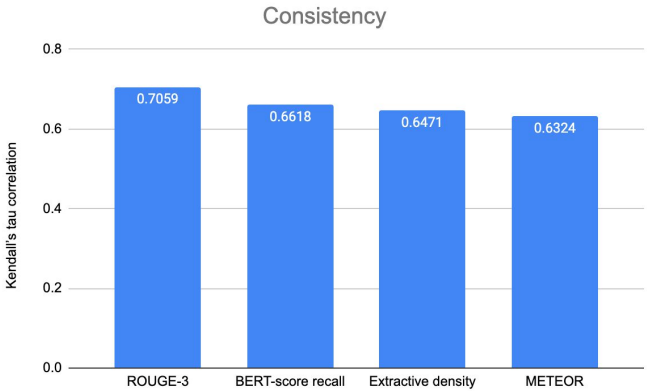
Metric Re-evaluation

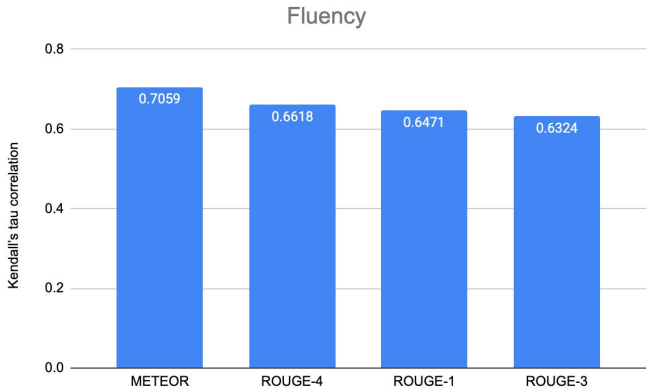
- Kendall's tau rank for system-level ranking as in Louis and Nenkova, 2013

Metric	Coherence	Consistency	Fluency	Relevance
ROUGE-1	0.2500	0.5294	**0.5240**	0.4118
ROUGE-2	0.1618	0.5882	0.4797	0.2941
ROUGE-3	0.2206	**0.7059**	**0.5092**	0.3529
ROUGE-4	**0.3088**	0.5882	**0.5535**	0.4118
ROUGE-L	0.0735	0.1471	0.2583	0.2353
ROUGE-su*	0.1912	0.2941	0.4354	0.3235
ROUGE-w	0.0000	0.3971	0.3764	0.1618
ROUGE-we-1	**0.2647**	0.4559	**0.5092**	**0.4265**
ROUGE-we-2	-0.0147	0.5000	0.3026	0.1176
ROUGE-we-3	0.0294	0.3676	0.3026	0.1912
S^3-pyr	-0.0294	0.5147	0.3173	0.1324
S^3-resp	-0.0147	0.5000	0.3321	0.1471
BertScore-p	0.0588	-0.1912	0.0074	0.1618
BertScore-r	0.1471	**0.6618**	0.4945	0.3088
BertScore-f	0.2059	0.0441	0.2435	**0.4265**
MoverScore	0.1912	-0.0294	0.2583	0.2941
SMS	0.1618	0.5588	0.3616	0.2353
SummaQA^	0.1176	**0.6029**	0.4059	0.2206
BLANC^	0.0735	0.5588	0.3616	0.2647
SUPERT^	0.1029	0.5882	0.4207	0.2353
BLEU	0.1176	0.0735	0.3321	0.2206
CHRF	**0.3971**	0.5294	0.4649	**0.5882**
CIDEr	0.1176	-0.1912	-0.0221	0.1912
METEOR	0.2353	**0.6324**	**0.6126**	**0.4265**
Length^	-0.0294	0.4265	0.2583	0.1618
Novel unigram^	0.1471	-0.2206	-0.1402	0.1029
Novel bi-gram^	0.0294	-0.5441	-0.3469	-0.1029
Novel tri-gram^	0.0294	-0.5735	-0.3469	-0.1324
Repeated unigram^	**-0.3824**	0.1029	-0.0664	-0.3676
Repeated bi-gram^	**-0.3824**	-0.0147	-0.2435	**-0.4559**
Repeated tri-gram^	-0.2206	0.1471	-0.0221	-0.2647
Stats-coverage^	-0.1324	0.3529	0.1550	-0.0294
Stats-compression^	0.1176	-0.4265	-0.2288	-0.0147
Stats-density^	0.1618	**0.6471**	0.3911	0.2941

Metric Re-evaluation

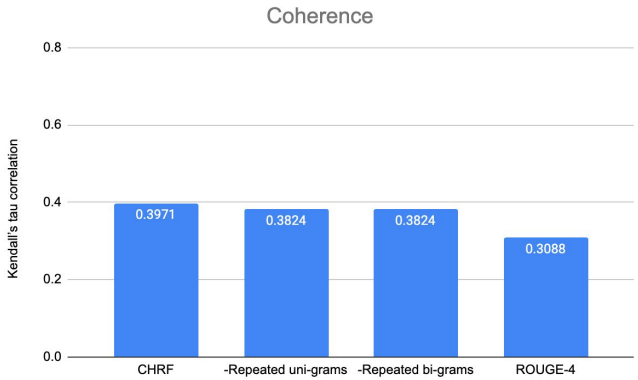
- Strong correlation with consistency and fluency perhaps due to **extractive** nature of dataset.

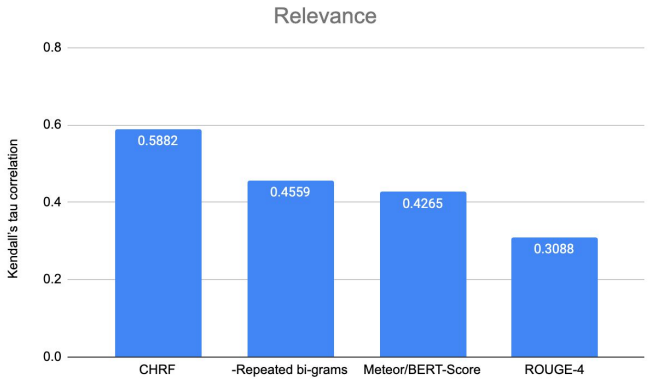




Metric Re-evaluation

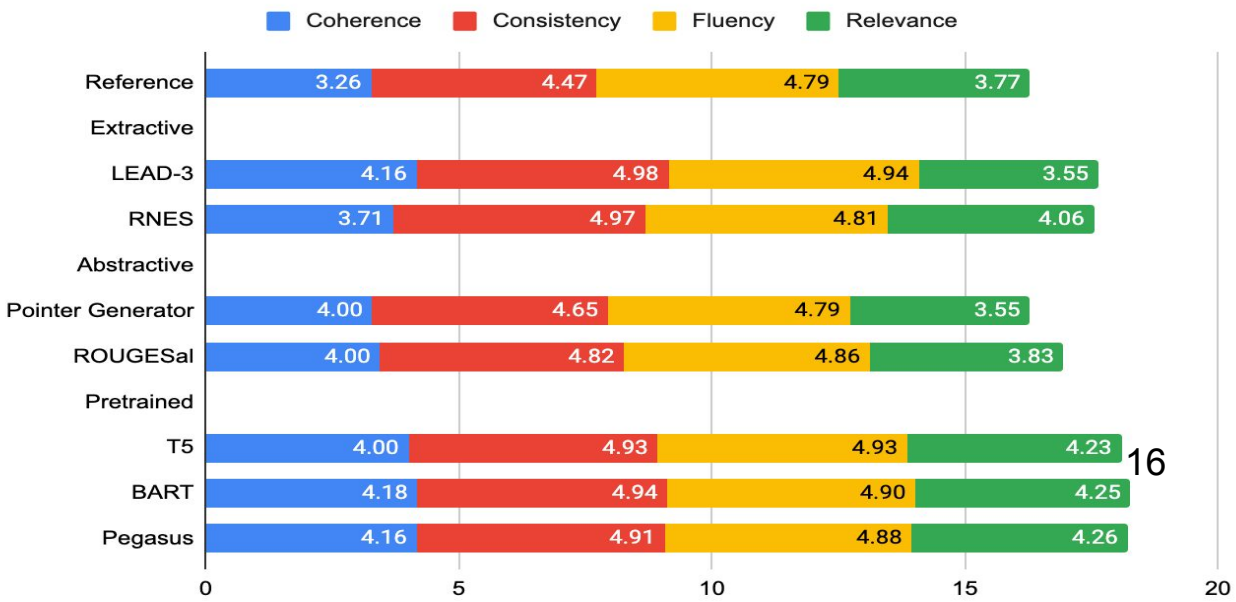
- Weaker correlations potentially to inherent subjectiveness of the dimension and the difficulty of collecting consistent human annotations





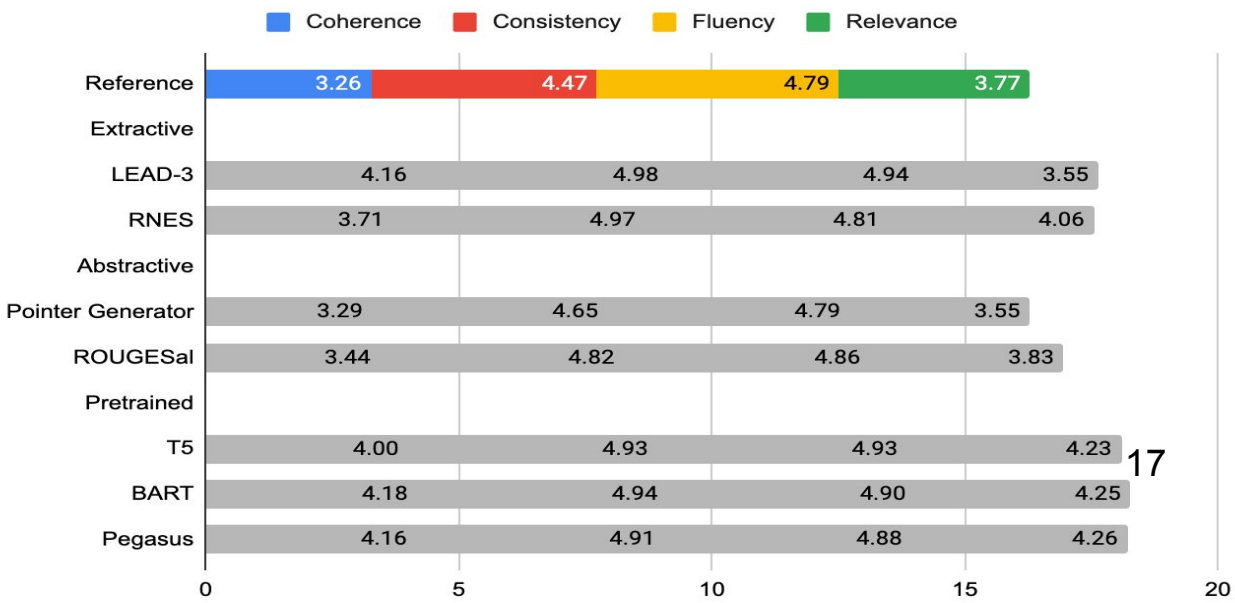
Model Re-evaluation

- Scores from 1 to 5 (best)



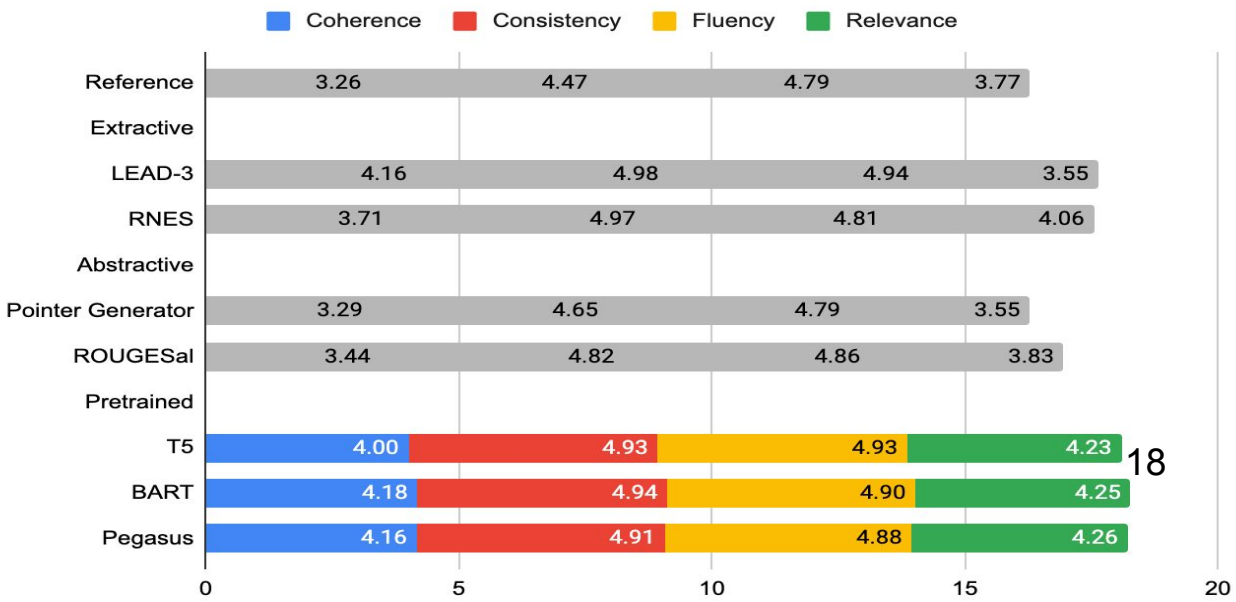
Model Re-evaluation

- Reference summaries are far from ideal



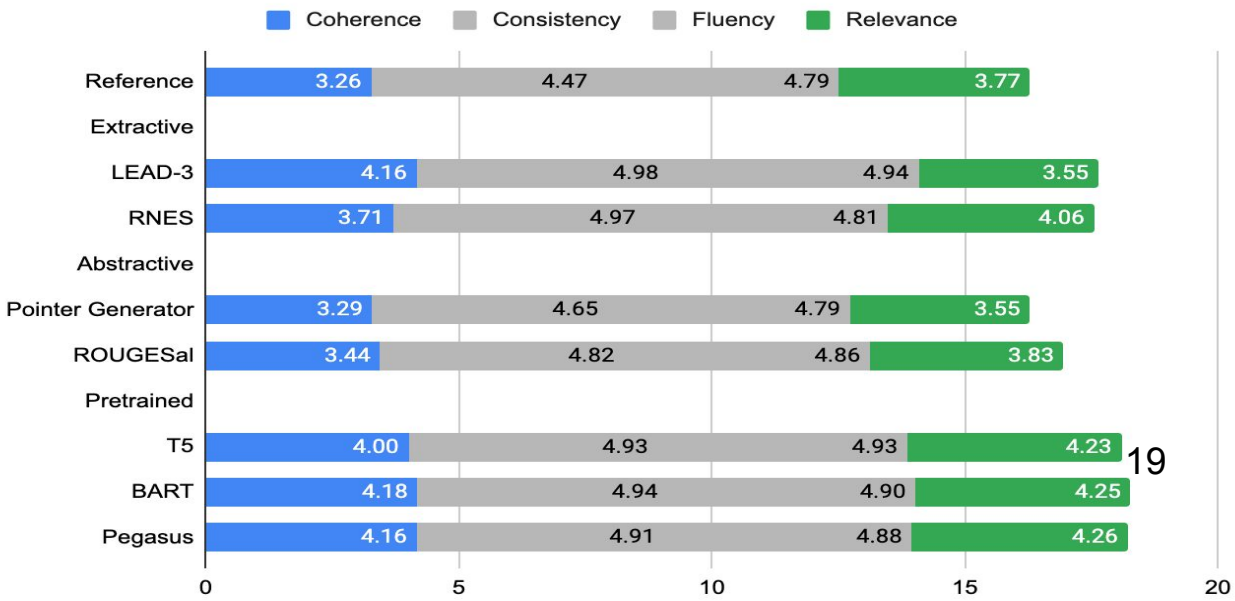
Model Re-evaluation

- Reference summaries are far from ideal
- Improvements with pretrained models



Model Re-evaluation

- Reference summaries are far from ideal
- Improvements with pretrained models
- Coherence and relevance can still be improved on this dataset



SummEval Toolkit

- Install

```
> pip install summ-eval
```

- Import

```
> from summ_eval.rouge_metric import RougeMetric
> rouge = RougeMetric()
```

- Evaluate!

```
> summaries = ["This is one summary", "This is another summary"]
> references = ["This is one reference", "This is another"]
> rouge_dict = rouge.evaluate_batch(summaries, references)
```

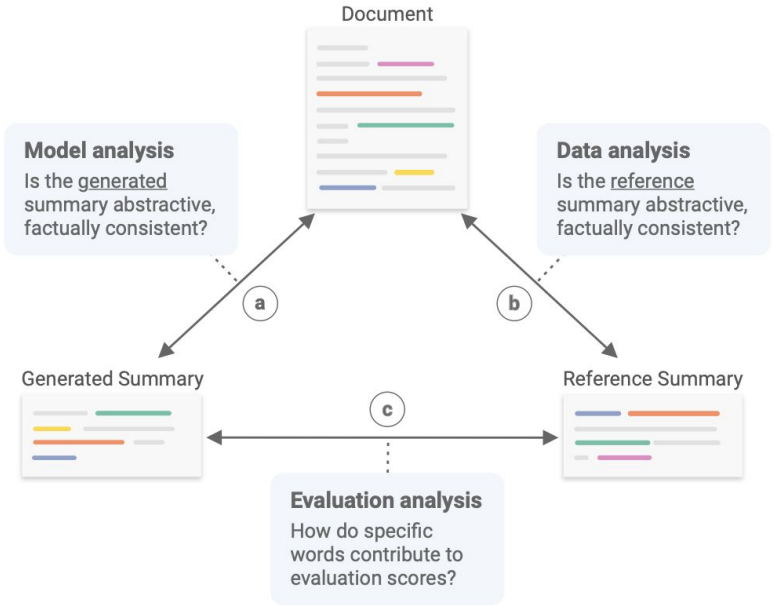
SummVis: Interactive Visual Analysis of Models, Data, and Evaluation

Vig et al., 2021

SummVis: Motivation

- Novel neural architectures, training strategies, and the availability of large-scale corpora haven been the driving force behind recent progress in abstractive text summarization.

- However, due to the **black-box** nature of neural models, **uninformative evaluation metrics**, and **scarce tooling** for model and data analysis, the true performance and failure modes of summarization models remain largely unknown.



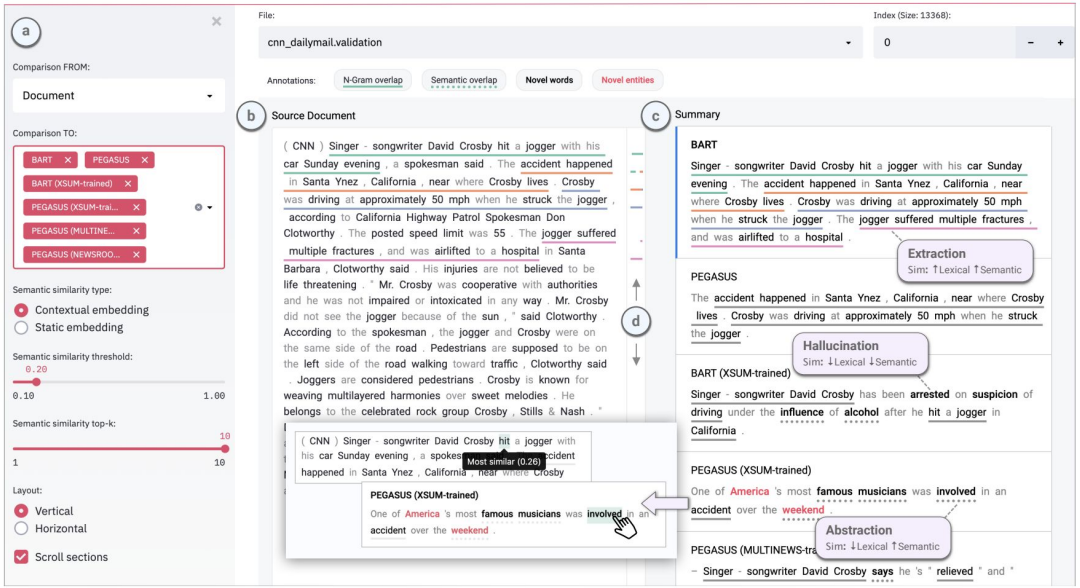
Tool Design

(a) Configuration panel

(b) Source document (or reference summary)

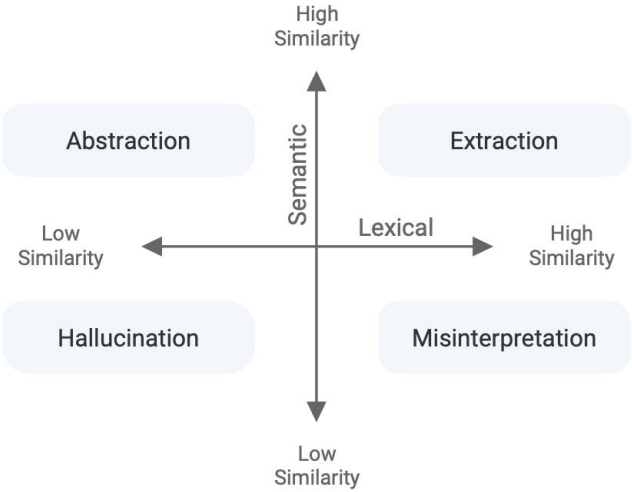
(c) Generated summaries (and/or reference summary)

(d) Scroll bar with global view of annotations.



Text Comparison

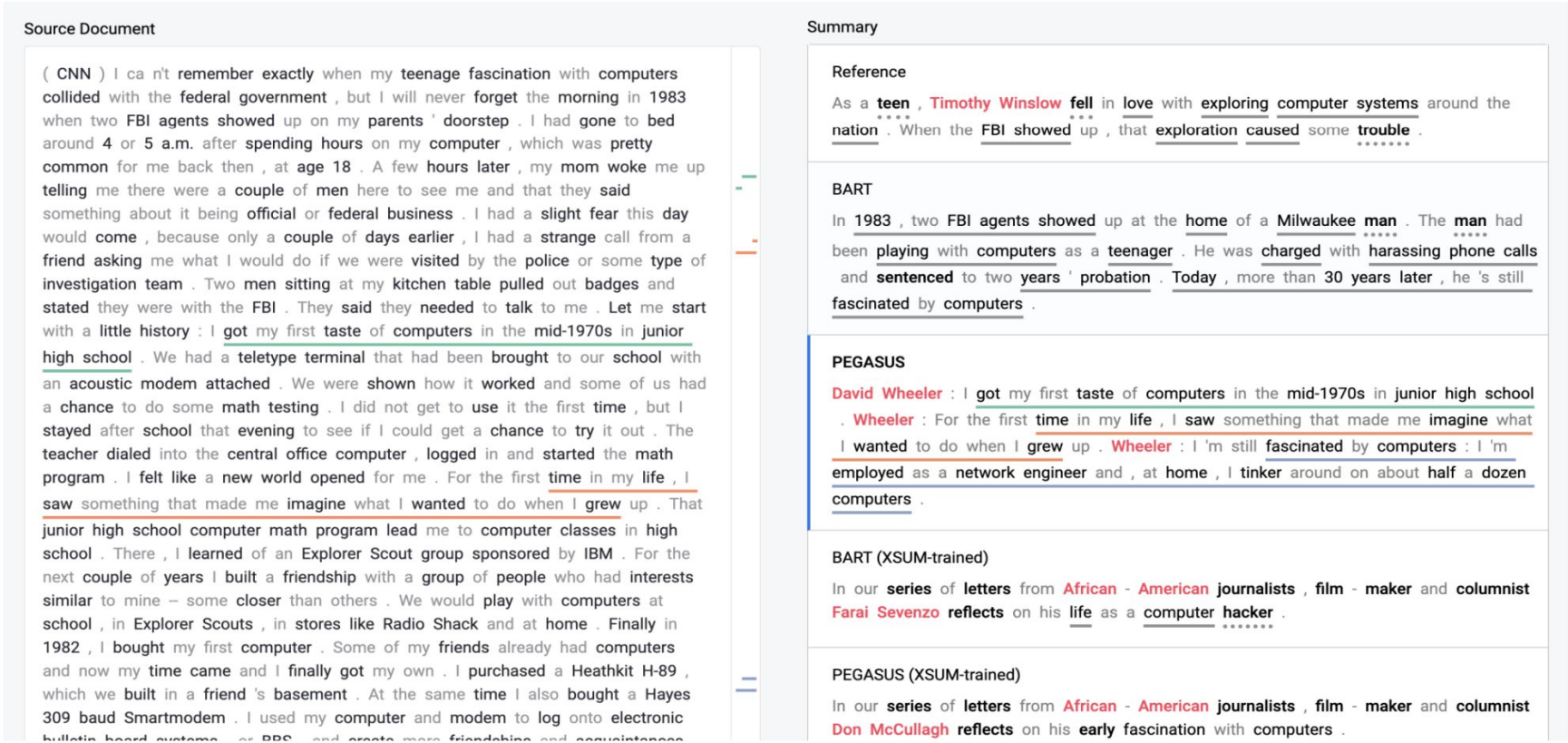
- Lexical Overlap
  - N-grams
- Semantic Overlap
  - Cosine similarity between word embeddings
    - Static word embeddings provided by spaCy (Honnibal et al., 2020)
    - Contextual embeddings from a pretrained RoBERTa (Liu et al., 2019) model
- Taxonomy
  - Two dimensions, four quadrants of behavior



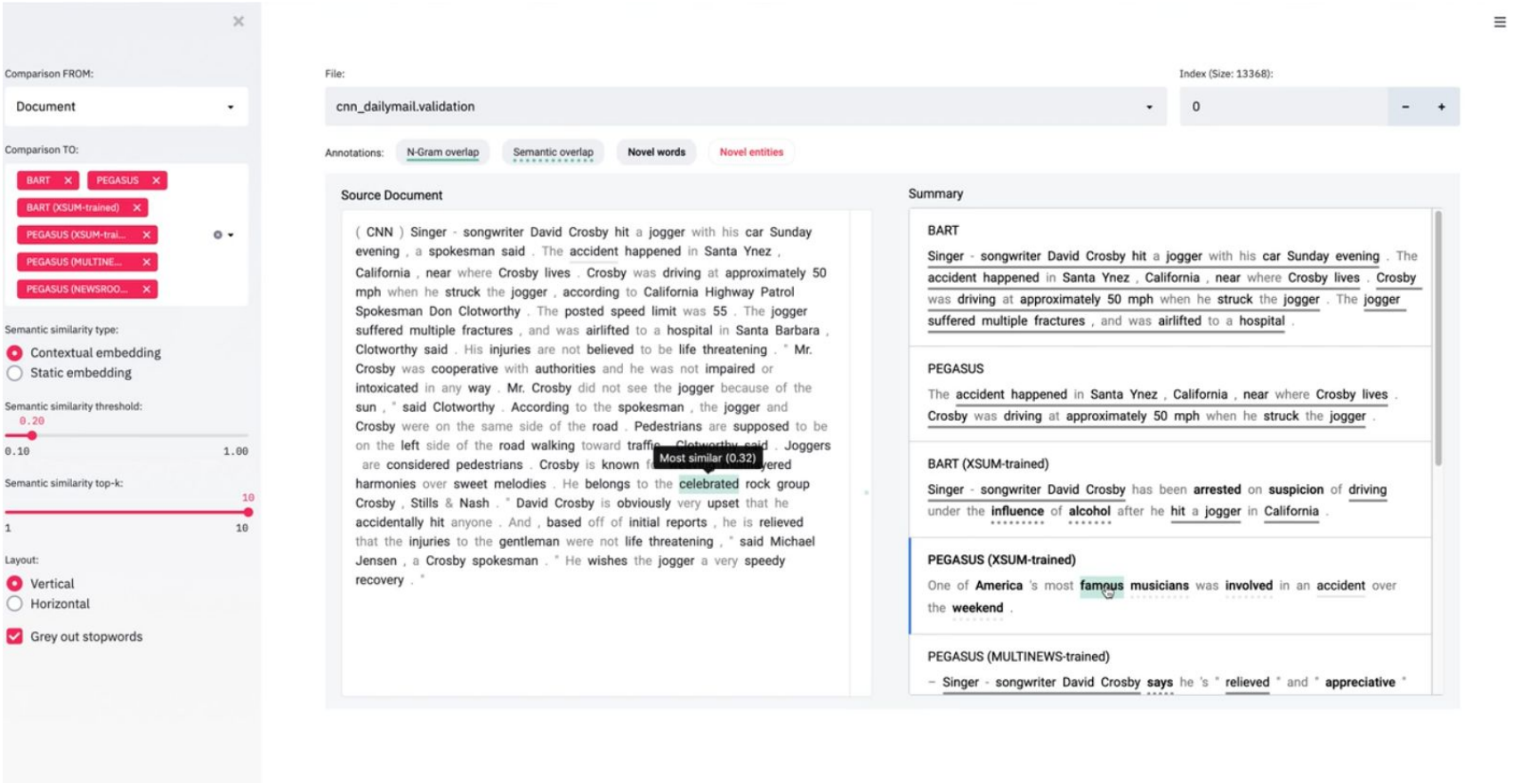
Tool Design

- Colored **underlines align n-grams** between source document and the selected summary.
- Novel words in the summary that do not appear in the source document are bolded, while **novel entities are bolded in red**.
- **Stopwords are grayed out** and are not used in the matching algorithms.
- **Dotted underlines** indicate tokens that are **semantically similar** to a token in the source document (above the threshold specified in the configuration panel).
- The user may hover over a token to see the most semantically similar tokens in the source document, or click on the token to auto-scroll the source document to the most similar token.

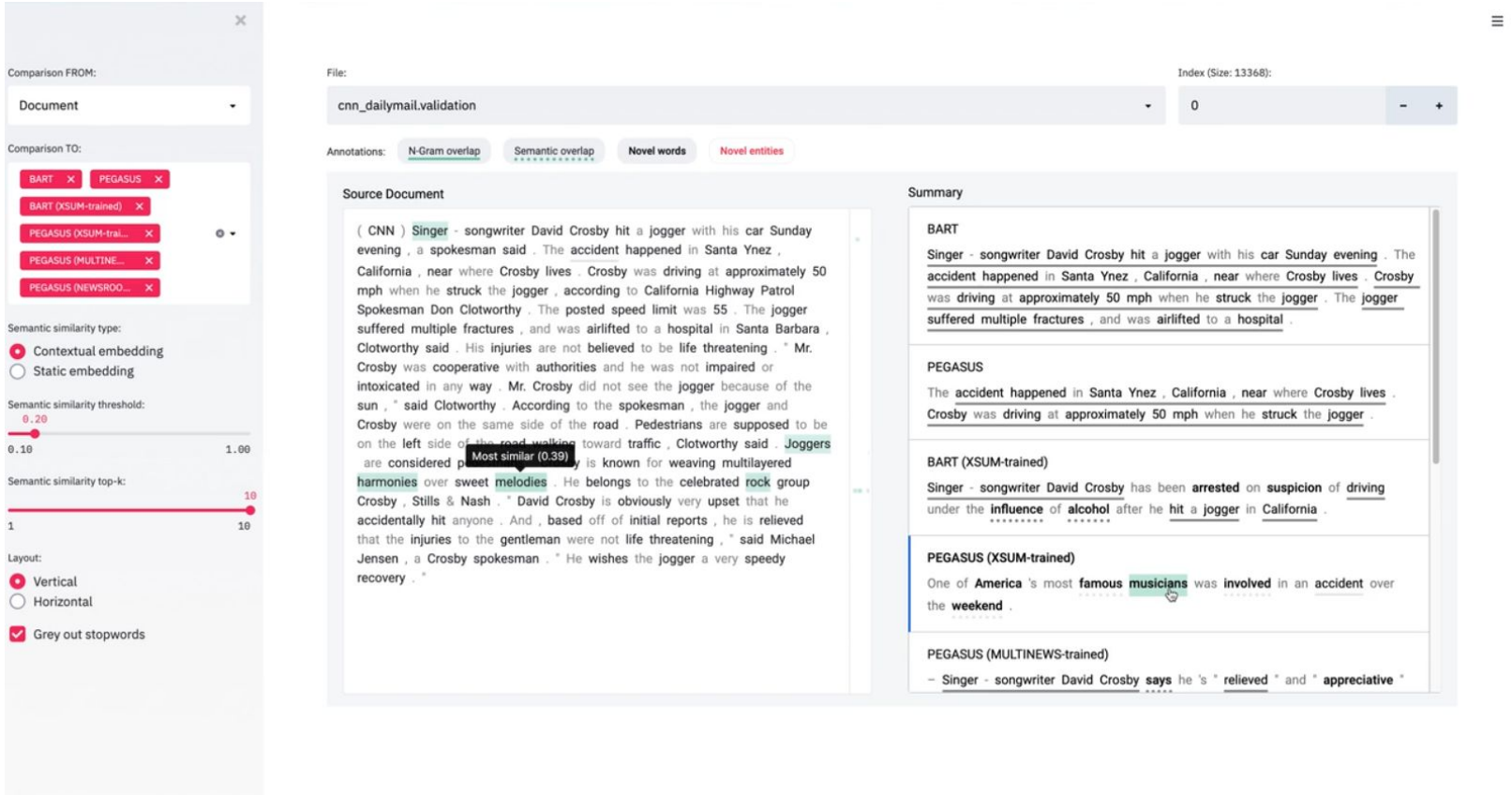
Case Study: Debugging Hallucination



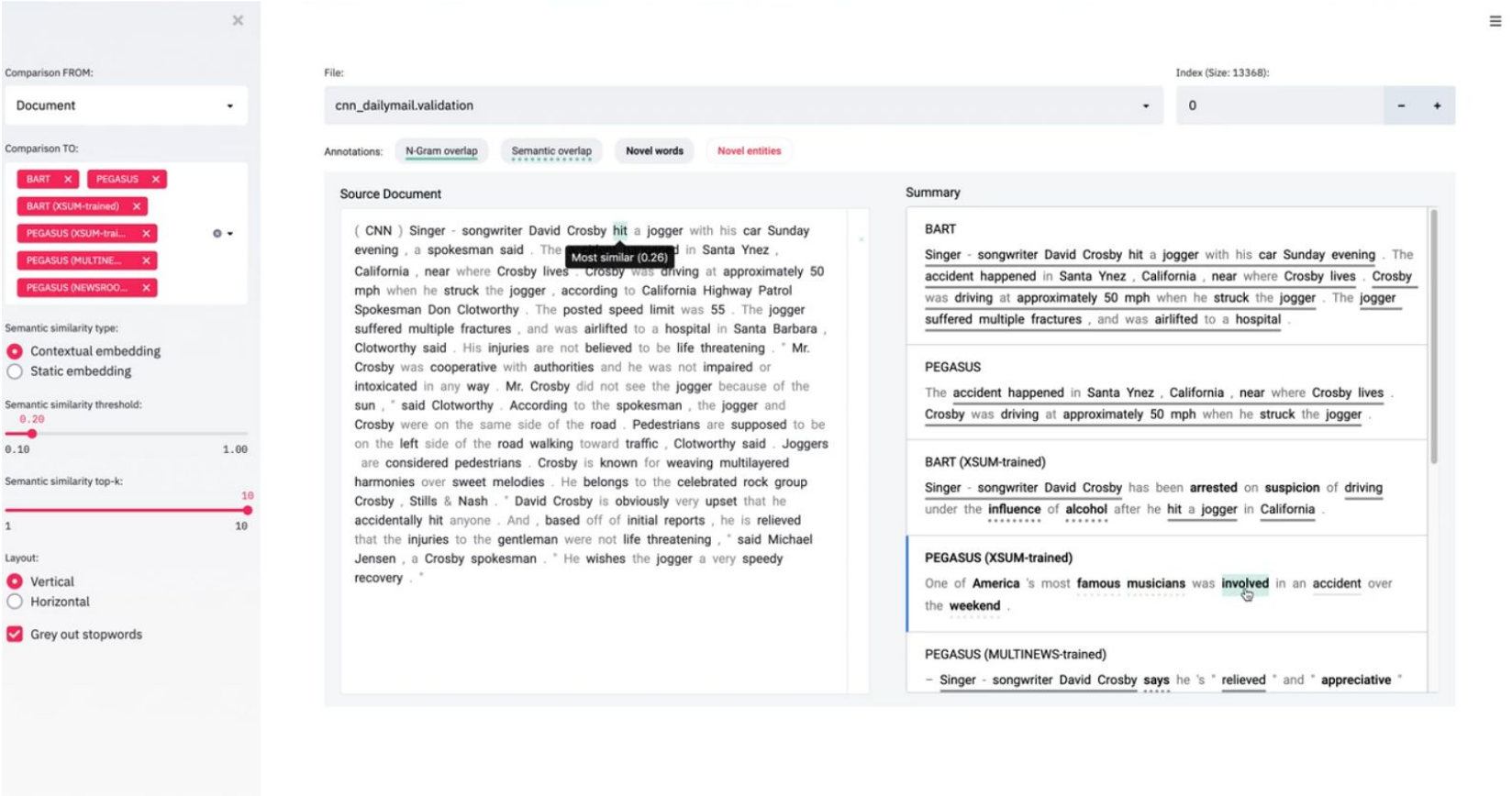
Case Study: Similarity



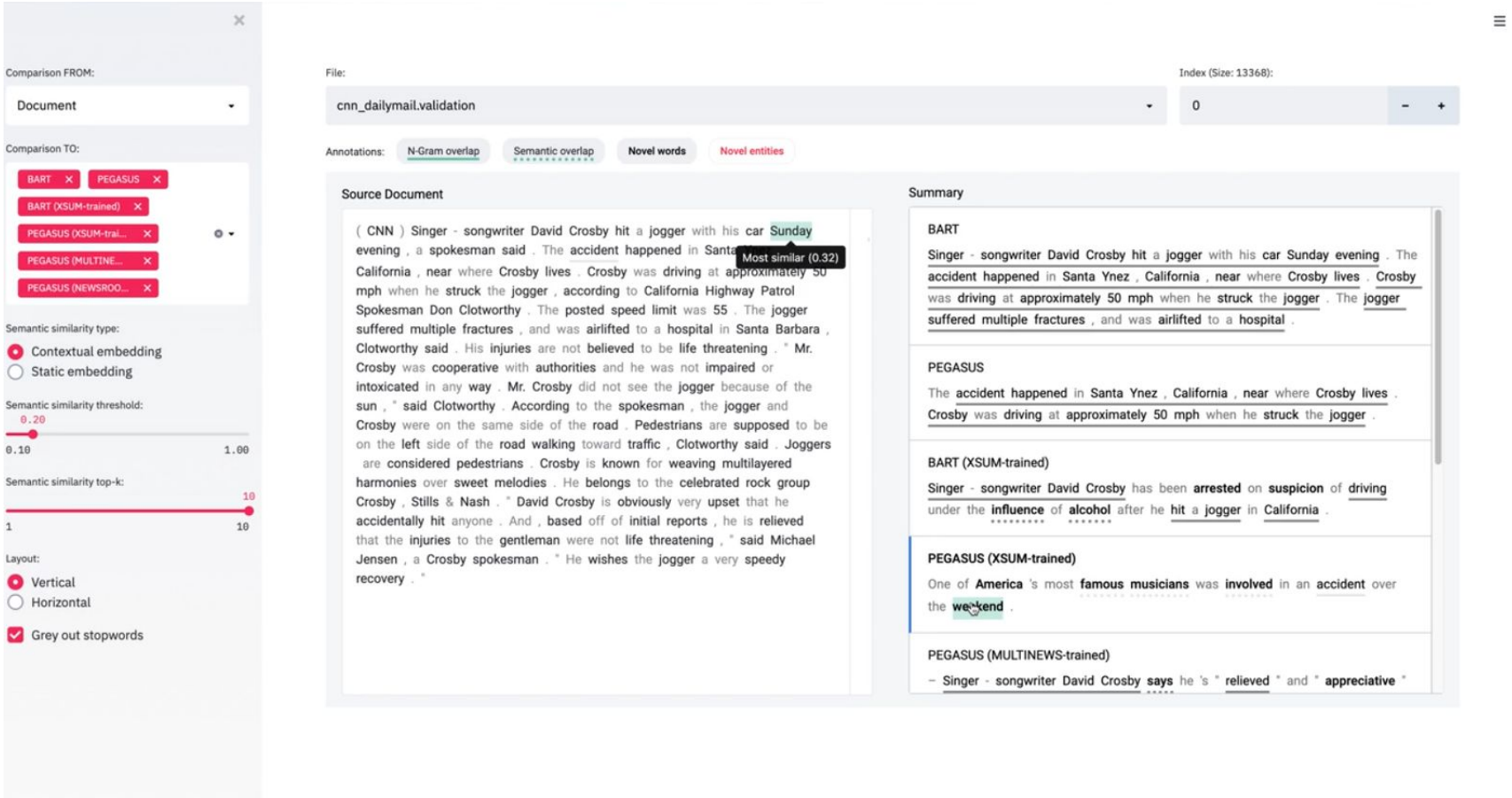
Case Study: Similarity



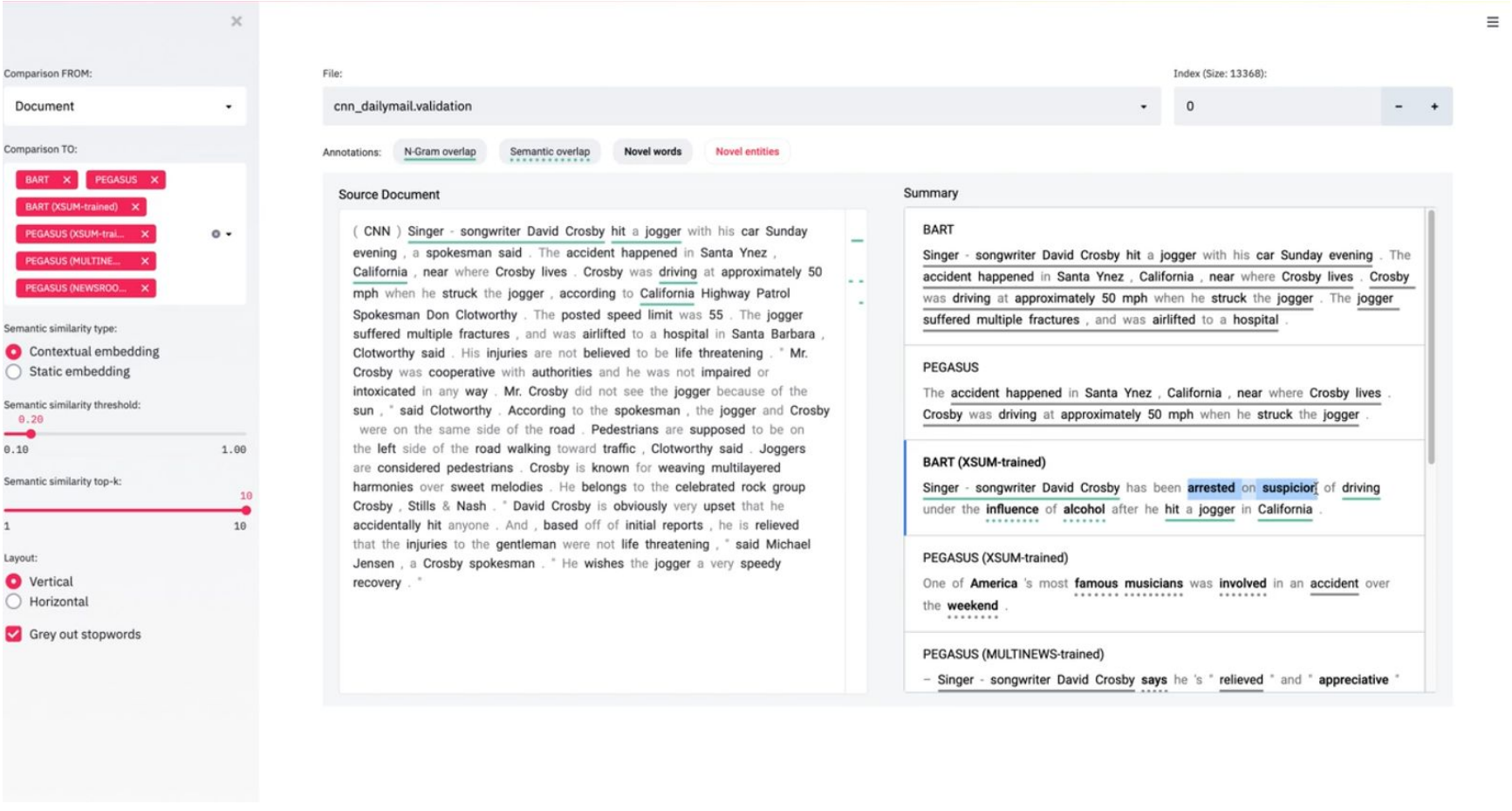
Case Study: Similarity



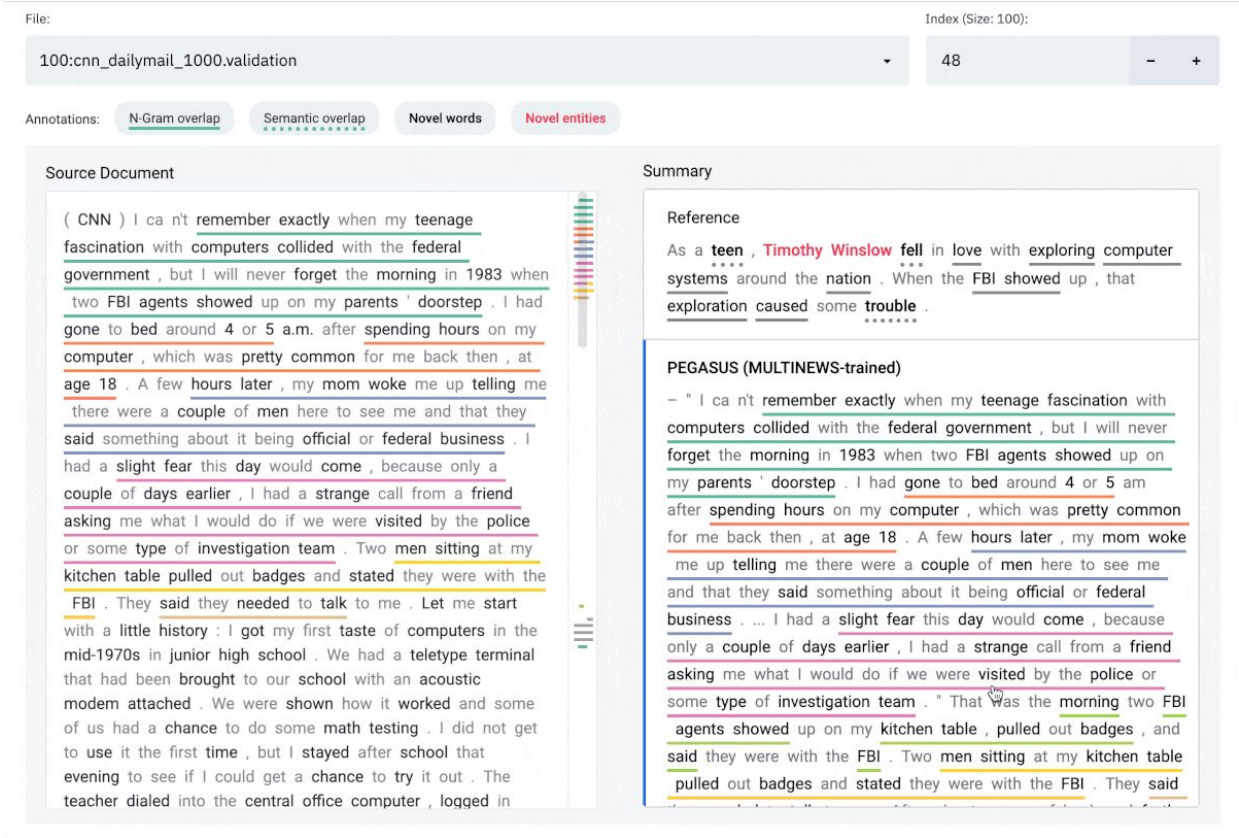
Case Study: Similarity



Case Study: Similarity



https://github.com/robustness-gym/summvis



Factual Consistency in Summarization / Dialogue

Preprint / Ongoing

Factual Evaluation: Summarization

1. Many interesting human evaluation datasets are being collected upon which evaluation metrics can be compared
    a. FRANK factuality benchmark
    b. FFCI: for focus, faithfulness, coherence, informativeness
    c. SummEval: coherence, consistency, fluency, relevance
    d. FALKE, Go Figure!, FEQA, QAGs, FactCC, On Faithfulness and Factuality, Entity-level factual consistency, Scarecrow
2. Differences in metric implementations lead to different conclusions
    a. QA > NLI: FEQA
    b. NLI > QA: On Faithfulness
    c. BertScore-based metrics > QA: FFCI
        i. For each summary sentence, uses average of top 3 sentences in the source. Then average over summary sentences.
    d. QA > BertScore-based metrics: QuestEval
        i. Likely uses reference-based BertScore
    e. BertScore and FactCC perform best: FRANK
        i. BertScore over source article but doesn't do sentence division/averaging over top sentences

Factual Evaluation: Summarization

- Are QA/QG approaches sensitive to different QA/QG models?
  - MixQG: Neural Question Generation with Mixed Answer Types
- What is the best way to select NLI-style comparison?
  - Average? Top-K? Sentence-level or token-level?
- Which metric is better under which setting?
- How to combine and leverage theses metrics for better summaries?

Factual Evaluation: Dialogue

- Fact-checking has been explored to verify formal single-sentence claims instead of casual conversational claims.
- Not all responses in a dialogue are carrying verifiable information.
- Kim et al. (2021) proposed the task of verification of colloquial claims which are stylistic modifications of the FEVER (Thorne et al., 2018) claims. However, their generated claims are not contextualized within dialogue.

FEVER: Google Search displays movie showtimes.
Colloquial Claim:
I can try google search to see
what movie to watch and get show times!

FEVER:
Unison (Celine Dion album) was originally released
by Atlantic Records.
Colloquial Claim:
I remember the Celine Dion album titled Unison.
It was released by Atlantic Records.

FEVER: Firefox is a desktop browser.
Colloquial Claim:
Yes, I use something called firefox for my desktop browser.

FEVER: Kung Fu Panda was released in theaters in 2006.
Colloquial Claim:
Have you watched Kung Fu Panda? It came out in 2006.

FEVER: San Francisco Bay Area contains many airports.
Colloquial Claim:
Sure, and yes there are lots of Bay Area airports!

Factual Evaluation: Dialogue

- Release a new evaluation dataset for factual evaluation in dialogue.
- Three sub-tasks: 1) verifiable claim detection, 2) evidence retrieval, and 3) claim verification.
- Point out the weaknesses of existing fact-checking models trained on non-dialogue domains.

Dialogue Context: I have family in Ireland! Have you ever been there?
Evidence: Ireland is an island in the North Atlantic.

Non-Verifiable Response: I haven't been but want to!
Verifiable Supported Response: I haven't. It is an island in the north Atlantic right?
Verifiable Refuted Response: I haven't been. Isn't it somewhere in north Pacific?
Verifiable NEI Response: I haven't been. I heard it's the most popular tourist location in Europe!

	Oracle-Evidence		Wiki-Evidence		DPR-Evidence	
Model	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
DNLI	42.5	35.4	38.1	31.1	37.0	29.1
DECODE	41.2	31.0	33.6	25.8	32.8	22.1
VitaminC	57.6	56.3	48.7	46..2	48.3	45.7
CorefBert-Colloquia	61.0	60.1	47.3	45.7	45.1	40.8
Colloquial	63.6	63.2	48.4	47.4	48.9	46.8

Validation

	Supported	Refuted	NEI-Factual	NEI-Personal	Total
Generated	1686	1047	150	1745	4628
Written	1656	2316	1836	0	5808
Total	3342	3363	1986	1745	10436

Test

	Supported	Refuted	NEI-Factual	NEI-Personal	Total
Generated	2446	1195	1274	1393	6308
Written	1493	2740	1268	0	5501
Total	3939	3935	2542	1393	11809

wu.jason@salesforce.com
Twitter: @SFResearch @jasonwu0731

Thank You