Machine Learning for Natural Language Processing

Automatic Summarization

UNIVERSITY OF
GOTHENBURG

CHALMERS

Richard Johansson

richard.johansson@cse.gu.se

summarization, headline generation

Summary: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

Document: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
The plane's only two occupants were unharmed, it added.
The Diario de Noticias newspaper quoted an eyewitness who said the plane had been flying at a low altitude over the beach, although he did not realise anything was wrong until other beachgoers began running.
One young witness told Reuters news agency: "I was near the water when I saw the plane. I called my parents, the plane fell on the sand and ran over two people, fatally hurting them and another was injured, I think, but I'm not sure, people were running away."
"The plane is still there, but the ambulances arrived quickly. I think maybe the fuel ran out because I find it weird that it landed on the beach."
Other reports said the victims had been sunbathing when the plane made its emergency landing.
The Associated Press news agency said the girl who died had been with her parents, who were unhurt. The agency quoted witnesses from local television broadcasts.
Joao Quadros, who was on the beach, tweeted photos of the aftermath, saying the plane had passed by his son by a matter of metres. There had been no noise, he said.
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
One wing seemed to be misaligned in those photos.
The cause of the emergency landing remains unclear.

example by Narayan et al. (2019)

is this task well-defined?

how do we summarize this text?

India

From Wikipedia, the free encyclopedia

Coordinates: 21°N 78°E

This article is about the Republic of India. For other uses, see India (disambiguation).
"Bharat" redirects here. For other uses, see Bharat (disambiguation).

India (Hindi: *Bhārat*), officially the **Republic of India** (Hindi: *Bhārat Gaṇarājya*),[23] is a country in South Asia. It is the second-most populous country, the seventh-largest country by land area, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the southwest, and the Bay of Bengal on the southeast, it shares land borders with Pakistan to the west;[f] China, Nepal, and Bhutan to the north; and Bangladesh and Myanmar to the east. In the Indian Ocean, India is in the vicinity of Sri Lanka and the Maldives; its Andaman and Nicobar Islands share a maritime border with Thailand and Indonesia.

Modern humans arrived on the Indian subcontinent from Africa no later than 55,000 years ago.[24] Their long occupation, initially in varying forms of isolation as hunter-gatherers, has made the region highly diverse, second only to Africa in human genetic diversity.[25] Settled life emerged on the subcontinent in the western margins of the Indus river basin 9,000 years ago, evolving gradually into the Indus Valley Civilisation of the third millennium BCE.[26] By 1200 BCE, an archaic form of Sanskrit, an Indo-European language, had diffused into India from the northwest, unfolding as the language of the *Rigveda*, and recording the dawning of Hinduism in India.[27] The Dravidian languages of India were supplanted in the northern and western regions.[28] By 400 BCE, stratification and exclusion by caste had emerged within Hinduism,[29] and Buddhism and Jainism had arisen, proclaiming social orders unlinked to heredity.[30] Early political consolidations gave rise to the loose-knit Maurya and Gupta Empires based in the Ganges Basin.[31] Their collective era was suffused with wide-ranging creativity,[32] but also marked by the declining status

Republic of India
Bhārat Gaṇarājya
(see other local names)

Flag

सत्यमेव जयते
State emblem

Motto: "Satyameva Jayate" (Sanskrit)
"Truth Alone Triumphs"[1]

Anthem: "Jana Gana Mana"[2][3]
"Thou Art the Ruler of the Minds of All People"[4][2]

0:00 CC MENU

National song
"Vande Mataram" (Sanskrit)
"I Bow to Thee, Mother"[a][1][2]

alternatives to “simple” summarization

- **aspect-based** summarization
  - summarize with respect to a fixed number of topics
- **query-based** summarization
  - summarize with respect to what the user is searching for

main technical approaches in summarization

extractive: select some parts

Stockholm

From Wikipedia, the free encyclopedia

For other uses, see Stockholm (disambiguation).
"Sthlm" redirects here. For the Swedish TV series, see Sthlm (TV series).

Stockholm [a][8] is the capital and most populous urban area of Sweden as well as in Scandinavia. 1 million people live in the municipality,[9] approximately 1.6 million in the urban area,[5] and 2.4 million in the metropolitan area.[9] The city stretches across fourteen islands where Lake Mälaren flows into the Baltic Sea. Outside the city to the east, and along the coast, is the island chain of the Stockholm archipelago. The area has been settled since the Stone Age, in the 6th millennium BC, and was founded as a city in 1252 by Swedish statesman Birger Jarl. It is also the county seat of Stockholm County.

- dominated until recently

main technical approaches in summarization

extractive: select some parts

Stockholm

From Wikipedia, the free encyclopedia

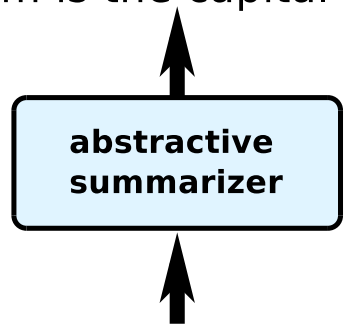
For other uses, see Stockholm (disambiguation).
"Sthlm" redirects here. For the Swedish TV series, see Sthlm (TV series).

Stockholm [a][8] is the capital and most populous urban area of Sweden as well as in Scandinavia. 1 million people live in the municipality,[9] approximately 1.6 million in the urban area,[5] and 2.4 million in the metropolitan area.[9] The city stretches across fourteen islands where Lake Mälaren flows into the Baltic Sea. Outside the city to the east, and along the coast, is the island chain of the Stockholm archipelago. The area has been settled since the Stone Age, in the 6th millennium BC, and was founded as a city in 1252 by Swedish statesman Birger Jarl. It is also the county seat of Stockholm County.

- dominated until recently

abstractive: read and rewrite

Stockholm is the capital of Sweden.



Stockholm

From Wikipedia, the free encyclopedia

For other uses, see Stockholm (disambiguation).
"Sthlm" redirects here. For the Swedish TV series, see Sthlm (TV series).

Stockholm [a][8] is the capital and most populous urban area of Sweden as well as in Scandinavia. 1 million people live in the municipality,[9] approximately 1.6 million in the urban area,[5] and 2.4 million in the metropolitan area.[9] The city stretches across fourteen islands where Lake Mälaren flows into the Baltic Sea. Outside the city to the east, and along the coast, is the island chain of the Stockholm archipelago. The area has been settled since the Stone Age, in the 6th millennium BC, and was founded as a city in 1252 by Swedish statesman Birger Jarl. It is also the county seat of Stockholm County.

- recently became viable
- mostly for short summaries

some well-known datasets for summarization

- datasets for supervised learning available recently
- some more suitable for extractive, some for abstractive
- examples:
  - **CNN/DailyMail** (Hermann et al., 2015): a few bullet points
  - **XSum** (Narayan et al., 2019): one-sentence summaries

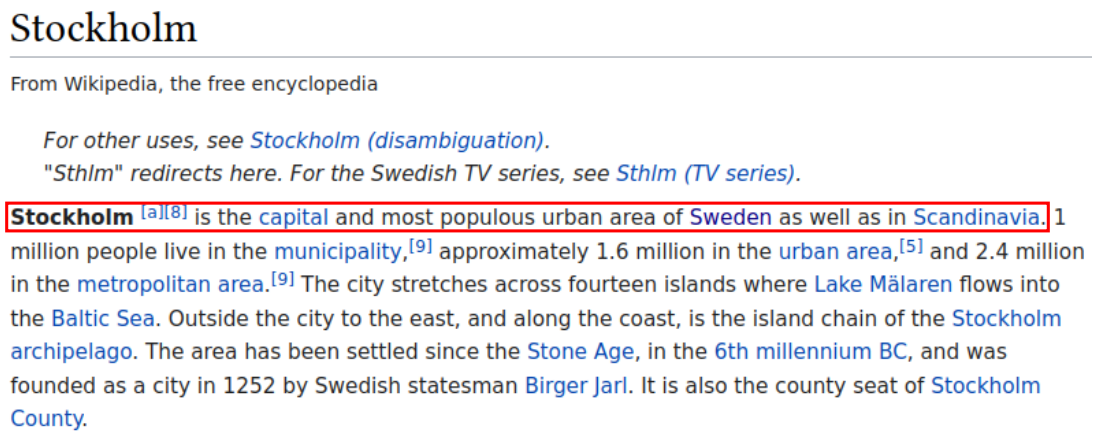
evaluation metrics and baselines for summarization

Model	R1	R2	RL
ORACLE	52.59	31.24	48.87
LEAD-3	40.42	17.62	36.67
Extractive			
SUMMARUNNER (Nallapati et al., 2017)	39.60	16.20	35.30
REFRESH (Narayan et al., 2018b)	40.00	18.20	36.60
LATENT (Zhang et al., 2018)	41.05	18.77	37.54
NEUSUM (Zhou et al., 2018)	41.59	19.01	37.98
SUMO (Liu et al., 2019)	41.00	18.40	37.20
TransformerEXT	40.90	18.02	37.17

(Liu and Lapata, 2019)

- **ROUGE** scores: word and bigram recall
- **ROUGE-L**: recall based on longest common subsequence
- the Lead-N baseline takes the first N sentences
- Oracle: the upper bound of an extractive system

extractive summarization: key ideas



Stockholm

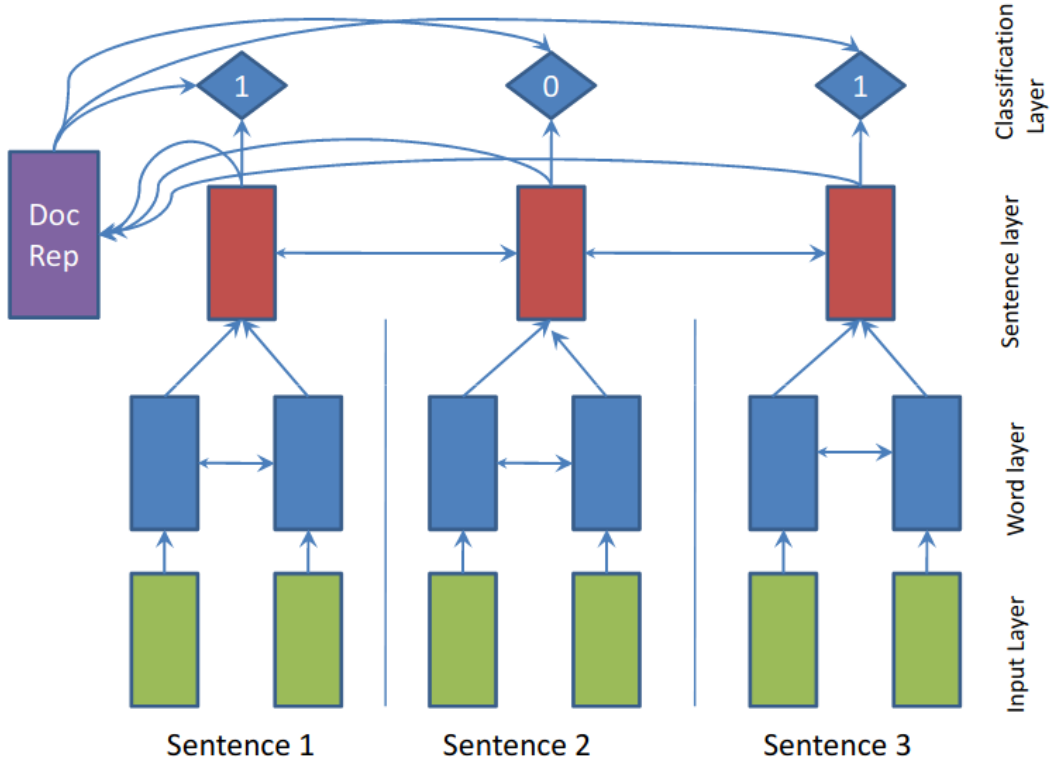
From Wikipedia, the free encyclopedia

For other uses, see Stockholm (disambiguation).
"Sthlm" redirects here. For the Swedish TV series, see Sthlm (TV series).

Stockholm [a][8] is the capital and most populous urban area of Sweden as well as in Scandinavia. 1 million people live in the municipality,[9] approximately 1.6 million in the urban area,[5] and 2.4 million in the metropolitan area.[9] The city stretches across fourteen islands where Lake Mälaren flows into the Baltic Sea. Outside the city to the east, and along the coast, is the island chain of the Stockholm archipelago. The area has been settled since the Stone Age, in the 6th millennium BC, and was founded as a city in 1252 by Swedish statesman Birger Jarl. It is also the county seat of Stockholm County.

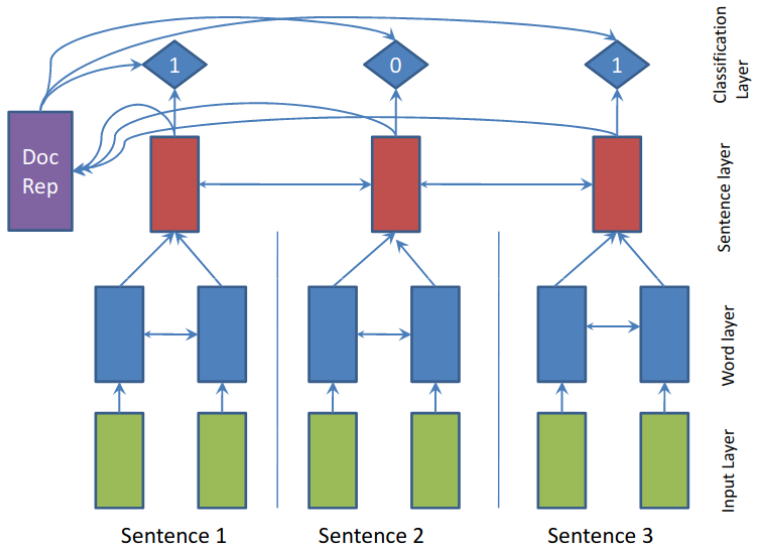
- select which parts of the text to include
  - typically sentences
- we want to select sentences that are **informative**, while minimizing **reduncancy**
- early systems e.g. Lin and Bilmes (2011) optimize these goals
- more recently, supervised learning is used

binary classifier for selecting sentences



- the SummaRuNNer system (Nallapati et al., 2017) treats sentence selection as a binary classification task

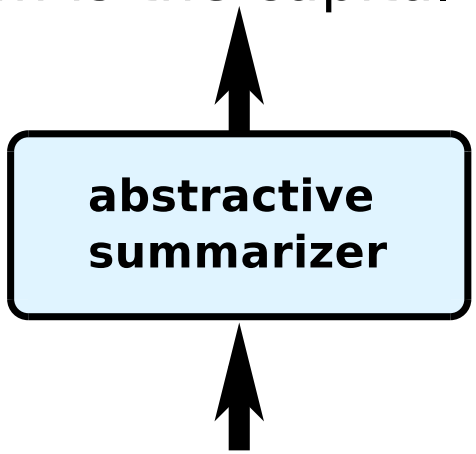
how do we train the sentence selector?



- what annotation do we use to train the sentence selector?
- Nallapati et al. (2017) use an **oracle**
  - try to include the sentences that maximize the ROUGE metric
  - in practice, a greedy algorithm is used
- **reinforcement learning** can be an alternative (Narayan et al., 2018): optimize ROUGE scores directly

abstractive summarization: key ideas

Stockholm is the capital of Sweden.



Stockholm

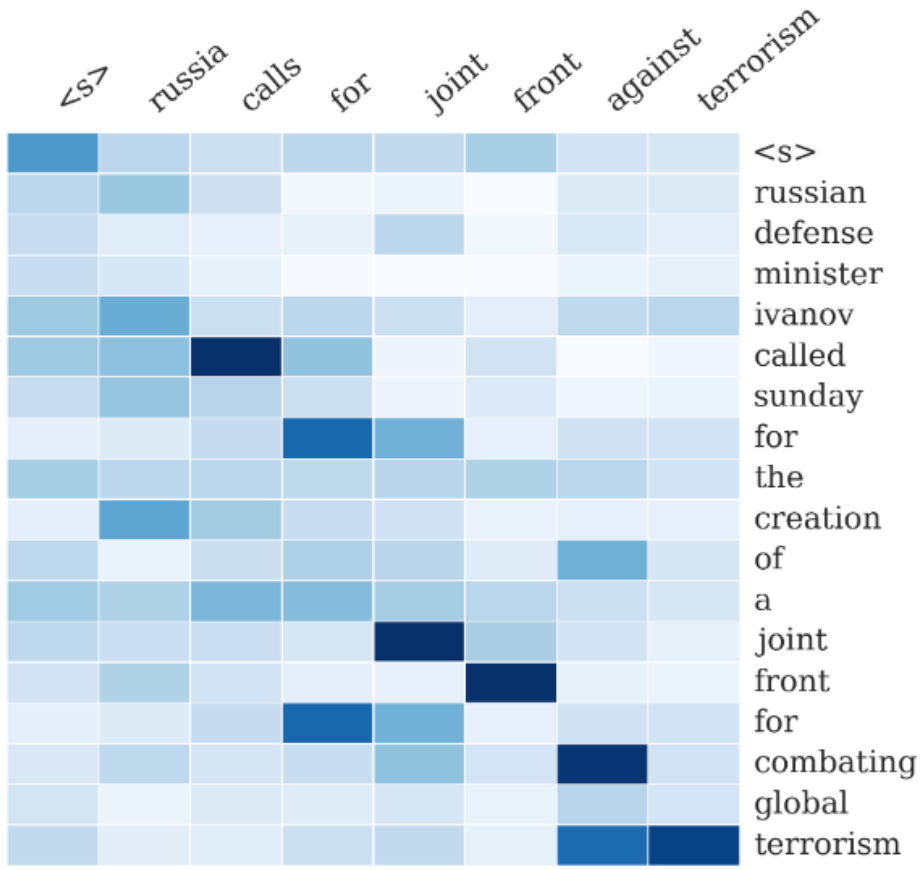
From Wikipedia, the free encyclopedia

For other uses, see Stockholm (disambiguation).
"Sthlm" redirects here. For the Swedish TV series, see Sthlm (TV series).

Stockholm [a][8] is the capital and most populous urban area of Sweden as well as in Scandinavia. 1 million people live in the municipality,[9] approximately 1.6 million in the urban area,[5] and 2.4 million in the metropolitan area.[9] The city stretches across fourteen islands where Lake Mälaren flows into the Baltic Sea. Outside the city to the east, and along the coast, is the island chain of the Stockholm archipelago. The area has been settled since the Stone Age, in the 6th millennium BC, and was founded as a city in 1252 by Swedish statesman Birger Jarl. It is also the county seat of Stockholm County.

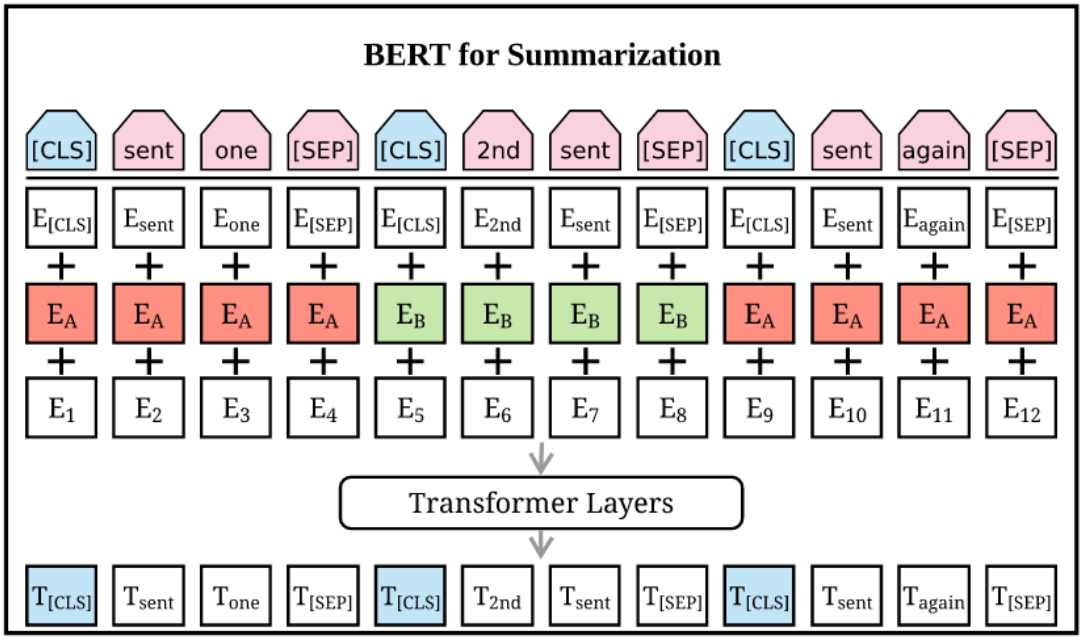
- more similar to conditional generation techniques we've seen
- text can be paraphrased and compressed

encoder/decoder solutions in abstractive summarization



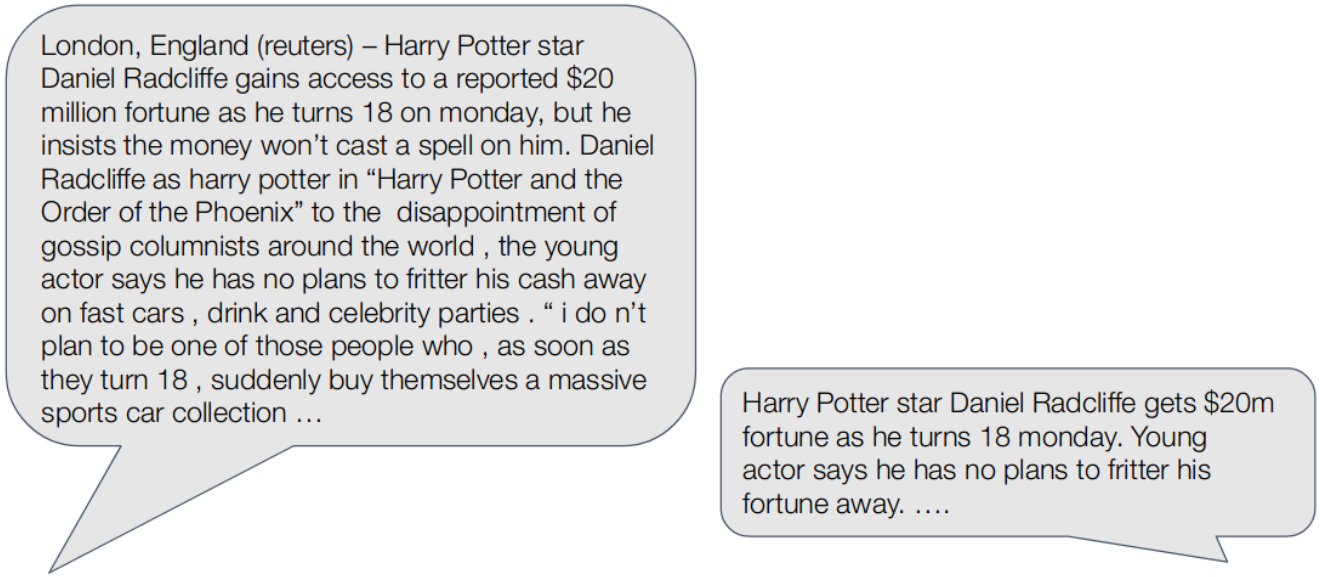
- Rush et al. (2015) introduced an encoder/decoder model for sentence summarization
- more recent work include copying mechanisms

examples of recent work in summarization



- Liu and Lapata (2019) adapt BERT for summarization
- key challenge: how to deal with long texts
- they use a **hierarchical** solution where a second Transformer is applied to the sequence of sentences
- results for extractive and abstractive summarization

summarization: take-home messages



[source]

- **technically** challenging: how to deal with long texts?
- **methodologically** challenging: what is a good summary?
- **extractive** methods dominated until recently, while **abstractive** methods have become more popular recently, for short summaries in particular
- **pre-trained models** are becoming more important

reading

- **Eisenstein** (Chapter 19) discusses summarization briefly
- the **BertSum** paper (Liu and Lapata, 2019) is good recent work including both extractive and abstractive techniques
- **BART** (Lewis et al., 2020) demonstrates the importance of pre-training tasks for generation applications including summarization and dialogue

references

K. M. Hermann, T. Kočiský, and E. Grefenstette et al. 2015. Teaching machines to read and comprehend. In *NIPS*.

M. Lewis, Y. Liu, and N. Goyal et al. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

H. Lin and J. Bilmes. 2011. A class of submodular functions for document summarization. In *ACL*.

Y. Liu and M. Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP*.

R. Nallapati, F. Zhai, and B. Zhou. 2017. SummaRuNNer: A RNN based sequence model for extractive summarization of documents. In *AAAI*.

S. Narayan, S. Cohen, and M. Lapata. 2019. What is this article about? Extreme summarization with topic-aware convolutional neural networks. *JAIR* 66.

S. Narayan, S. B. Cohen, and M. Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL*.

A. M. Rush, S. Chopra, and J. Weston. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP*.