ATNLP: Lecture 3
Summarisation

Shay Cohen

January 16, 2026

School of informatics

Summarisation

Georgina Rannard
Climate and science reporter
———

9 January 2025

What is probably the world's oldest ice, dating back 1.2m years ago, has been dug out from deep within Antarctica.

Working at temperatures of -35C, a team of scientists extracted a 2.8km-long cyclinder, or core, of ice - longer than eight Eiffel Towers end-to-end.

Suspended inside the ice are ancient air bubbles which scientists hope will help solve an enduring mystery about our planet's climate history.

The European scientists worked over four Antarctic summers, racing against seven nations to be first to reach the rock under the frozen continent.

Scientists dug up the ancient ice and stored it inside frozen caves on the ice sheet

Produce a **concise and coherent summary** of a longer document or multiple documents, to **capture essential information**, themes or points presented in the original document while **reducing its length**.

Summarisation uses

Useful for creating, for example:

- Outlines or abstracts for documents and articles (long tradition in NLP with news summarisation)

- Summaries for online conversations (Slack, e-mail)

- Action items for a meeting

- Simplifying documents by compressing them

Class exercise (history of tools)

The history of mankind using tools dates back millions of years and is a defining characteristic of human evolution. The earliest known tools, discovered in Africa, date to around 3.3 million years ago. These simple stone tools, known as Oldowan tools, were used by early hominins such as Australopithecus and later Homo habilis to cut meat and break bones.

Around 1.76 million years ago, early humans developed more advanced Acheulean tools, characterized by hand axes and cleavers. These tools, used by Homo erectus, represented a leap in cognitive abilities and craftsmanship. As human societies evolved, so did toolmaking. The discovery of fire allowed early humans to harden wooden tools and cook food, improving nutrition and survival rates.

The Neolithic Revolution (around 10,000 BCE) marked a major shift, as agriculture led to permanent settlements and new tools for farming, such as plows and sickles. Metallurgy soon followed, with the Bronze Age (c. 3300 BCE) and Iron Age (c. 1200 BCE) introducing stronger and more efficient tools and weapons.

The Industrial Revolution (18th–19th century) brought mechanization, mass production, and power-driven tools, transforming industries and daily life. In the 20th and 21st centuries, digital and robotic tools have revolutionized work, from computers to AI-driven machinery.

From stone tools to artificial intelligence, mankind's use of tools has shaped civilization, enabling survival, progress, and innovation across history.

Summary from an LLM

Humans have used tools for millions of years, starting with simple stone implements in the Oldowan and Acheulean periods. The development of agriculture, metallurgy, and mechanization led to increasingly advanced tools, from plows and swords to industrial machines. In the modern era, digital and robotic tools have revolutionized work and society, continuing humanity's long history of innovation.

Four components in summarisation
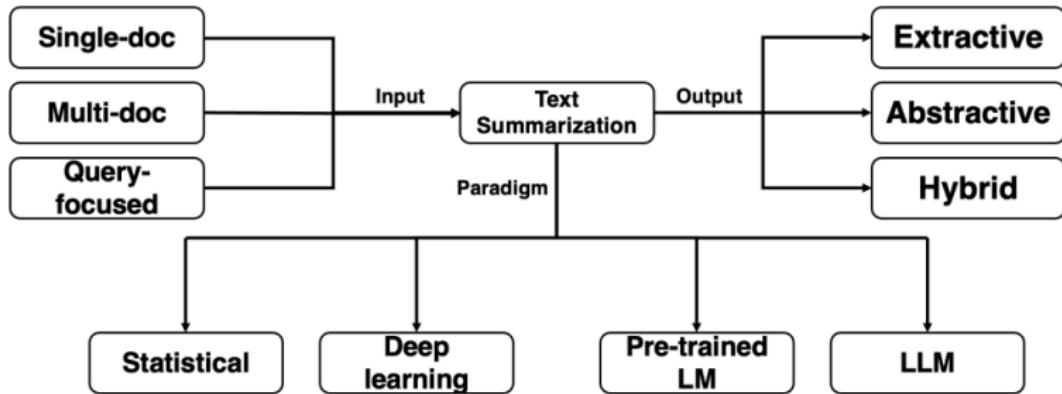
When someone comes to develop a new summarisation system, they need to decide on:

- The type of summarisation to apply
- Data to learn
- Model and learning
- Evaluation

Types of summarisation models

- Multi-document versus single document
- Extractive versus abstractive
- Generic (unconditioned) versus query-focused (conditioned) or controllable
- Supervised versus unsupervised
- Multi-modal versus single modality

Figure adapted from Zhang et al. (2024)

Multi-document versus single document

- Single document summarisation use a single source, for example, a news article

- Multi document summarisation uses multiple sources or documents for the summary (all sources are related to a theme or topic)
  Food for thought: Which is easier and why?

Multi-document versus single document

- Single document summarisation use a single source, for example, a news article

- Multi document summarisation uses multiple sources or documents for the summary (all sources are related to a theme or topic)
  Food for thought: Which is easier and why? Issues with MDS:
    - Multiple documents: information overlap, how to ensure we do not repeat information?
    - Multiple documents: how do we ensure we keep the information consistent? Views/opinions/different times?
    - More to summarise, longer context! Always an issue with LLMs

Query-focused summarisation and controllable summarisation

Both use an additional input or parameter to summarise
Query-focused:

- Summarise a document based on a specific query

Controllable:

- Control parameters such as length, aspect, sentiment

Can be formulated as part of the prompt, or as additional information we condition on, or part of the encoding in an encoder-decoder summariser

Extractive versus abstractive summarisation

- Extractive - use a subset of the sentences from the document as the candidate summary
- Abstractive - synthesise a summary which is not tied to the exact wording in the article

Food for thought: Why do we need extractive summarisation?

Extractive versus abstractive summarisation

- Extractive - use a subset of the sentences from the document as the candidate summary
- Abstractive - synthesise a summary which is not tied to the exact wording in the article

Food for thought: Why do we need extractive summarisation?
"High precision"

First Shot at Modelling Extractive Summarisation

We have a document $d = (s_1, \ldots, s_n)$ with n sentences

Use classification: $y_i = 1$ if s_i should be in the (extractive) summary, and 0 if not

Quite simple and indeed serves as the basic model for many summarisers... but presents some challenges

Summarisation as Binary Labelling

Two main questions we will discuss:

- Where do we get the labels to train the model?

- What objective do we choose to train with?

Extractive Summarisation - the Naive Way

- Where do we get the labels to train the model?

Compute a score (ROUGE) between all sentences in an article and the summary, and choose the highest scoring ones as 1 labels

Extractive Summarisation - the Naive Way

- Where do we get the labels to train the model?

Compute a score (ROUGE) between all sentences in an article and the summary, and choose the highest scoring ones as 1 labels

- What objective do we choose to train with?

Maximise the log-likelihood of the data with a binary classification objective

Abstractive summarisation

Treat the problem of summarisation as a text generation
problem (rather than extracting sentences as before)
See et al. (2017) use a pointer-generator network as follows:

- The base is a sequence-to-sequence model

Abstractive summarisation

Treat the problem of summarisation as a text generation problem (rather than extracting sentences as before)
See et al. (2017) use a pointer-generator network as follows:

- The base is a sequence-to-sequence model
- The probability of generating a word is

$$P(w) = p_{gen}P_{vocab}(w) + (1 - p_{gen}) \sum_{i:\ w_i = w} a_i^t$$

  where t is the current time step, and p_{gen} indicates a mixture (or "switch") between either generating a word from the base decoder or copying a word based on the attention weights (a)

Abstractive summarisation

Treat the problem of summarisation as a text generation problem (rather than extracting sentences as before)
See et al. (2017) use a pointer-generator network as follows:

- The base is a sequence-to-sequence model
- The probability of generating a word is

$$P(w) = p_{gen}P_{vocab}(w) + (1 - p_{gen}) \sum_{i \,:\, w_i = w} a_i^t$$

  where t is the current time step, and p_{gen} indicates a mixture (or "switch") between either generating a word from the base decoder or copying a word based on the attention weights (a)

- Coverage mechanism: to avoid repetition (a big problem with older seq-to-seq models), a coverage vector is fed into the decoder: $c^t = \sum_{t'=0}^{t-1} a^{t'}$

CNN/DailyMail dataset (Hermann et al., 2015)

- Training data: pairs of news articles ($\approx$800 words on average) and summaries (aka story highlights), usually 3 or 4 sentences long ($\approx$56 words on average)

- CNN: 100k pairs; Daily Mail: 200k pairs

- Highlights were sourced from journalists in compressed, "telegraphic", manner

- The highlights need not to form a coherent summary — each highlight is relatively stand-alone, with little co-referencing

- Available at https://github.com/abisee/cnn-dailymail

Example

Most blacks say MLK's vision fulfilled, poll finds WASHINGTON (CNN) – More than two-thirds of African-Americans believe Martin Luther King Jr.'s vision for race relations has been fulfilled, a CNN poll found – a figure up sharply from a survey in early 2008.

The CNN-Opinion Research Corp. survey was released Monday, a federal holiday honoring the slain civil rights leader and a day before Barack Obama is to be sworn in as the first black U.S. president.

The poll found 69 percent of blacks said King's vision has been fulfilled in the more than 45 years since his 1963 'I have a dream' speech – roughly double the 34 percent who agreed with that assessment in a similar poll taken last March.

But whites remain less optimistic, the survey found. 'Whites don't feel the same way – a majority of them say that the country has not yet fulfilled King's vision,' CNN polling director Keating Holland said. However, the number of whites saying the dream has been fulfilled has also gone up since March, from 35 percent to 46 percent.

Highlights: [Paraphrased]   [Verbatim]

- **69% of blacks** *polled* say **Martin Luther King Jr's vision** *realised*
- *Slim majority of white people say King's vision is not fulfilled*
- King gave *his* **"I have a dream" speech** *in 1963*

Some results

Models	ROUGE		
	1	2	L
seq-to-seq+attn	31.33	11.81	28.83
pointer-generator	36.44	15.66	33.42
pointer-generator + coverage	39.53	17.28	36.38
lead-3 baseline	40.34	17.70	36.57
BERTSUMABS	41.72	19.39	38.76

What is LEAD-3?

Some results

Models	ROUGE		
	1	2	L
seq-to-seq+attn	31.33	11.81	28.83
pointer-generator	36.44	15.66	33.42
pointer-generator + coverage	39.53	17.28	36.38
lead-3 baseline	40.34	17.70	36.57
BERTSUMABS	41.72	19.39	38.76

What is LEAD-3? Taking the first three sentences in the
document as the candidate summary

Human Summaries for CNN/DailyMail

Article: Queen Victoria spent her holidays in Osborne House on the Isle of Wight. ... She would travel to Portsmouth by train and then by ferry to Ryde. From Ryde there was a railway line that passed not far from Osborne House but the nearest station was at Wootton, more than two miles from the property. So, in 1875, a station was built at Whippingham, the closest point on the line to Osborne House – just to serve the Royal residence. ... The building is now a five-bedroom family home, currently on the market for £625,000, while the track has become a cycle path. ...

Human-written Summary:

- Queen Victoria's holiday residence was Osborne House on the Isle of Wight
- But her journeys there involved train and ferry ride and then another train ride to a station more than two miles from the property
- In 1875, a station was built at Whippingham just to serve Royal residence
- Building is now a five-bedroom home, currently on the market for £625,000

Automatic Summaries for CNN/DailyMail

Article: Queen Victoria spent her holidays in Osborne House on the Isle of Wight. ... She would travel to Portsmouth by train and then by ferry to Ryde. From Ryde there was a railway line that passed not far from Osborne House but the nearest station was at Wootton, more than two miles from the property. So, in 1875, a station was built at Whippingham, the closest point on the line to Osborne House – just to serve the Royal residence. ... The building is now a five-bedroom family home, currently on the market for £625,000, while the track has become a cycle path. ...

Automatically-written Summary (Pointer-Generator, See et al., 2017):

- Queen Victoria spent her holidays in Osborne House on the Isle of Wight.

- She would travel to Portsmouth by train and then by ferry to ryde.

- Building is now a five-bedroom family home, currently on the market for £625,000.

The XSum Dataset

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.

[*6 sentences with 139 words are abbreviated from here.*]

Other reports said the victims had been sunbathing when the plane made its emergency landing.

[*Another 4 sentences with 67 words are abbreviated from here.*]

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.

[*Last 2 sentences with 19 words are abbreviated.*]

Source: https://www.bbc.co.uk/news/world-europe-40806376

Datasets

Many datasets are available for training and testing
summarisation models, including:

XWikis (Perez-Beltrachini and Lapata, 2021), cross-lingual document-summary pairs
in 4 languages derived from Wikipedia

MLSUM (Scialom et al., 2020), document-summary pairs, 5 languages news outlets

XSum (Narayan et al., 2018), 227K BBC articles with single-sentence summaries

NewsRoom (Grusky et al., 2018), 1.3M article-summary pairs written by editors

arXiv, **PubMed** (Cohan et al., 2018), abstract and paper body (113K and 215K)

BigPatent (Sharma et al., 2019) 1.3 million US patents with human summaries

WikiHow (Koupaee and Wang, 2018) 200K instructions with single-sentence
summaries

Reddit TIFU (Kim et al., 2019) 129K stories with descriptive summaries

Evaluation

How to evaluate summaries? A few dimensions for evaluation:

- Human versus automatic
- Word overlap versus soft overlap
- Reference-free versus reference-based

Each of these dimensions has more complexities

ROUGE (Lin, 2004)

Recall-Oriented Understudy for Gisting Evaluation:

$$ROUGE\text{-}N = \frac{\sum_{S \in ref} \sum_{gram_n \in S} countmatch(gram_n)}{\sum_{S \in ref} \sum_{gram_n \in S} count(gram_n)}$$

Food for thought: Why are we not considering precision?

ROUGE (Lin, 2004)

Recall-Oriented Understudy for Gisting Evaluation:

$$ROUGE\text{-}N = \frac{\sum_{S \in ref} \sum_{gram_n \in S} countmatch(gram_n)}{\sum_{S \in ref} \sum_{gram_n \in S} count(gram_n)}$$

Food for thought: Why are we not considering precision? Recall is more important to decide whether the information in the candidate summary captures the information in the reference summary. However, there is a version of ROUGE which calculates F_1

There is also ROUGE-L, which finds the longest common string between the reference and the candidate summary

BERTScore (Zhang et al., 2019)

Rather than relying on "hard" n-grams, use BERT to softly score similarity

Let x be a sequence of symbols for the reference, and $\hat{x}$ the candidate.

Then, the BERT scores are defined similarly to recall, precision, F1:

$R = \frac{1}{|x|} \sum_{i=1}^{|x|} \max_j x_i^\top \hat{x}_j \quad P = \frac{1}{|\hat{x}|} \sum_{j=1}^{|\hat{x}|} \max_i x_i^\top \hat{x}_j$

$F = \frac{2PR}{P + R}$

The maximum $\max_i x_i^\top \hat{x}_j$ tell us "is there any word in x that matches the j word in $\hat{x}$? - precision
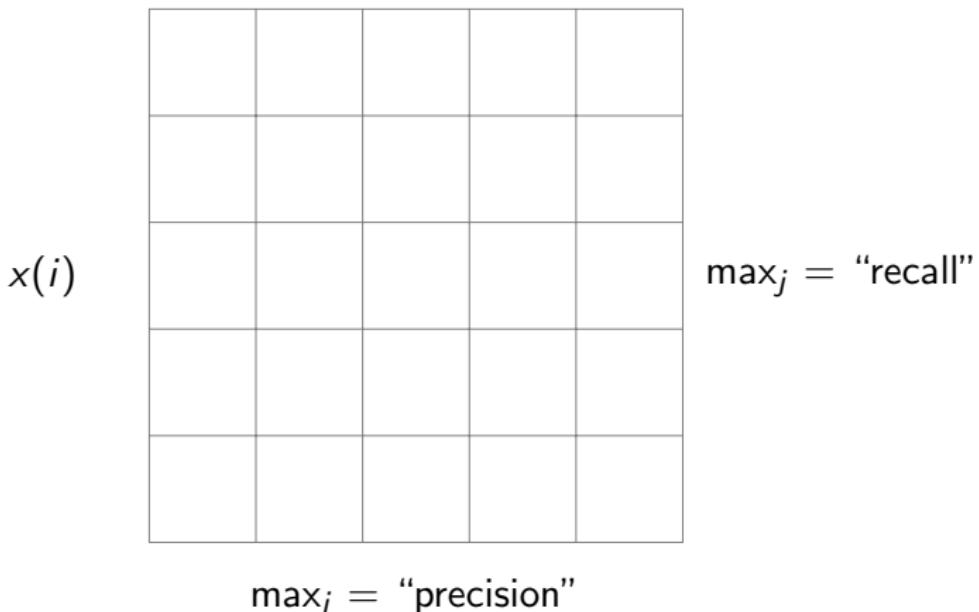
The maximum $\max_j x_i^\top \hat{x}_j$ tell us "is there any word in $\hat{x}$ that matches the i word in x? - recall

Illustration of BERTScore

Every cell in the matrix corresponds to dot product of the form $x_i^\top \hat{x}_j$

$\hat{x}(j)$

$x(i)$

$\max_j$ = "recall"

$\max_i$ = "precision"

A Valid Criticism

We are optimising the wrong metric! Especially for "informativeness"

> **Gold summary:** The book was long and interesting.
> **Summary 1:** Interesting long book.
> **Summary 2:** The book was long and boring.

The ROUGE score for summary 1 might be lower than summary 2.

Similar case to BLEU in Machine Translation

Human evaluations are needed

Evaluation of summarisation

These metrics, even BERTScore, are quite lexical in their nature
Food for thought: Can we do better?

Evaluation of summarisation

These metrics, even BERTScore, are quite lexical in their nature
Food for thought: Can we do better?

- Rely on human evaluations
- Several approaches:
    - Ranking two summaries and asking which one is better (for example, reference and candidate)
    - Generate questions based on the reference summary, ask humans to answer them based on the candidate summary
    - Dimensions inspected: relevance, fluency, coherence, faithfulness (factual correctness), and conciseness.
- We can also ask a machine to answer such questions

The evaluation of summarisation by itself is a complex problem that spun a lot of work!

LLMs as a Judge

- Ask an LLM to judge a summary
- Same style as human evaluation (pairwise ranking, absolute scores), only now with LLMs
- The problem of "evaluation" is considered easier than "summarisation" itself, so we can hope LLMs do it reasonably well

The same issues as usual: sensitivity to the prompt, over-reliance on surface features, lack of consistency

Summarisation with Long Text

Long text still poses an issue for summarisation (summarising books? summarising long scientific articles?)

Example approach (hierarchical-style summarisation):

- Break the document into chunks
- Summarise each separately
- Now treat the concatenation as a new document, and repeat

Summarisation with Long Text

- LLMs give impressive summaries for short documents
- For long documents, they may disproportionately focus on content at the start and end.
- Even if the middle part contains important information, it may be underrepresented in the summary ("lost-in-the-middle" problem in LLMs).

GPT-3 News Summarisation (Goyal et al., 2022)

- Humans prefer GPT-3 summaries over other automatic summaries
- It is difficult to rely on automatic metrics to evaluate GPT-3 summaries: automatic metrics show the other systems are "better"

Dataset	BRIO		T0		GPT3	
	Best ↑	Worst ↓	Best ↑	Worst ↓	Best ↑	Worst ↓
CNN	36	24	8	67	58	9
BBC	20	56	30	29	57	15

Table 3: Percentage of times a summarization system is selected as the best or worst according to majority vote (may be tied). Human annotators have a clear preference for GPT3-D2 for both CNN and BBC style summaries.

Summarisation These Days

We portrayed more of an "evolution" of the area of summarisation over the years, as a dedicated task. Now:

- Commericial LLMs are pretty good at summarising medium-length texts (papers, reports, articles) - by prompting! Not specialised LLMs (probably)

- Still suffer from the problem of hallucination

- Can have "query-focused" summarisation – summarise a specific part of the text or respond to a question

- (Most likely) relying on high-quality summarisation data to finetune or adapt the pretrained models

- The focus of the summarisation community is now on two aspects: (a) multilingual data; (b) better, more fine-grained evaluation

Summary (by an LLM - you be the judge)

- Summarization condenses long text into concise, coherent summaries for applications like news, meetings, and emails.
- Types: Single vs. multi-document (handling redundancy & inconsistencies) and extractive vs. abstractive (sentence selection vs. text generation).
- Extractive summarization is modeled as a classification problem, selecting key sentences based on metrics like ROUGE.
- Abstractive summarization uses deep learning models like pointer-generator networks to generate novel text while avoiding repetition.
- Evaluation methods include ROUGE (word overlap), BERTScore (semantic similarity), and human evaluation (ranking and faithfulness).
- LLMs as judges rank and score summaries, mimicking human assessment but facing issues like prompt sensitivity.
- Summarizing long texts is challenging; hierarchical methods break content into smaller sections for processing.
- GPT-3 excels in summarization, preferred by humans over traditional models, though automatic metrics struggle to reflect its quality.

References

- A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models, Zhang et al. (2024)
- Teaching machines to read and comprehend, Hermann et al. (2015)
- Get to the point: Summarization with pointer-generator networks, See et al. (2017)
- Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, Narayan et al. (2018)
- ROUGE: A package for automatic evaluation of summaries, Lin (2004)
- BERTScore: Evaluating text generation with BERT, Zhang et al. (2019)
- News summarization and evaluation in the era of GPT-3, Goyal et al. (2022)