

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

1

What to summarize? Single vs. multiple documents

- **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

2

Query-focused Summarization & Generic 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

3

Summarization for Question Answering: Snippets

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

The image shows a Google search interface. The search bar contains the text "what is die brücke?". Below the search bar, it says "Search" and "About 5,910,000 results (0.28 seconds)". The search results are categorized by "Everything", "Images", "Maps", "Videos", "News", "Shopping", "Applications", and "More". The "Everything" category is selected, showing a snippet for "Die Brücke - Wikipedia, the free encyclopedia". The snippet includes the title, a link to the Wikipedia page, and a short description: "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". Below this, there are more search results for "Die Brücke (film) - Wikipedia, the free encyclopedia" and "Die Brücke - Die Brücke Art".

4

Summarization for Question Answering: 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

5

Extractive summarization & Abstractive summarization

- **Extractive summarization:**
 - create the summary from phrases or sentences in the source document(s)
- **Abstractive summarization:**
 - express the ideas in the source documents using (at least in part) different words

6

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]

7

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

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

Lin and Hovy 2003

- Intrinsic metric for automatically evaluating summaries
 - Based on BLEU (a metric used for machine translation)
 - Not as good as human evaluation
 - But much more convenient (and some correlation with humans)
- Given a document D, and an automatic summary X:
 1. Have N humans produce a set of reference summaries of D
 2. Run system, giving automatic summary X
 3. What percentage of the bigrams from the reference summaries appear in X?

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

8

A ROUGE example:

Q: “What is water spinach?”

Human 1: Water spinach is a green leafy vegetable grown in the tropics.

Human 2: Water spinach is a semi-aquatic tropical plant grown as a vegetable.

Human 3: Water spinach is a commonly eaten leaf vegetable of Asia.

- System answer: Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

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

9