TEXT MINING

L08. SUMMARIZATION

SUZAN VERBERNE 2021

Universiteit Leiden

QUIZ ABOUT WEEK 7

- Name the kind of task that is used to learn word embeddings (e.g. by pre-training a BERT model)

a. Sequential processing

b. Language modelling

c. Recurrence

d. Semantic role labelling

QUIZ ABOUT WEEK 7

- Which statements about context in sequential mdoels are true?
  a. LSTMs are sequential models with longer memories than traditional RNNs
  b. BERT models can make use of longer contexts than LSTMs
  c. The attention mechanism in Transformer models has quadratic complexity relative to the input length
  d. The maximum input length for BERT models is limited by computational memory

QUIZ ABOUT WEEK 7

- Consider a text classification task for Chinese newspaper texts with limited labeled data (300 items). What do we need to build a classifier using transfer learning?

  a. A pre-trained Chinese BERT model

  b. A large collection of Chinese texts to train a BERT model

  c. GPU computing, a lot of time, and the 300 items for supervised learning

  d. GPU computing, a pre-trained Chinese BERT model and the 300 items for supervised fine-tuning

TODAY’S LECTURE

- Automatic summarization
  - Extractive summarization methods
  - Abstractive summarization methods
  - Challenges
  - Evaluation
- Argument mining (by Michiel)

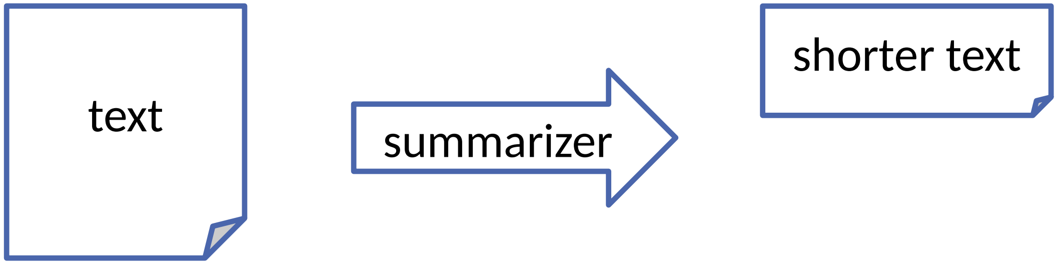
AUTOMATIC SUMMARIZATION

Universiteit Leiden

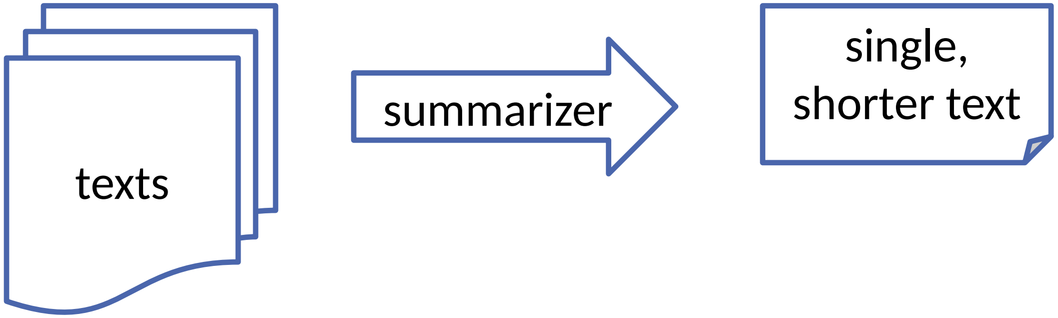
Suzan Verberne 2021

SINGLE-DOCUMENT VS. MULTI-DOCUMENT

- Single-document summarization



- Multi-document summarization



SINGLE-DOCUMENT VS. MULTI-DOCUMENT

- Single-document summarization:
  - News articles
  - Scientific articles
  - Meeting reports (minutes)
- Multi-document summarization:
  - Output of a search engine
  - News about a single topic from multiple sources

EXTRACT VS. ABSTRACT

- An extract is a summary composed completely of material from the source
- An abstract is a summary that contains material not originally in the source, but shorter paraphrases

SUMMARIZATION METHODS

EXTRACTIVE VS ABSTRACTIVE

Extractive summarization

Select the most important nuggets (sentences)

This is a classification or ranking task

- Classification: for each sentence, select it: yes/no
- Ranking: assign a score to each sentence, then select the top-k

EXTRACTIVE VS ABSTRACTIVE

Extractive summarization

Select the most important nuggets (sentences)

This is a classification or ranking task

- Classification: for each sentence, select it: yes/no
- Ranking: assign a score to each sentence, then select the top-k

Abstractive summarization

Learn a text-to-text-transformation model (cf. translation)

- Training data: pairs of longer and shorter texts
- Sequence-to-sequence models: Learning a mapping between an input sequence and an output sequence

EXTRACTIVE VS ABSTRACTIVE

Extractive summarization

- Feasible / easy to implement
- Reliable (literal re-use of text)
- But limited in terms of fluency
- and fixes required after sentence selection

EXTRACTIVE VS ABSTRACTIVE

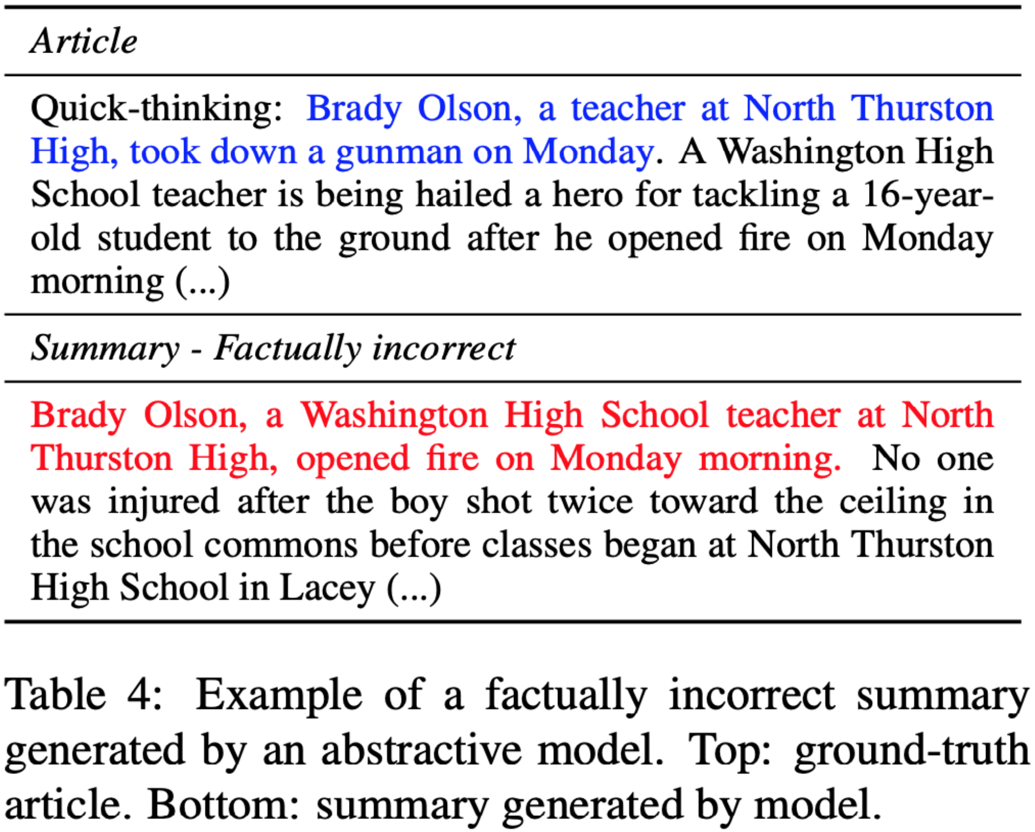
Extractive summarization

- Feasible / easy to implement
- Reliable (literal re-use of text)
- But limited in terms of fluency
- and fixes required after sentence selection

Abstractive summarization

- More natural/fluent result
- But a lot of training data needed
- And risk of untrue content

EXAMPLE OUTPUT



Article
Quick-thinking: Brady Olson, a teacher at North Thurston High, took down a gunman on Monday. A Washington High School teacher is being hailed a hero for tackling a 16-year-old student to the ground after he opened fire on Monday morning (...)
Summary - Factually incorrect
Brady Olson, a Washington High School teacher at North Thurston High, opened fire on Monday morning. No one was injured after the boy shot twice toward the ceiling in the school commons before classes began at North Thurston High School in Lacey (...)

Table 4: Example of a factually incorrect summary generated by an abstractive model. Top: ground-truth article. Bottom: summary generated by model.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

BASELINE SUMMARIZATION SYSTEM

- Take the first three sentences from the document
  - Strong baseline!
  - “State-of-the- art models only slightly outperform the Lead-3 baseline, which generates summaries by extracting the first three sentences of the source document.” (Kryściński et al, 2019)

EXTRACTIVE SUMMARIZATION

SENTENCE SELECTION METHODS

- Unsupervised methods:
  - Centrality-based
  - (Graph-based)
- Supervised methods:
  - Feature-based
  - (Neural-network based)

UNSUPERVISED SENTENCE SELECTION

- Centrality-based methods for sentence selection
  - Measure the cosine similarity between each sentence and the document
    - Either using the sparse vector space with words as dimensions
    - Or the dense vector space using embeddings representations
  - Select the sentences with the highest similarity (the most representative sentences)

SUPERVISED SENTENCE SELECTION

- Feature engineering + classifier (e.g. SVM)
- Features:
  - position in the document
  - word count
  - word lengths
  - word frequencies
  - punctuation
  - representativeness (similarity to full document/title)
  - etc.

PROBLEMS WITH SENTENCE SELECTION

- Selecting sentences that contain unresolved references to sentences not included in the summary or not explicitly included in the original document:
  - “Our investigations have shown this to be true.”
  - “There are three distinct methods to be considered.”

PROBLEMS WITH SENTENCE SELECTION

- Selecting sentences that contain unresolved references to sentences not included in the summary or not explicitly included in the original document:
  - “Our investigations have shown this to be true.”
  - “There are three distinct methods to be considered.”

- Improvements that might be needed after sentence selection:
  - Sentence ordering
  - Sentence revision
  - Sentence fusion
  - Sentence compression

ABSTRACTIVE SUMMARIZATION

SEQUENCE-TO-SEQUENCE LEARNING

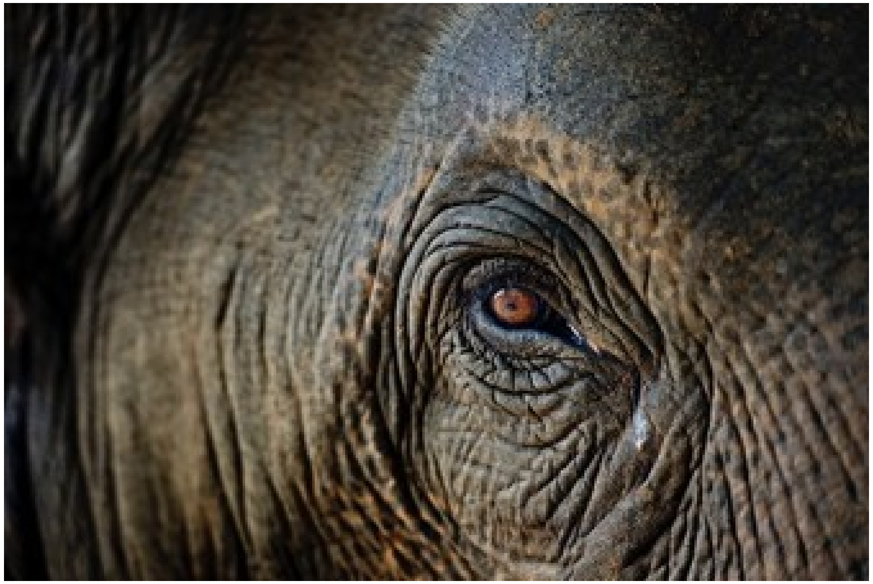
- Summarization as translation task:
- Translate a longer text into a shorter text
- Sequence-to-sequence models (RNNs/LSTMs/Transformers)

- Training data: pairs of longer and shorter texts. Examples:
  - For scientific documents: full article and abstract
  - For news texts: full article and snippet
  - For sentence compression: Parallel sentences derived from FLICKR30K data set (human generated image captions)

optional: see J&M chp 10. Machine Translation: https://web.stanford.edu/~jurafsky/slp3/10.pdf

SEQUENCE-TO-SEQUENCE LEARNING

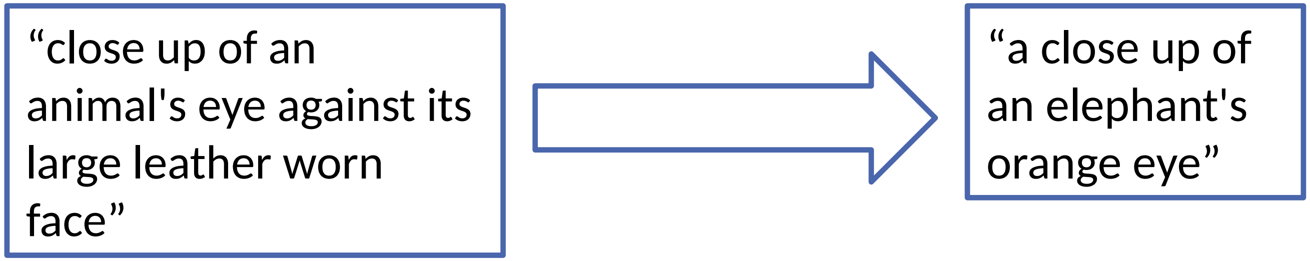
- Example: Parallel sentences derived from FLICKR30K data set (human generated image captions)



- “close up of an animal's eye against its large leather worn face”
- “elephant with a brown eye hyper focused in the camera”
- “a close up view of a pretty brown eye of an elephant”
- “a eye of an elephant that is looking forward”
- “a close up of an elephant's orange eye”

SEQUENCE-TO-SEQUENCE LEARNING

- Train a sequence-to-sequence model to learn the translation from a longer to a shorter sentence with the same meaning

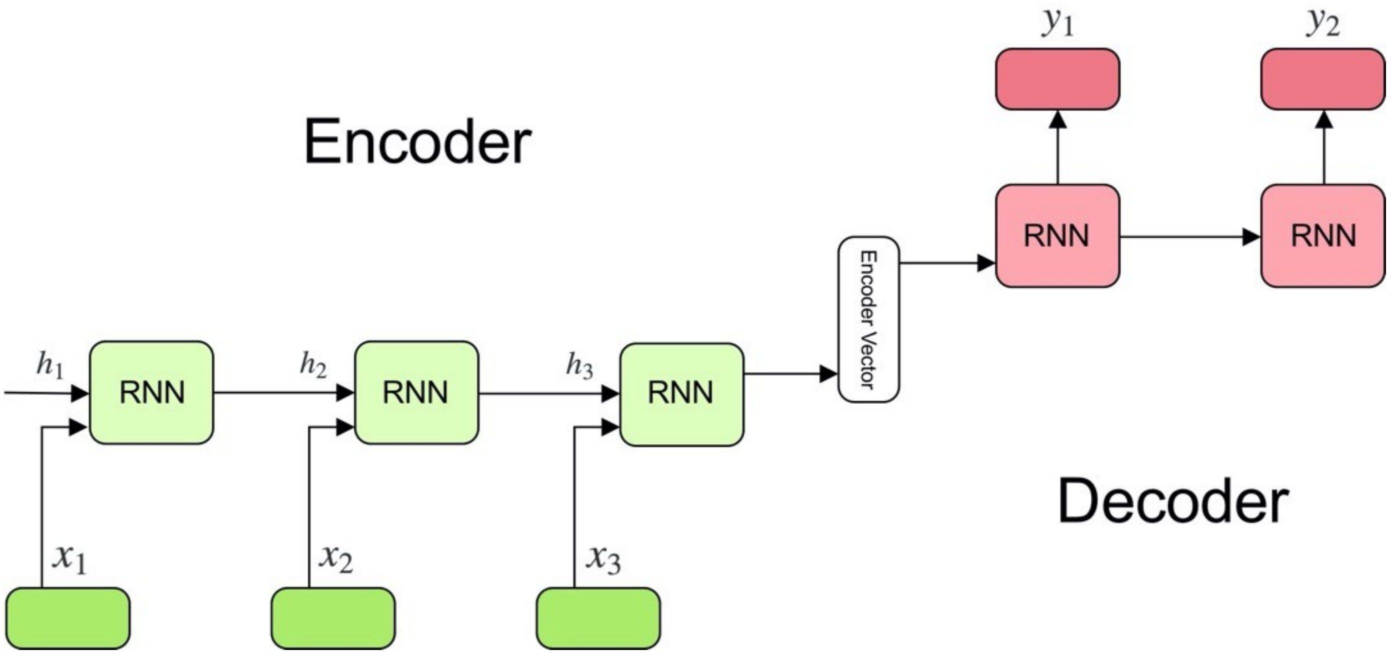


"close up of an animal's eye against its large leather worn face"

"a close up of an elephant's orange eye"

Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., & Xiang, B. (2016, August). Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* (pp. 280-290).

ENCODER-DECODER ARCHITECTURE



- The ‘encoder vector’ is the final hidden state from the encoder part of the model
  - This vector contains informative representations for all input elements to help the decoder make accurate predictions
- The model can map sequences of different lengths to each other

EXAMPLE RESULTS

Original

a man flipping in the air with a snowboard above a snow covered hill

many toilets without its upper top part near each other on a dark background

A table with three place settings with meat , vegetables and side dishes on it

A woman is leaning over a toilet , while her arms are inside a lawn and garden trash bag .

EXAMPLE RESULTS

Original	**Generated summary**
a man flipping in the air with a snowboard above a snow covered hill	A snowboarder is doing a trick on a snowy slope .
many toilets without its upper top part near each other on a dark background	A row of toilets sitting on a tiled floor .
A table with three place settings with meat , vegetables and side dishes on it	A table topped with plates of food and a glass of wine .
A woman is leaning over a toilet , while her arms are inside a lawn and garden trash bag .	A woman is cleaning a toilet in a park .

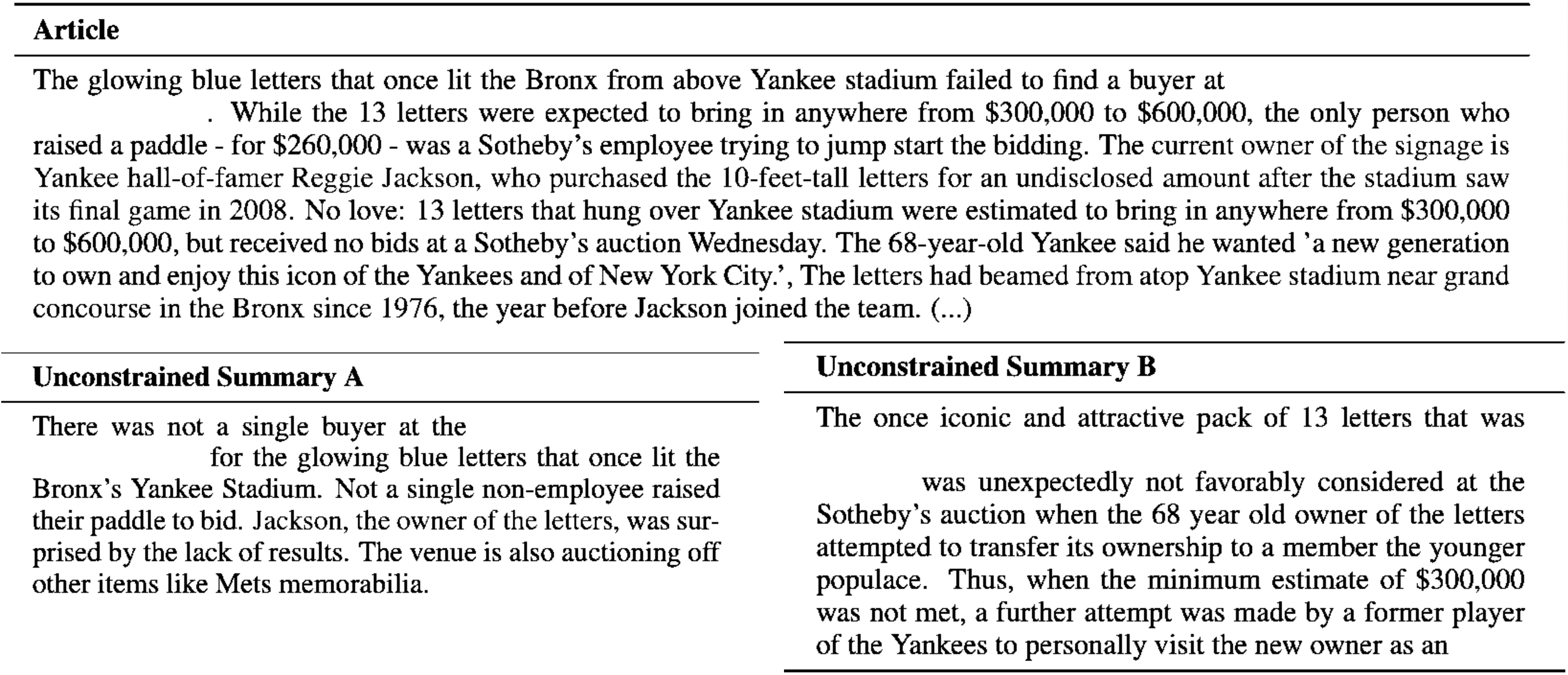
CHALLENGES OF SUMMARIZATION

RESEARCH CHALLENGES

- Task subjectivity/ambiguity
- Training data bias
- Evaluation

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

TASK SUBJECTIVITY/AMBIGUITY



Article

The glowing blue letters that once lit the Bronx from above Yankee stadium failed to find a buyer at . While the 13 letters were expected to bring in anywhere from $300,000 to $600,000, the only person who raised a paddle - for $260,000 - was a Sotheby's employee trying to jump start the bidding. The current owner of the signage is Yankee hall-of-famer Reggie Jackson, who purchased the 10-feet-tall letters for an undisclosed amount after the stadium saw its final game in 2008. No love: 13 letters that hung over Yankee stadium were estimated to bring in anywhere from $300,000 to $600,000, but received no bids at a Sotheby's auction Wednesday. The 68-year-old Yankee said he wanted 'a new generation to own and enjoy this icon of the Yankees and of New York City.', The letters had beamed from atop Yankee stadium near grand concourse in the Bronx since 1976, the year before Jackson joined the team. (...)

Unconstrained Summary A

There was not a single buyer at the for the glowing blue letters that once lit the Bronx's Yankee Stadium. Not a single non-employee raised their paddle to bid. Jackson, the owner of the letters, was surprised by the lack of results. The venue is also auctioning off other items like Mets memorabilia.

Unconstrained Summary B

The once iconic and attractive pack of 13 letters that was was unexpectedly not favorably considered at the Sotheby's auction when the 68 year old owner of the letters attempted to transfer its ownership to a member the younger populace. Thus, when the minimum estimate of $300,000 was not met, a further attempt was made by a former player of the Yankees to personally visit the new owner as an

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

TRAINING DATA BIAS

- Most used benchmark sets for training and evaluation summarization models are based on news data
  - E.g. the CNN/DailyMail (Nallapati et al., 2016) dataset

TRAINING DATA BIAS

- Most used benchmark sets for training and evaluation summarization models are based on news data
  - E.g. the CNN/DailyMail (Nallapati et al., 2016) dataset
- In newspaper articles, the most important information is in the first paragraph
  - This is used as a feature in summarization models
  - And this is why Lead-3 is such a strong baseline
- In other domains than newspaper data (e.g. books, legal documents, reviews), this characteristic does not always apply

EVALUATION

HOW WOULD YOU EVALUATE A SUMMARIZER?

- You have developed a summarization model
- How do you evaluate it?

EVALUATION OF SUMMARIZATION

- Compare to reference summaries
- Ask human judges

COMPARE TO REFERENCE SUMMARIES

- Extractive summarization (selecting sentences/responses) ▭ precision and recall against human reference data
  - set of true sentences in reference summary
  - set of selected sentences by system

COMPARE TO REFERENCE SUMMARIES

- Extractive summarization (selecting sentences/responses) □ precision and recall against human reference data
  - set of true sentences in reference summary
  - set of selected sentences by system

$$\text{precision} = \frac{|\text{selected} \cap \text{true}|}{|\text{selected}|} \qquad \text{recall} = \frac{|\text{selected} \cap \text{true}|}{|\text{true}|}$$

COMPARE TO REFERENCE SUMMARIES

- Abstractive summarization ▭ compute overlap with human reference summary
- ROUGE metrics
  - “Recall-Oriented Understudy for Gisting Evaluation”
  - Measures quality of a summary by comparison with reference summaries (literal)

COMPARE TO REFERENCE SUMMARIES

- ROUGE-N: the proportion of n-grams from the reference summaries that occur in the automatically created summary (‘recall-oriented’)

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Refs}\}} \sum_{\text{gram}_n \in S} \text{count}_{match}(\text{gram}_n)}{\sum_{S \in \{\text{Refs}\}} \sum_{\text{gram}_n \in S} \text{count}(\text{gram}_n)}$$

COMPARE TO REFERENCE SUMMARIES

- ROUGE-N: the proportion of n-grams from the reference summaries that occur in the automatically created summary (‘recall-oriented’)

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Refs}\}} \sum_{\text{gram}_n \in S} \text{count}_{match}(\text{gram}_n)}{\sum_{S \in \{\text{Refs}\}} \sum_{\text{gram}_n \in S} \text{count}(\text{gram}_n)}$$

  - n is the length of the n-gram, and $\text{Count}_{\text{match}}(\text{gram}_{\text{n}})$ is the number of n-grams co-occurring in a candidate summary and a set of reference summaries
  - the number of n-grams in the denominator increases as we add more references. This is intuitive and reasonable because there might exist multiple good summaries.
  - the numerator sums over all reference summaries. This effectively gives more weight to matching n-grams occurring in multiple references

COMPARE TO REFERENCE SUMMARIES

Toy example:

- Reference Summary: police killed the gunman
- System A: police kill the gunman
- System B: the gunman kill police

ROUGE-2:

COMPARE TO REFERENCE SUMMARIES

Toy example:

- Reference Summary: police killed the gunman
- System A: police kill the gunman
- System B: the gunman kill police

ROUGE-2

- A: $ police, police kill, kill the, the gunman, gunman $ → 3/5
- B: $ the, the gunman, gunman kill, kill police, police $ → 1/5
- → System A is better than system B according to ROUGE-2

EVALUATION OF SUMMARIZATION

Method:

- Compare to reference summaries
  - For extractive summarization ▭ Precision and Recall
  - For abstractive summarization ▭ ROUGE

- Ask human judges

ASK HUMAN JUDGES TO GRADE SUMMARIES

- Criteria to rate a summary:
  - Relevance: selection of important content from the source
  - Consistency: factual alignment between the summary and the source
  - Fluency: quality of individual sentences
  - Coherence: collective quality of all sentences
- Ask multiple judges per summary

CHALLENGES IN EVALUATION (ABSTRACTIVE)

- ROUGE often has weak correlation with human judgments
- But human judgments for relevance (importance) and fluency are strongly correlated to each other

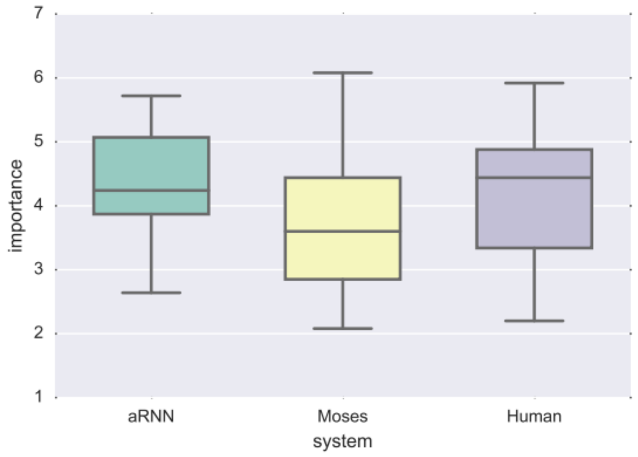


Figure 2: Importance scores given by human subjects to the two systems and human description.

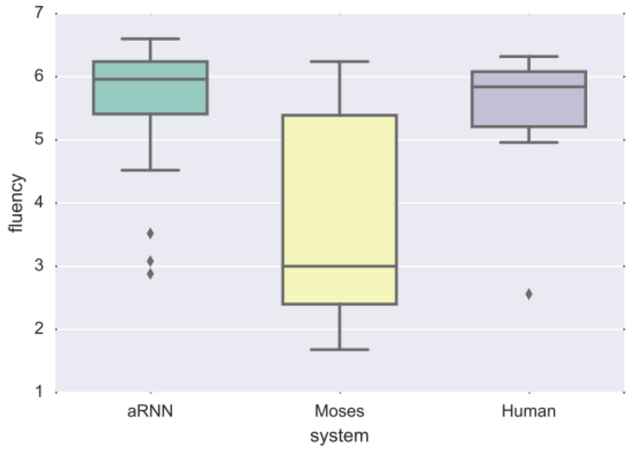


Figure 3: Fluency scores given by human subjects to the two systems and human description.

Wubben et al. (2016). Abstractive Compression of Captions with Attentive Recurrent Neural Networks. Proceedings of the 9th International Natural Language Generation conference (INLG)

CONCLUSIONS

SUZAN VERBERNE 2021

HOMEWORK

- Read:
  - Kryściński et al (2019). Neural text summarization: A critical evaluation
- Complete assignment 2: sequence labelling (CRF)

HOMEWORK

- Complete assignment 2: sequence labelling
- Send in via Brightspace before or on Monday November 15:
  - Submit your report as PDF and your python code as separate files.
  - Your report should not be longer than 3 pages

AFTER THIS LECTURE...

- You can explain the differences between extractive and abstractive summarization
- You can define the Lead-3 baseline and explain its success in benchmark data
- You can explain centrality-based summarization
- You can list a number of features that are relevant for sentence selection
- You can define the ROUGE metric to evaluate an automatic summarizer
- You can describe how you would evaluate an automatic summarizer
- You can explain the challenges in evaluation of summarization

ARGUMENT MINING

PRESENTED BY MICHIEL

Universiteit Leiden