



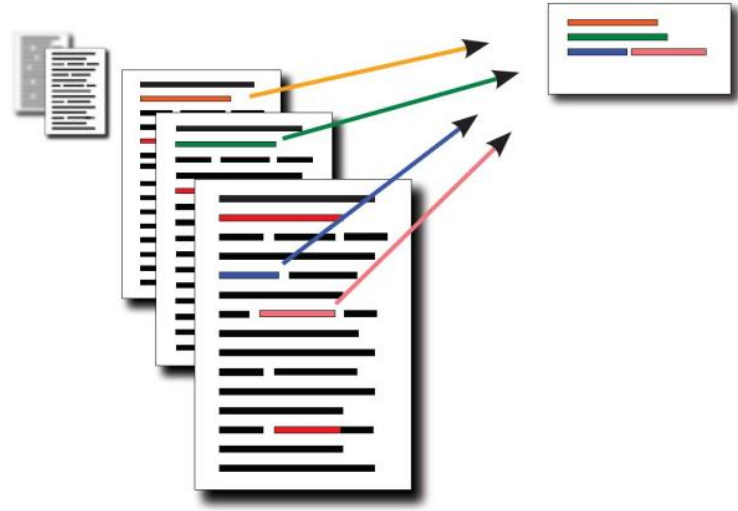
Language Understanding

11 - Summarization

Hossein Zeinali

Text Summarization

- **Goal:** produce an abridged version of a text that contains information that is important or relevant to a user.
- **Summarization Applications:**
 - **Outlines or abstracts** of any document, article, etc.
 - **Summaries** of email threads
 - **Action items** from a meeting
 - **Simplifying** text by compressing sentences



What to Summarize?

- **Single-document summarization**

- Given a single document, produce
 - Abstract
 - Outline
 - Headline

- **Multiple-document summarization**

- Given a group of documents, produce a gist of the content:
 - A series of news stories on the same event
 - A set of web pages about some topic or question



Single-document Summarization

Document

Cambodian leader Hun Sen on Friday rejected opposition parties ' demands for talks outside the country , accusing them of trying to `` internationalize " the political crisis .

Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen 's party to form a new government failed .

Opposition leaders Prince Norodom Ranariddh and Sam Rainsy , citing Hun Sen 's threats to arrest opposition figures after two alleged attempts on his life , said they could not negotiate freely in Cambodia and called for talks at Sihanouk 's residence in Beijing .Hun Sen , however , rejected that .``

I would like to make it clear that all meetings related to Cambodian affairs must be conducted in the Kingdom of Cambodia , " Hun Sen told reporters after a Cabinet meeting on Friday .`` No-one should internationalize Cambodian affairs .

It is detrimental to the sovereignty of Cambodia , " he said .Hun Sen 's Cambodian People 's Party won 64 of the 122 parliamentary seats in July 's elections , short of the two-thirds majority needed to form a government on its own .Ranariddh and Sam Rainsy have charged that Hun Sen 's victory in the elections was achieved through widespread fraud .They have demanded a thorough investigation into their election complaints as a precondition for their cooperation in getting the national assembly moving and a new government formed

Summary

Cambodian government rejects opposition's call for talks abroad



Multiple-document Summarization

Documents

Fingerprints and photos of two men who boarded the doomed Malaysia Airlines passenger jet are being sent to U.S. authorities so they can be compared against records of known terrorists and criminals. The cause of the plane's disappearance has baffled investigators and they have not said that they believed that terrorism was involved, but they are also not ruling anything out. The investigation into the disappearance of the jetliner with 239 passengers and crew has centered so far around the fact that two passengers used passports stolen in Thailand from an Austrian and an Italian. The plane which left Kuala Lumpur, Malaysia, was headed for Beijing. Three of the passengers, one adult and two children, were American.

(CNN) -- A delegation of painters and calligraphers, a group of Buddhists returning from a religious gathering in Kuala Lumpur, a three-generation family, nine senior travelers and five toddlers. Most of the 227 passengers on board missing Malaysia Airlines Flight 370 were Chinese, according to the airline's flight manifest. The 12 missing crew members on the flight that disappeared early Saturday were Malaysian. The airline's list showed the passengers hailed from 14 countries, but later it was learned that two people named on the manifest -- an Austrian and an Italian -- whose passports had been stolen were not aboard the plane. The plane was carrying five children under 5 years old, the airline said.

:

Vietnamese aircraft spotted what they suspected was one of the doors belonging to the ill-fated Malaysia Airlines Flight MH370 on Sunday, as troubling questions emerged about how two passengers managed to board the Boeing 777 using stolen passports. The discovery comes as officials consider the possibility that the plane disintegrated mid-flight, a senior source told Reuters. The state-run Thanh Nien newspaper cited Lt. Gen. Vo Van Tuan, deputy chief of staff of Vietnam's army, as saying searchers in a low-flying plane had spotted an object suspected of being a door from the missing jet. It was found in waters about 56 miles south of Tho Chu island, in the same area where oil slicks were spotted Saturday.

Summary

Flight MH370, carrying 239 people vanished over the South China Sea in less than an hour after taking off from Kuala Lumpur, with two passengers boarded the Boeing 777 using stolen passports. Possible reasons could be an abrupt breakup of the plane or an act of terrorism. The government was determining the "true identities" of the passengers who used the stolen passports. Investigators were trying to determine the path of the plane by analysing civilian and military radar data while ships and aircraft from seven countries scouring the seas around Malaysia and south of Vietnam.

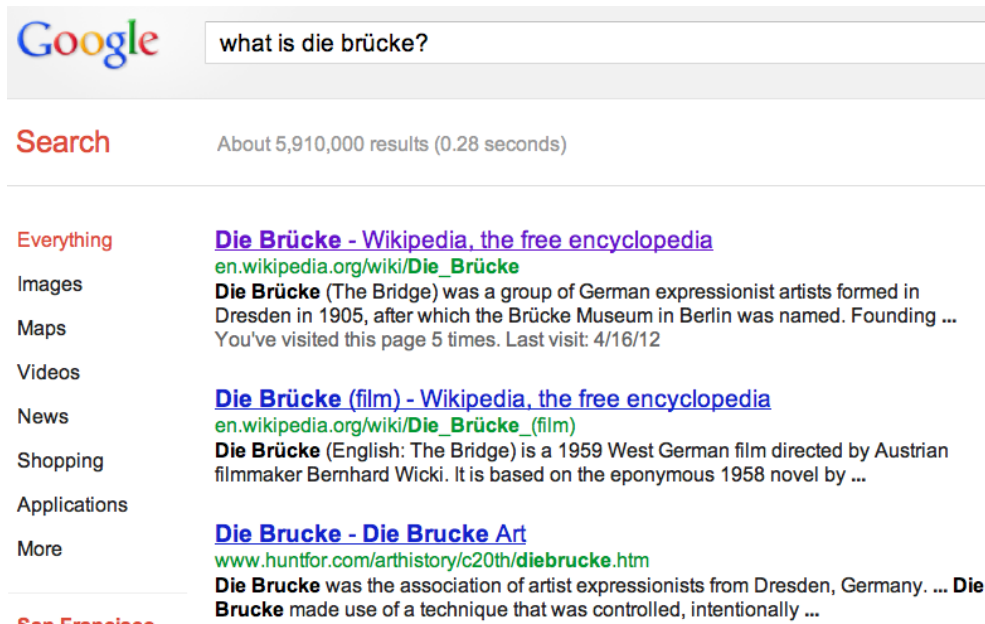
Multiple-document Summarization for the topic of "Malaysia Airlines Disappearance"

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Summarization for QA: Snippets

- Create **snippets** summarizing a web page for a query
 - Google: 156 characters (about 26 words) plus title and link



Google search results for the query "what is die brücke?". The search bar shows the query, and the search button is labeled "Search". The results show "About 5,910,000 results (0.28 seconds)". The first result is titled "Die Brücke - Wikipedia, the free encyclopedia" with the URL "en.wikipedia.org/wiki/Die_Brücke". The snippet describes the group of German expressionist artists formed in Dresden in 1905. The second result is titled "Die Brücke (film) - Wikipedia, the free encyclopedia" with the URL "en.wikipedia.org/wiki/Die_Brücke_(film)". The snippet describes the 1959 West German film directed by Austrian filmmaker Bernhard Wicki. The third result is titled "Die Brücke - Die Brücke Art" with the URL "www.huntfor.com/arthistory/c20th/diebrucke.htm". The snippet describes the association of artist expressionists from Dresden, Germany.

Google

what is die brücke?

Search About 5,910,000 results (0.28 seconds)

Everything [Die Brücke - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Die_Brücke
Die Brücke (The Bridge) was a group of German expressionist artists formed in Dresden in 1905, after which the Brücke Museum in Berlin was named. Founding ...
You've visited this page 5 times. Last visit: 4/16/12

Images

Maps

Videos

News [Die Brücke \(film\) - Wikipedia, the free encyclopedia](#)
[en.wikipedia.org/wiki/Die_Brücke_\(film\)](https://en.wikipedia.org/wiki/Die_Brücke_(film))
Die Brücke (English: The Bridge) is a 1959 West German film directed by Austrian filmmaker Bernhard Wicki. It is based on the eponymous 1958 novel by ...

Shopping

Applications

More [Die Brücke - Die Brücke Art](#)
www.huntfor.com/arthistory/c20th/diebrucke.htm
Die Brücke was the association of artist expressionists from Dresden, Germany. ... Die Brücke made use of a technique that was controlled, intentionally ...



Summarization for QA: Multiple Documents

- Create **answers** to complex questions summarizing multiple documents.
 - Instead of giving a snippet for each document
 - Create a cohesive answer that combines information from each document



Ways of Doing Summarization

- **Extractive summarization:**

- **Select parts** (typically sentences) of the original text to form a summary
 - E.g. Snippets used by Google
- Easier
- Restrictive (no paraphrasing)
- The great majority of past works

- **Abstractive summarization:**

- **Generate new** text using natural language generation techniques
- More difficult
- More flexible (human-like)
- Recent Seq2Seq models



Kinds of Summarization

- **Generic summarization:**

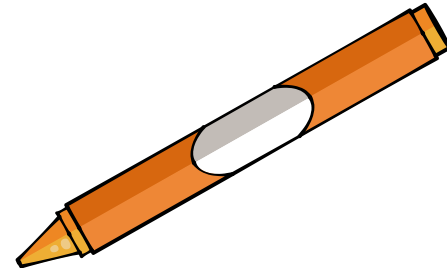
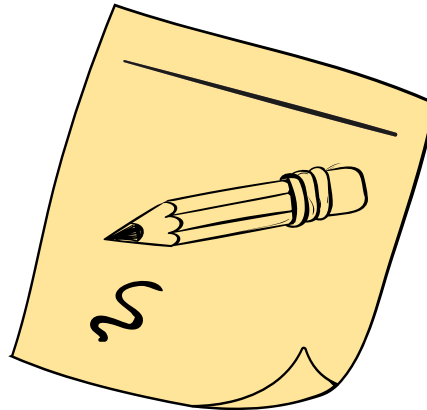
- Summarize the content of a document

- **Query-focused summarization:**

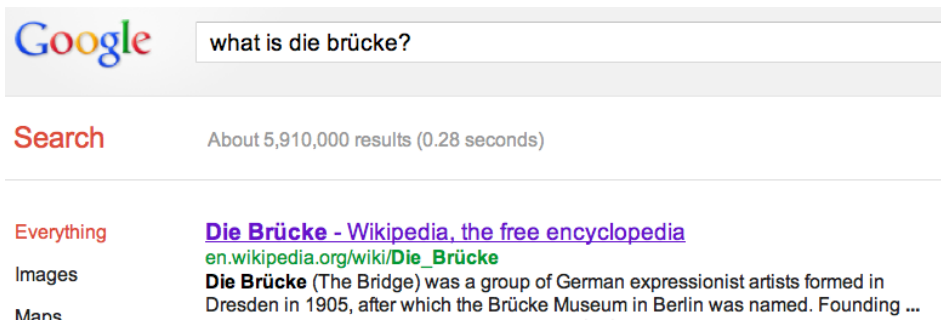
- Summarize a document with respect to an information need expressed in a user query.
- A kind of complex question answering:
 - Answer a question by summarizing a document that has the information to construct the answer



Summarization Methods



Simple Baseline: Take the First Sentence



Die Brücke

From Wikipedia, the free encyclopedia

For other uses, see [Die Brücke \(disambiguation\)](#).

Die Brücke (The Bridge) was a group of [German expressionist](#) artists formed in [Dresden](#) in 1905, after which the [Brücke Museum in Berlin](#) was named. Founding members were [Fritz Bleyl](#), [Erich Heckel](#), [Ernst Ludwig Kirchner](#) and [Karl Schmidt-Rottluff](#). Later members were [Emil Nolde](#), [Max Pechstein](#) and [Otto Mueller](#). The seminal group had a major impact on the evolution of [modern art](#) in the 20th century and the creation of expressionism.^[1]

Die Brücke is sometimes compared to the [Fauves](#). Both movements shared interests in [primitivist](#) art. Both



Snippets: Query-focused Summaries

Was cast-metal movable type invented in korea?

About 591,000 results (0.14 seconds)

[Movable type - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Movable_type

Jump to [Metal movable type](#): Transition from wood type to **metal** type occurred in 1234 ... The following description of the **Korean** font **casting** ... In the early fifteenth century, however, the **Koreans invented** a form of **movable type** that has ...

[History of printing in East Asia - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/History_of_printing_in_East_Asia

The following description of the **Korean** font **casting** process was recorded by the ... While **metal movable type** printing was **invented in Korea** and the oldest ...

[Korea, 1000–1400 A.D. | Heilbrunn Timeline of Art History | The ...](#)

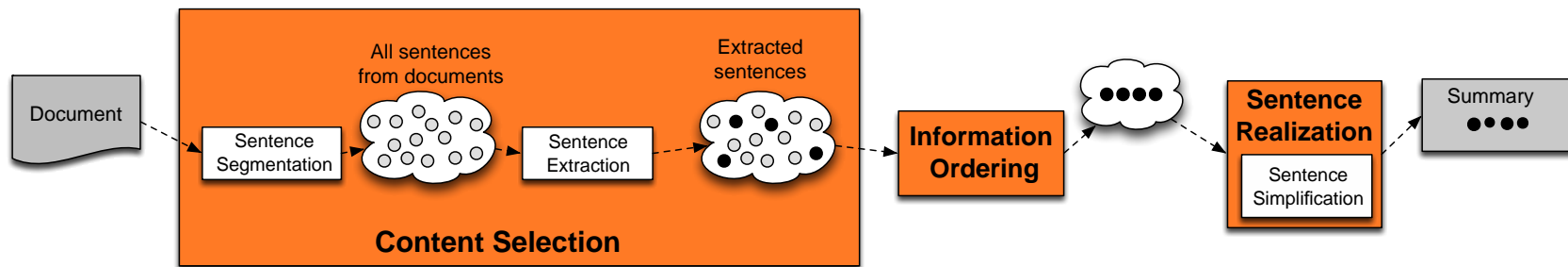
www.metmuseum.org/toah/ht/?period=07®ion=eak

The **invention** and use of **cast-metal movable type** in **Korea** in the early thirteenth century predates by two centuries Gutenberg's **invention** of metal **movable type** ...



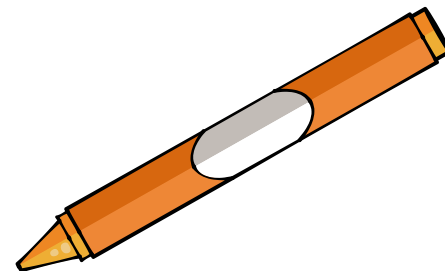
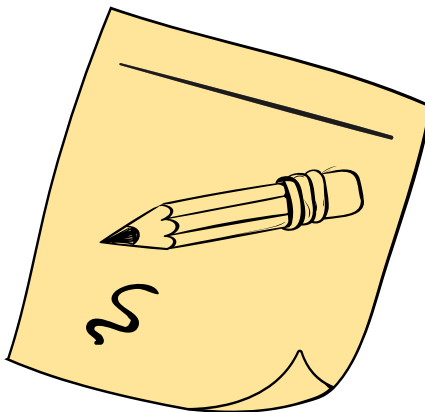
Summarization: Three Stages

1. **Content selection**: choose sentences to extract from the document
2. **Information ordering**: choose an order to place them in the summary
3. **Sentence realization**: clean up the sentences





Neural Text Summarization



Modeling Approach

- A **Language Model** predicts **next word**, given words predicted so far:

$$p(y_t | y_1, \dots, y_{t-1})$$

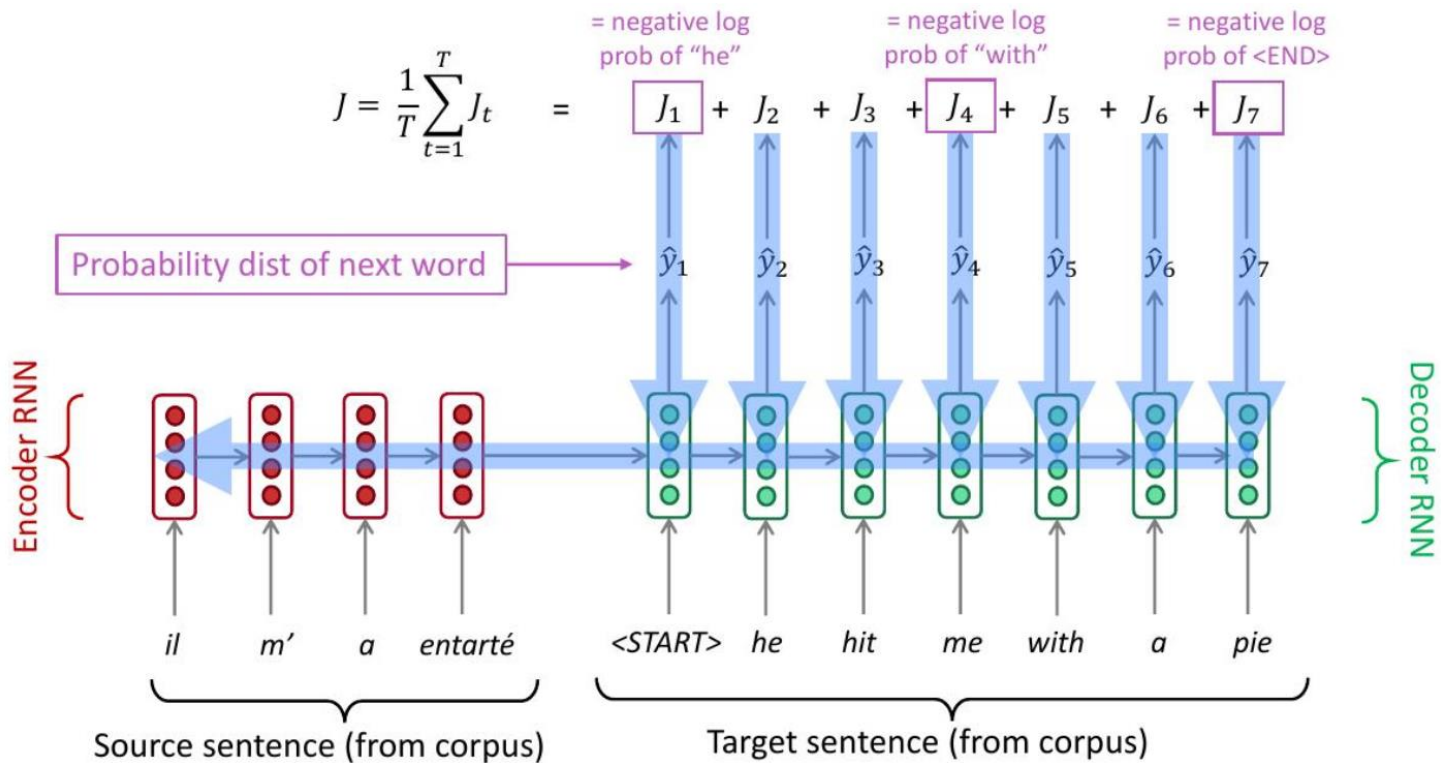
- A **Conditional Language Model** predicts next word, given words so far, and **also some other input x** :

$$p(y_t | y_1, \dots, y_{t-1}, x)$$

- For example an RNN can be used to model the above probability!
 - Machine Translation (x =source sentence, y =target sentence)
 - Summarization (x =input text, y =summarized text)
- **Task Definition:**
 - Given input text x , write summary y which is shorter and contains main information of x .



Modeling Approach



The CNN/Daily Mail Dataset

- Training data consists of **pairs** of news articles (average 800 words) and summaries (aka **story highlights**), usually 3 or 4 sentences long (average 56 words).
- CNN 100K pairs; Daily Mail 200K pairs.
- Highlights have been written by journalists, in compressed telegraphic manner.
- Does not form a coherent summary — each highlight relatively stand-alone, little co-referencing.
- Download from <https://github.com/abisee/cnn-dailymail>

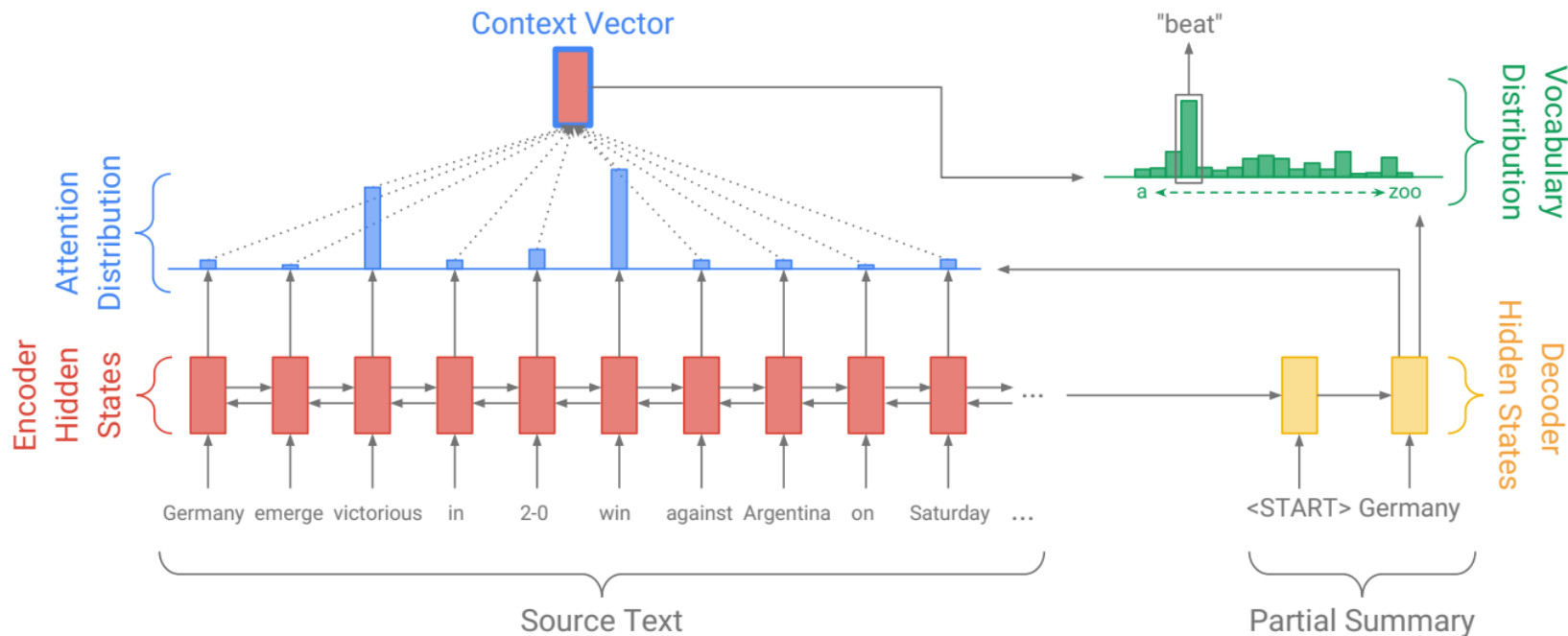


The CNN/Daily Mail Dataset

- **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:**
 - **69 percent of blacks** polled say **Martin Luther King Jr's vision** realized.
 - Slim majority of whites say King's vision not fulfilled.
 - King gave his "**I have a dream**" speech in 1963.



Sequence-to-Sequence Attentional Model



Source: Get To The Point: Summarization with Pointer-Generator Networks
AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Sequence-to-Sequence Attentional Model

- **Encoder:** single-layer bidirectional LSTM produces a sequence of **hidden states** h_i .
- **Decoder:** single-layer unidirectional LSTM receives word embedding of previous word emitted by decoder and has **decoder state** s_t .
- **Attention distribution:**
$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{attn})$$
$$a^t = \text{softmax}(e^t)$$
- **Context vector:** weighted sum of encoder hidden states $h_t^* = \sum_i a_i^t h_i$
- **Vocabulary distribution:** probability distribution over all words in the vocabulary, $P_{vocab} = \text{softmax}(V'(V[s_t; h_t^*] + b) + b')$
- **Training loss** for time step t is negative log-likelihood of target word w_t^* ,
$$loss_t = -\log P(w_t^*)$$

Source: Get To The Point: Summarization with Pointer-Generator Networks

AUT, Language Understanding Course, Fall 2022, Hossein Zeinali



Pointer-Generator Network

- Adding a copying mechanism (useful for rare words and phrases) to the Seq2Seq model.
- Model allows both **copying words by pointing**, and **generating words** from a fixed vocabulary.
- The attention distribution a^t and context vector h_t^* are calculated as before.
- On each decoder step, calculate p_{gen} , **probability of generating next word** (rather than copying it).

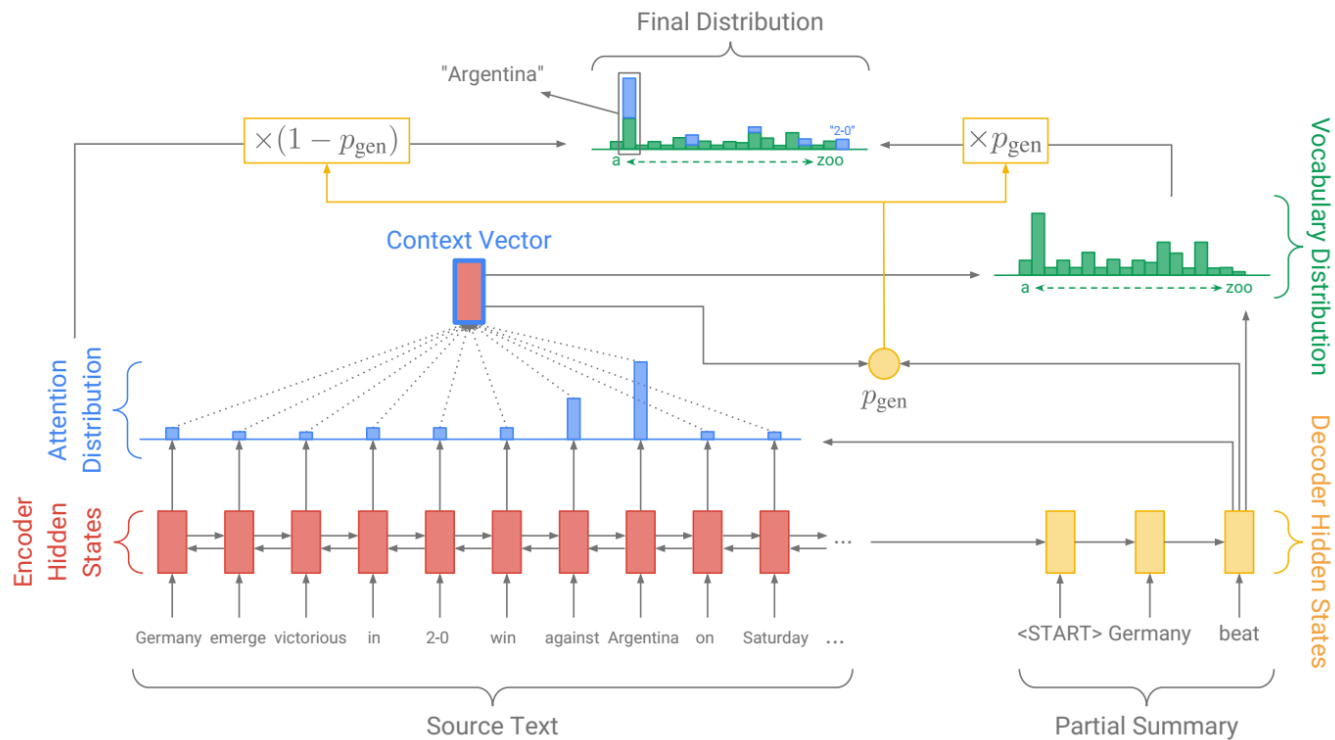
$$p_{gen} = \sigma(w_{h^*}^T h_t^* + w_s^T S_t + w_x^T x_t + b_{ptr})$$

- Next, p_{gen} is used to choose between generating a word from the vocabulary, or copying a word from the input sequence:

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



Pointer-Generator Network



Methods' Output Comparison

Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, *muhammadu buhari* told cnn's christiane amannpour that **he plans to aggressively fight corruption that has long plagued nigeria** and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, **he said his administration is confident it will be able to thwart criminals** and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. **the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.**

Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to **destabilize nigeria's economy**. UNK says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.



Coverage Mechanism

- Attempts to generate less repetitive summaries
- Penalizes repeatedly attending to same parts of the source text
- **Coverage vector** c^t tells us what has been attended so far:

$$c^t = \sum_{j=0}^{t-1} a^j$$

- Use coverage vector as extra input to attention mechanism:

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{attn})$$

- **Coverage loss** penalizes overlap between coverage vector c^t and new attention distribution a^t :

$$covloss_t = \sum_i \min(a_i^t, c_i^t)$$



Summarization Evaluation: ROUGE

- ROUGE stands for Recall Oriented Understudy for Gisting Evaluation
- Intrinsic metric for automatically evaluating summaries
 - Based on BLEU, but much more convenient
 - Not as good as human evaluation
- Given a document D , and an automatic summary X :
 - Have N humans produce a set of reference summaries of D
 - Run system, giving automatic summary X
 - What percentage of the n -grams from the reference summaries appear in X ?



Summarization Evaluation: ROUGE

$$ROUGE_N = \frac{\sum_{S \in \{RefSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{RefSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

- Like BLEU, it is based on **n-gram overlap**
- ROUGE has no brevity penalty and is based on **recall**
- Most commonly-reported ROUGE scores: ROUGE-1 **unigram** overlap, ROUGE-2 **bigram** overlap, and ROUGE-L **longest common subsequence** overlap



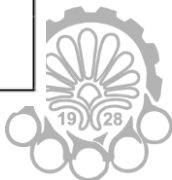
CNN/Daily Mail Corpus

- The dataset contains online news articles paired with multi-sentence summaries
 - News articles with 781 tokens on average
 - Summaries with 3.75 sentences or 56 tokens on average
- The processed version contains:
 - 287,226 training pairs
 - 13,368 validation pairs
 - 11,490 test pairs
- Models are usually evaluated with full-length F1-scores of ROUGE-1, ROUGE-2, ROUGE-L.
- Can be used for both Extractive and Abstractive models



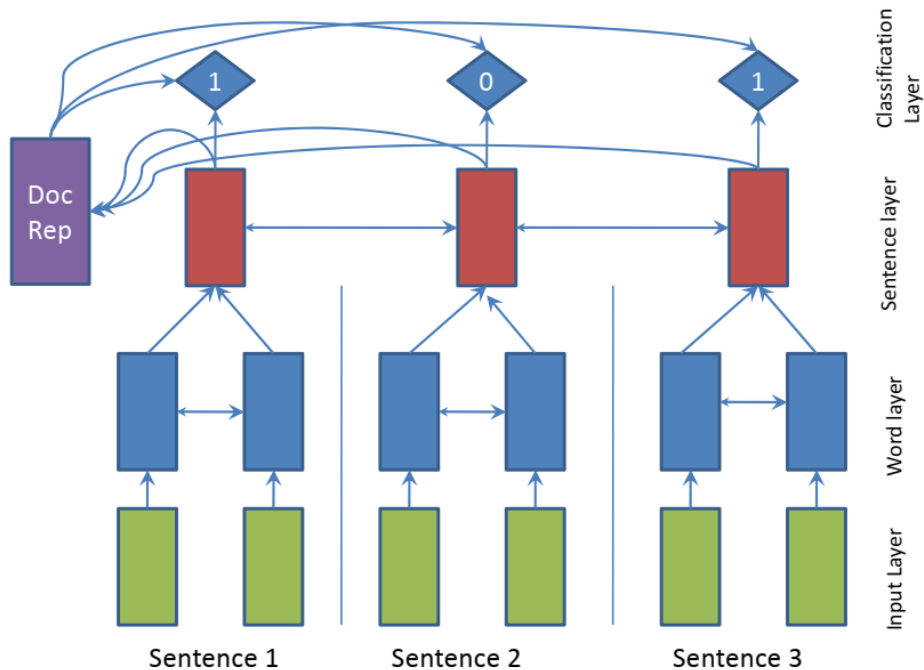
Comparison Results

Lead-3 baseline uses first three article sentences as the summary.	ROUGE		
	1	2	L
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08
seq-to-seq + attn baseline (50k vocab)	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 (ours)	40.34	17.70	36.57
lead-3 baseline (Nallapati et al., 2017)*	39.2	15.7	35.5
extractive model (Nallapati et al., 2017)*	39.6	16.2	35.3



Neural Extractive Models: SummaRuNNer

- Treat extractive summarization as a sequence classification problem

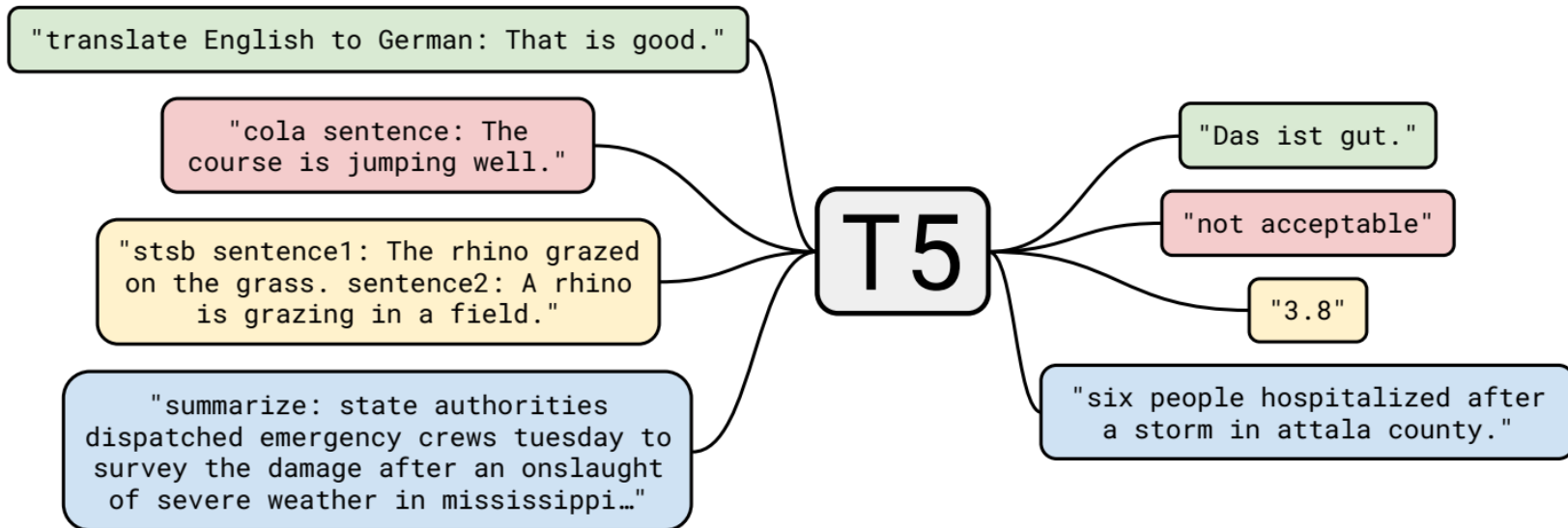


Results on CNN/Daily Mail Corpus

	Rouge-1	Rouge-2	Rouge-L
Lead-3	39.2	15.7	35.5
(Nallapati et al. 2016)	35.4	13.3	32.6
SummaRuNNer-abs	37.5	14.5	33.4
SummaRuNNer	39.6\pm0.2*	16.2\pm0.2*	35.3 \pm 0.2



Text-to-Text Transfer Transformer



Text-to-Text Transfer Transformer

- Train on supervised and unsupervised tasks

Original text

Thank you ~~for~~ ~~inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>



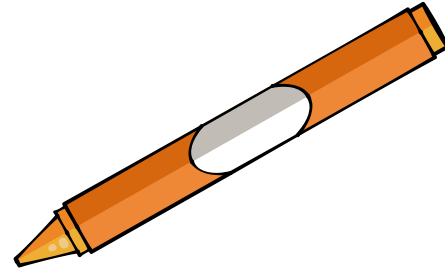
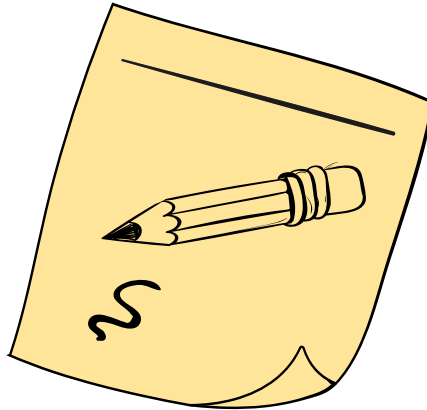
T5 Results

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	39.77	24.04

The ROUGE-2-F score is reported for CNNDM



Speech Summarization



Speech Summarization

- Speech summarization entails ‘summarizing’ speech
 - Identify important information relevant to users and the story
 - Represent the important information
 - Present the extracted/inferred information as an addition or substitute to the story
- Why speech summarization?
 - Multimedia production and size are increasing: need less time-consuming ways to archive, extract, use and browse speech data - speech summarization, a possible solution
 - Due to temporal nature of speech, difficult to scan like text
 - User-specific summaries of broadcast news is useful
 - Summarizing voicemails can help us better organize voicemails



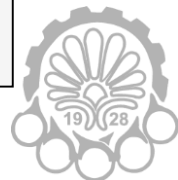
Comparing Speech and Text Summarization

• Alike

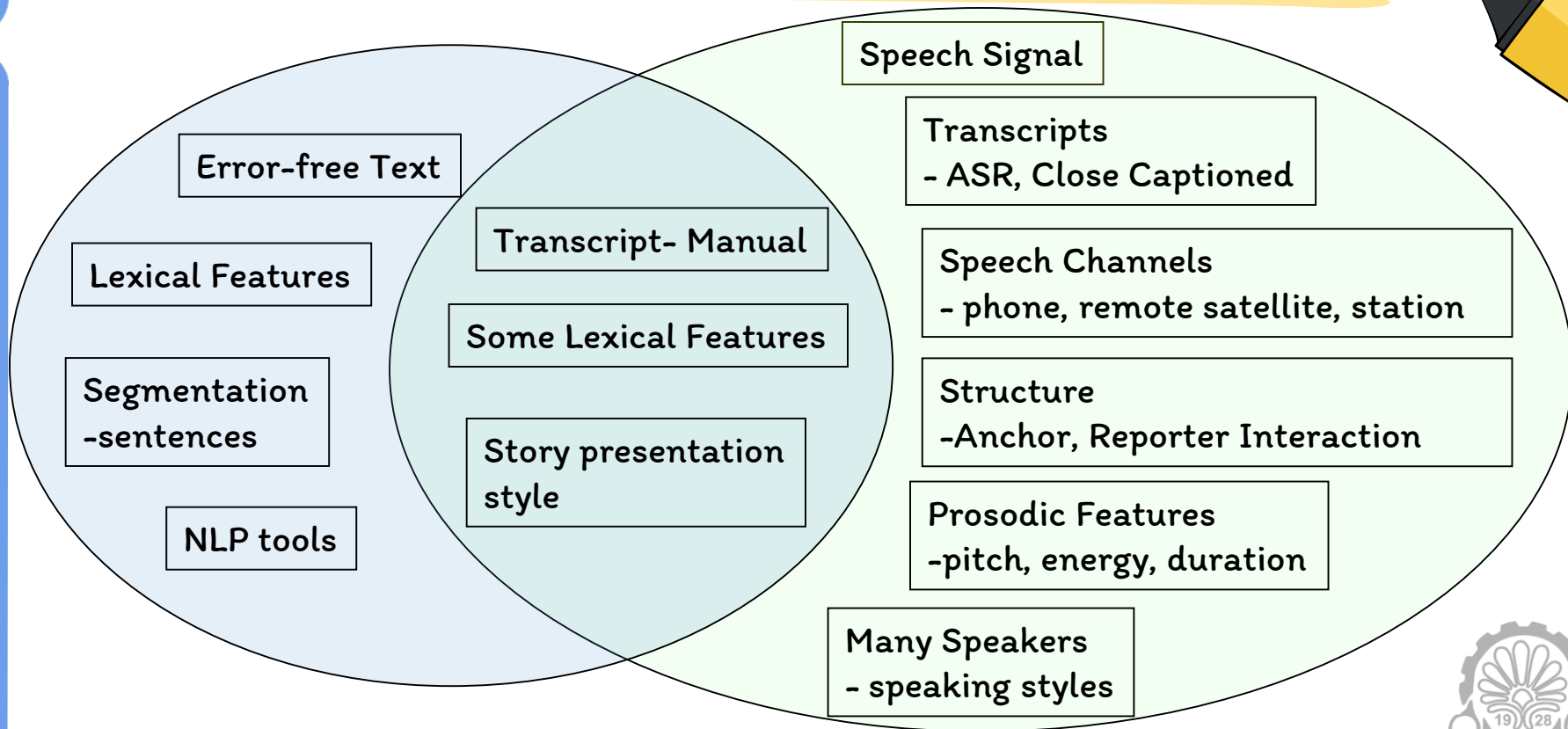
- Identifying important information
- Some lexical, discourse features
- Extraction or generation or compression

• Different

- Speech Signal
- Prosodic features
- NLP tools?
- Segments – sentences?
- Generation?
- Errors
- Data size



Text vs. Speech Summarization (NEWS)

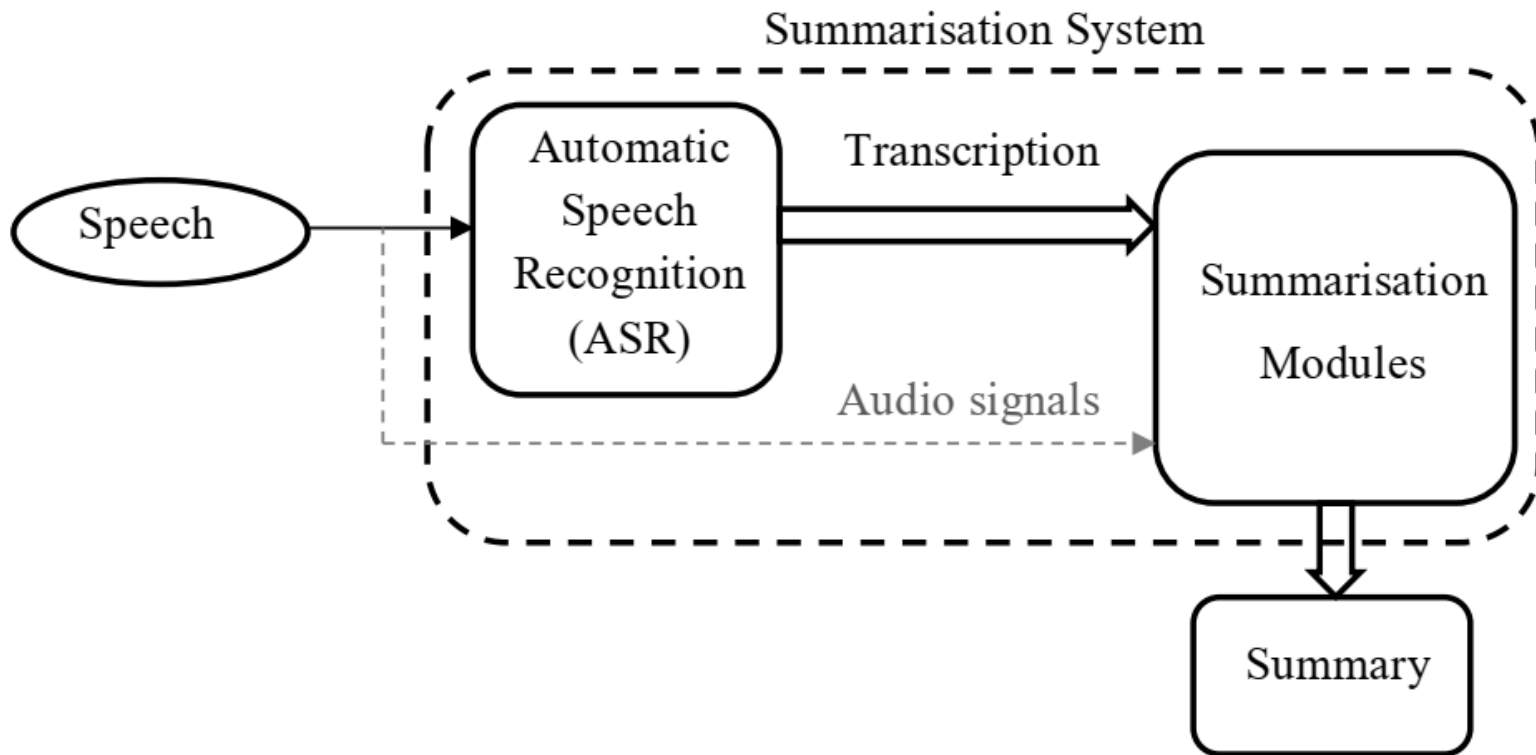


Speech Summarization

- Speech Summarization seems to be mostly based on extractive summarization
- Extraction of words, sentences, content units
- Some compression methods have also been proposed
- Generation as in some text-summarization techniques is not available/feasible
 - Mainly due to the nature of the content



General Structure of a System





Thanks for your attention



References and IP Notice

- Daniel Jurafsky and James H. Martin, “Speech and Language Processing”, 3rd ed., 2019
- Some of the slides were selected from Daniel Jurafsky's slides.
- Some of the slides were selected from Mirella Lapata's slides.
- Some of the slides were selected from Chan Young Park's slides.
Some of the slides were selected from Sameer Maskey's slides.
- Some graphics were selected from [Slidesgo](#) template

