

Natural language processing Applications

Week 7: Text Summarization



fit@hcmus

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

- ❑ Introduction to Text Summarization
- ❑ Text Summarization Methods
- ❑ Evaluation Measures for Text Summarization



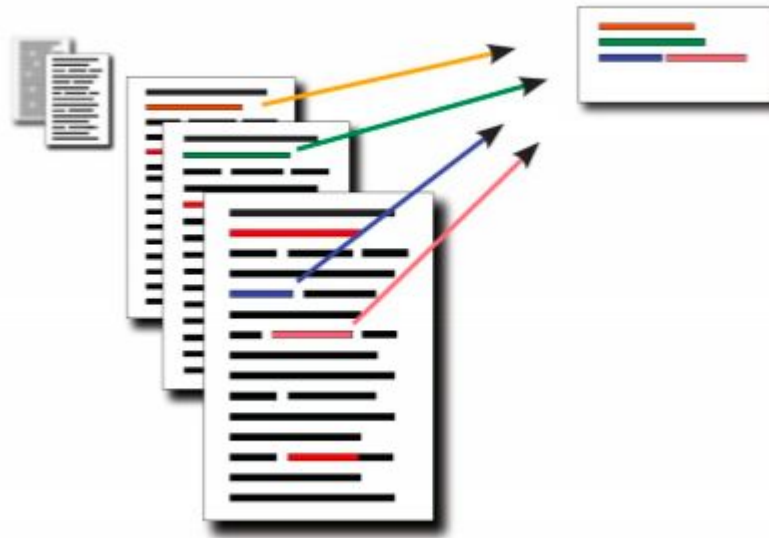
NLPA – Text Summarization

INTRODUCTION TO TEXT SUMMARIZATION



Introduction to Text Summarization

- ❑ Objective: To create a condensed version of the most relevant information found in the document or relevant information to the user.



Introduction to Text Summarization (cont)

- ❑ Text Summarization Applications:
 - ❑ Sketching or summarizing documents, articles...
 - ❑ Summarizing emails
 - ❑ Summarizing meetings
 - ❑ Simplifying documents by sentence compression
 - ❑ ...



Introduction to Text Summarization(cont)

- ❑ Input:
 - ❑ Single document
 - ❑ Multiple documents
- ❑ Output:
 - ❑ Documents using extractive methods
 - ❑ Documents using abstractive methods
- ❑ Focus:
 - ❑ Generic Text Summarization
 - ❑ Query-based Text Summarization

Introduction to Text Summarization(cont)

- ❑ Single-document Summarization: generating a new document from a single input document, approaches:
 - ❑ Abstract
 - ❑ Outline
 - ❑ Headline
- ❑ Multi-document Summarization: representing a set of documents with a short piece of text, approaches:
 - ❑ A series of summaries of the same event.
 - ❑ A series of summaries some topics or questions



Introduction to Text Summarization(cont)

❑ 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.

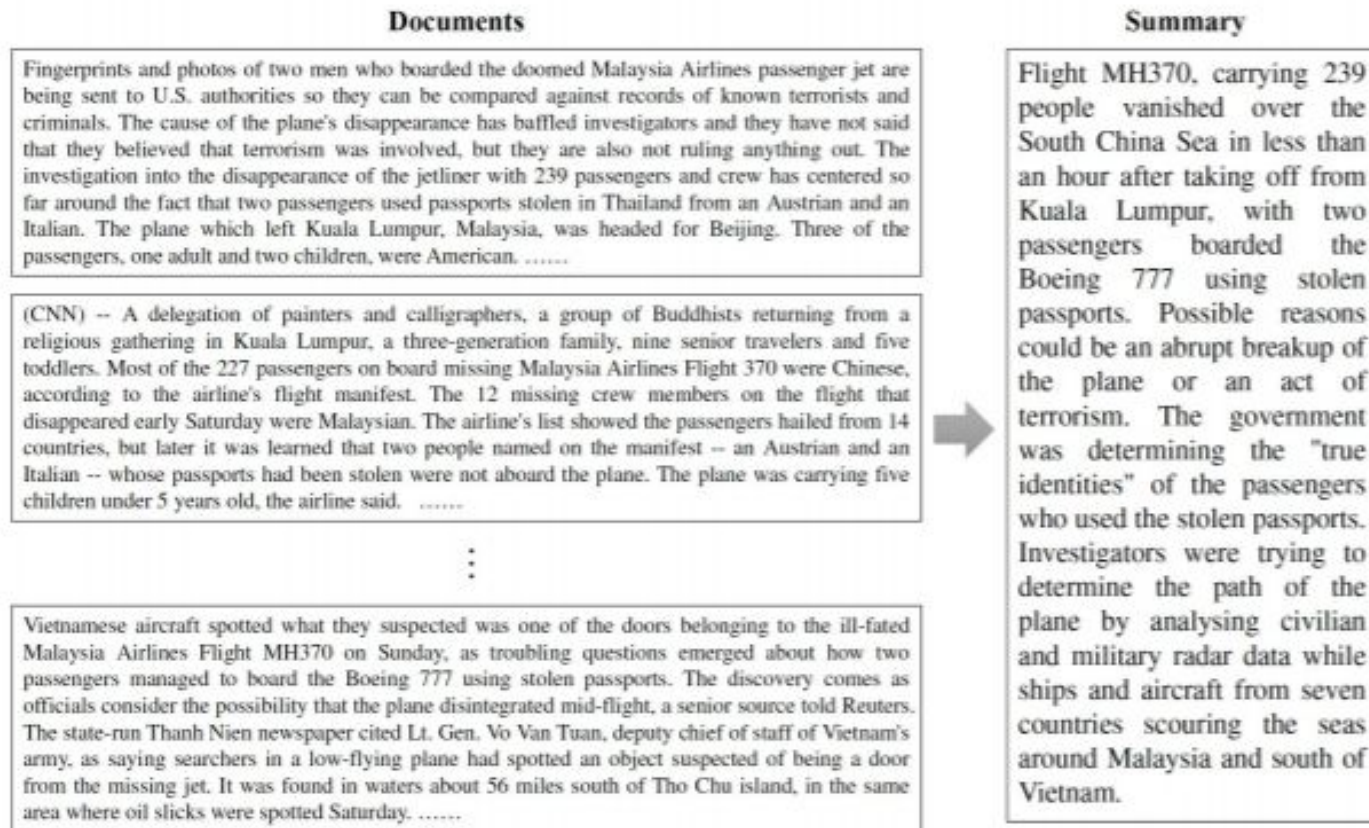
It is detrimental to the sovereignty of Cambodia," he said. Hun Sen's Cambodian People's Party won 64 of the 122 parliamentary seats in July's elections, short of the two-thirds majority needed to form a government on its own. Ranariddh and Sam Rainsy have charged that Hun Sen's victory in the elections 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

Introduction to Text Summarization(cont)

❑ Multi-document Summarization



Introduction to Text Summarization(cont)

- ❑ Generic Text Summarization:
 - ❑ Summarize the content of a text
- ❑ Query-based Text Summarization:
 - ❑ Summarize a document according to the information required by a query.
 - ❑ A kind of complex question answering: answer a question by summarizing a document to put into the answer.

Introduction to Text Summarization(cont)

- ❑ Extractive Summarization:
 - ❑ Generate summary using existing phrases or sentences in the source document
- ❑ Abstractive summarization:
 - ❑ Generate summary by paraphrasing ideas in the source document



NLPA – Text Summarization

TEXT SUMMARIZATION METHODS



Text Summarization methods

- ❑ Basic method: get the first sentence

Google search results for "what is die brücke?". The search bar shows the query and the results count: "About 5,910,000 results (0.28 seconds)". The first result is "Die Brücke - Wikipedia, the free encyclopedia" with the URL "en.wikipedia.org/wiki/Die_Brücke". The snippet below the title reads: "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 ...". Below the search results, the full Wikipedia article snippet is displayed under the heading "Die Brücke". The snippet starts with "From Wikipedia, the free encyclopedia" and includes a disambiguation note: "For other uses, see *Die Brücke* (disambiguation).". The main text of the snippet states: "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]". The snippet ends with "Die Brücke is sometimes compared to the Fauves. Both movements shared interests in primitivist art. Both".

Google search results for "what is die brücke?".

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

Everything [Die Brücke - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Die_Brücke
Images **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 ...
Maps

Die Brücke

From Wikipedia, the free encyclopedia

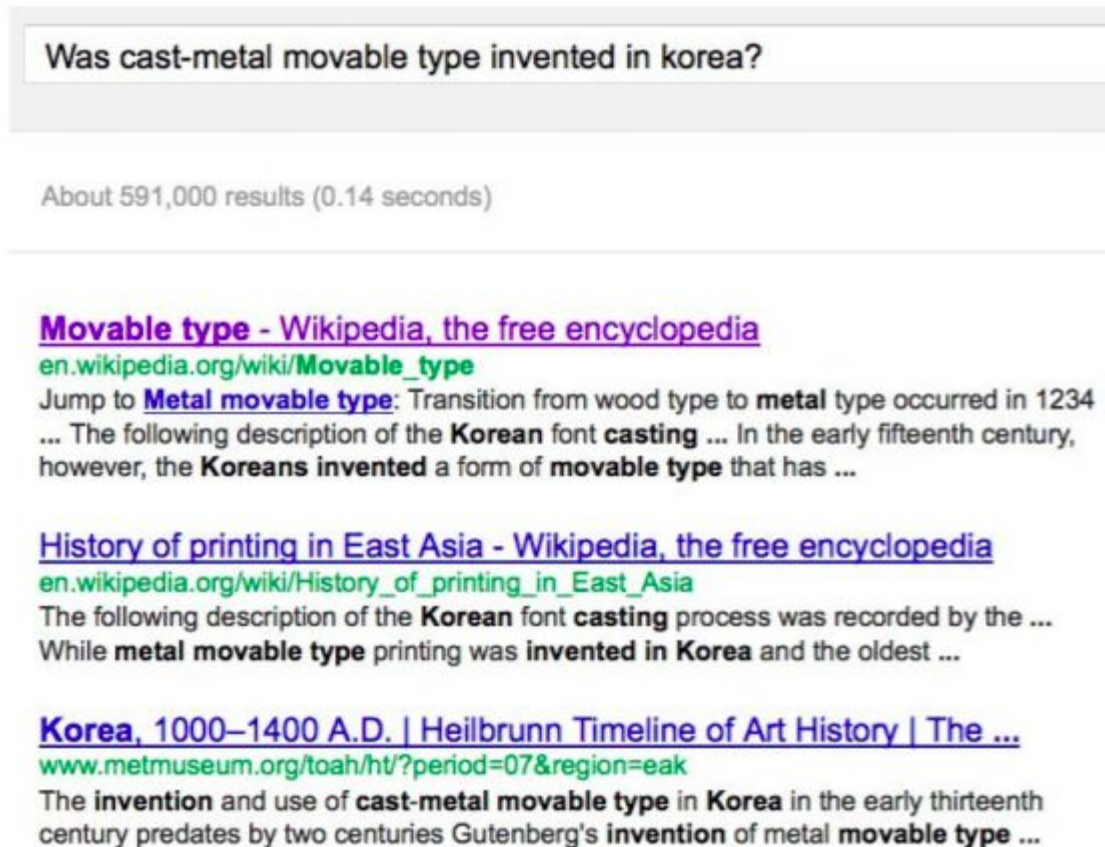
For other uses, see [Die Brücke](#) (disambiguation).

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

Text Summarization methods

❑ Snippets method



Was cast-metal movable type invented in korea?

About 591,000 results (0.14 seconds)

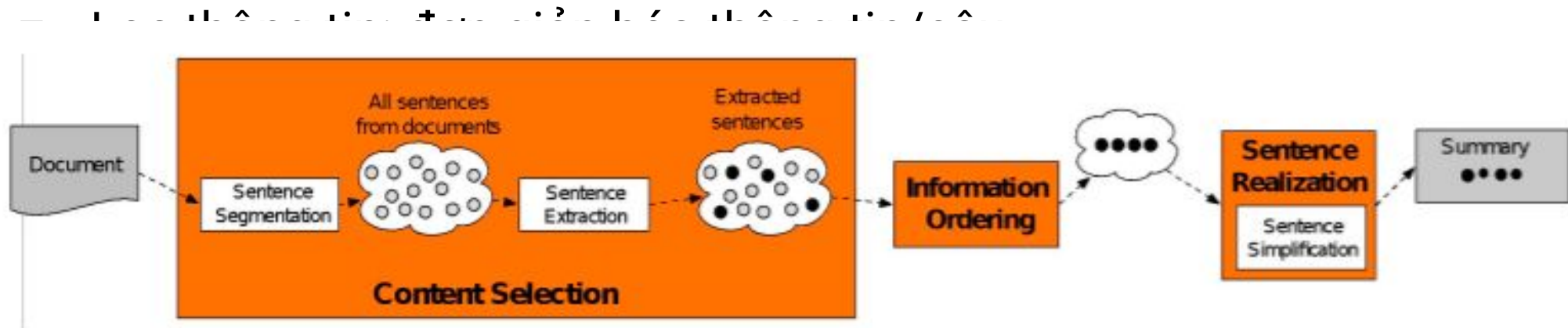
[Movable type - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Movable_type
Jump to [Metal movable type](#): Transition from wood type to **metal** type occurred in 1234 ... The following description of the **Korean** font **casting** ... In the early fifteenth century, however, the **Koreans 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** ...

Text Summarization methods

- ❑ Steps of Text Summarization:
 - ❑ Content selection: select sentences to extract from the document
 - ❑ Information sorting: choose the orders for the information in the summary.



Text Summarization methods

- ❑ Content selection:

- ❑ Luhn (1958): Choose sentences with salient information or informative sentences

- ❑ The method of identifying salient information:

- ❑ *tf-idf*: the weight of term w_i in document j : $\text{weight}(w_i) = \text{tf}_{ij} * \text{idf}_i$

- ❑ Theme sign: choose a smaller set from salient words

- ❑ Mutual information

- ❑ Log-likelihood ratio (Dunning (1993), Lin and Hovy (2000))

$$\text{weight}(w_i) = \begin{cases} 1 & \text{if } -2 \log \lambda(w_i) > 10 \\ 0 & \text{otherwise} \end{cases}$$

Text Summarization methods

- ❑ Content selection:

- ❑ The method of identifying informative words (Conroy, Schlesinger, and O'Leary 2006)

- ❑ Log-likelihood ratio

- ❑ Appearing in the query

$$weight(w_i) = \begin{cases} 1 & \text{if } -2 \log \lambda(w_i) > 10 \\ 1 & \text{if } w_i \in \text{question} \\ 0 & \text{otherwise} \end{cases}$$

- ❑ Weighting sentences by weights of words in the sentence

$$weight(s) = \frac{1}{|S|} \sum_{w \in S} weight(w)$$

Text Summarization methods

Content selection:

Sentence extraction using unsupervised methods (Rada Mihalcea,

AC

$$\text{Similarity}(S_i, S_j) = \frac{|W_k|_{W_k \in S_i \& W_k \in S_j}}{\log(|S_i|) + \log(|S_j|)}$$

3: BC-Hurricane Gilbert, 09-11 339
4: BC-Hurricane Gilbert, 0348
5: Hurricane Gilbert heads toward Dominican Coast
6: By Ruddy Gonzalez
7: Associated Press Writer
8: Santo Domingo, Dominican Republic (AP)
9: Hurricane Gilbert Swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12: Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14: Tropical storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
15: The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17: The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.
19: There were no reports on casualties.
20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24: The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

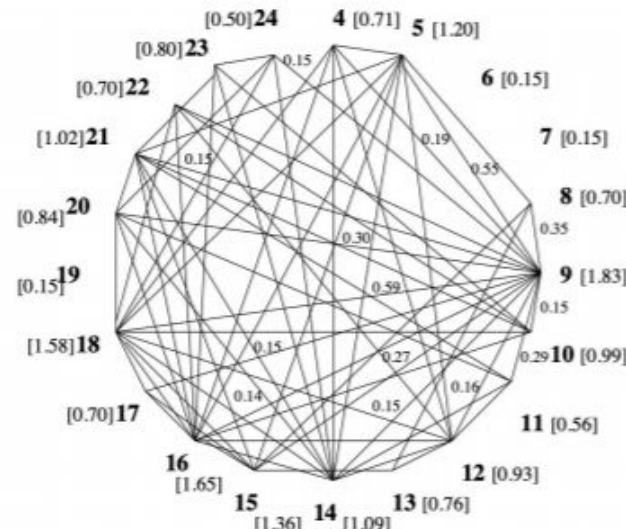


Figure 1: Sample graph build for sentence extraction from a newspaper article.

Text Summarization methods

- ❑ Content selection:
 - ❑ Sentence extraction using supervised methods:
 - ❑ Data: a set of training data consisting of summarized documents and source documents
 - ❑ Alignment: sentences in summarized documents and source documents
 - ❑ Feature extraction:
 - ❑ Position (ex: the first sentence)
 - ❑ Sentence length
 - ❑ Informative words, signal phrases
 - ❑ Cohesion



Text Summarization methods

- ❑ Content selection:

- ❑ Sentence extraction using supervised methods (cont):

- ❑ Training: a classifier (whether put the sentence into the summary: Yes or No)

- ❑ Problem:

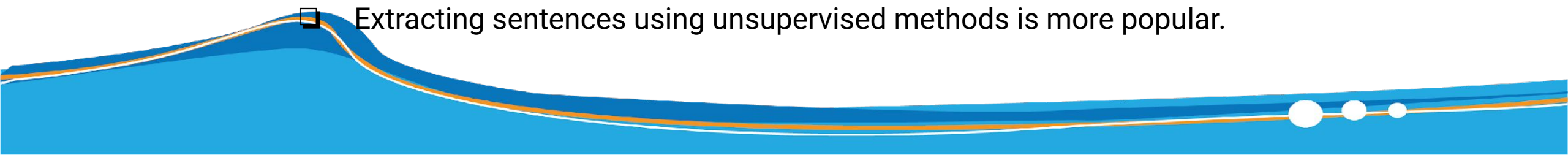
- ❑ Lack of data

- ❑ Difficult to align

- ❑ The performance is no better than performance of unsupervised methods

- ❑ In reality:

- ❑ Extracting sentences using unsupervised methods is more popular.



NLPA – Text Summarization

EVALUATION MEASURES FOR TEXT SUMMARIZATION



Evaluation Measures for Text Summarization

❑ ROUGE (Recall Oriented Understudy for Gisting Evaluation)

❑ An intrinsic evaluation metric for automatic summarization:

- ❑ A modification of BLEU (an evaluation metric for machine translation)
- ❑ Not as good as human judgement
- ❑ Convenient

❑ For a document D and a summary X:

- ❑ There are N summaries created manually from D (referential summary)
- ❑ Run summary system to generate summary X
- ❑ Per

$$ROUGE-2 = \frac{\sum_{s \in \{RefSummaries\}} \sum_{\text{bigrams } i \in S} \min(count(i, X), count(i, S))}{\sum_{s \in \{RefSummaries\}} \sum_{\text{bigrams } i \in S} count(i, S)}$$

n X

Evaluation Measures for Text Summarization

- ❑ ROUGE (Recall Oriented Understudy for Gisting Evaluation)
 - ❑ Question: “What is water spinach?”
 - ❑ System output: “Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.”
 - ❑ Referential summary (gold corpus):
 - ❑ Human 1: Water spinach is a green leafy vegetable grown in the tropics.
 - ❑ Human 2: Water spinach is a leafy vegetable commonly grown as a vegetable.
 - ❑ Human 3: Water spinach is a leafy vegetable commonly eaten in tropical areas of Asia.

$$\text{ROUGE-2} = \frac{3 + 3 + 6}{10 + 9 + 9} = 12/28 = .43$$

NLPA – Text Summarization

ABSTRACTIVE TEXT SUMMARIZATION



Abstractive Text Summarization

- ❑ Machine Translation using Neural Networks:
 - ❑ Architecture of Seq2seq model with attention (Bahdanau, 2014)

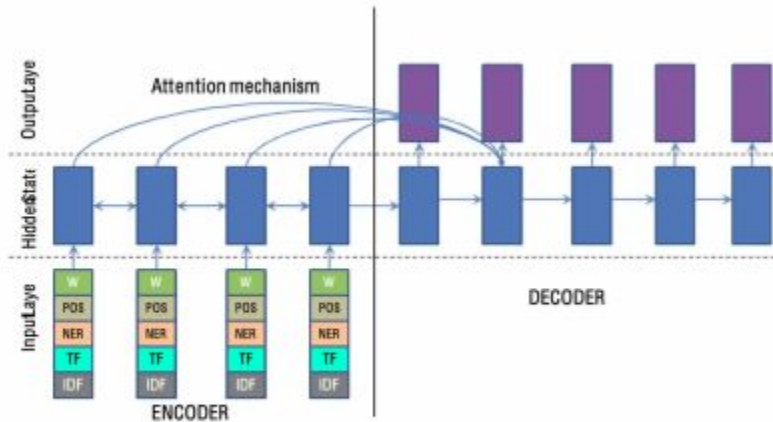


Figure 1: Feature-rich-encoder: We use one embedding

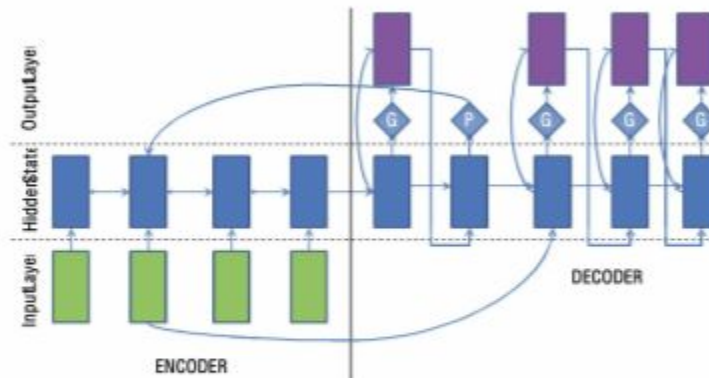


Figure 2: Switching generator/pointer model: When the

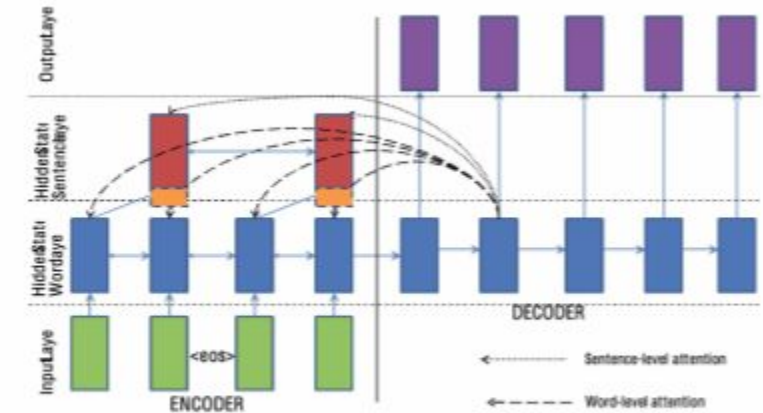


Figure 3: Hierarchical encoder with hierarchical attention: