

Natural Language Processing

CSE 447

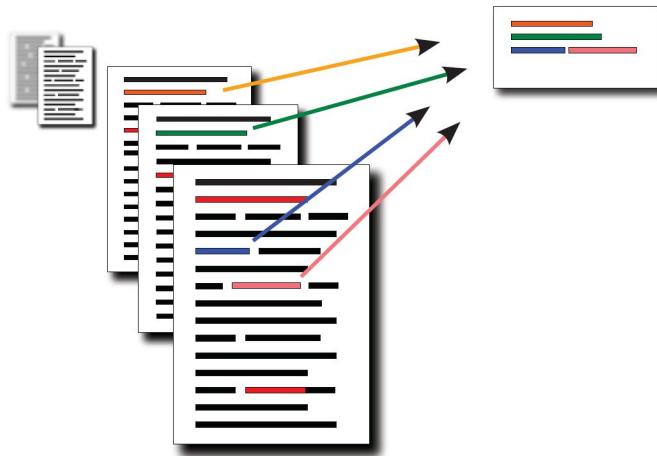
Text summarization

Chan Young Park

chanyoun@cs.cmu.edu

Text Summarization

- **Goal:** produce an abridged version of a text that contains information that is *important* or *relevant* to a user.



Text Summarization

- Applications of Summarization
 - outlines or abstracts of any document, article, etc
 - summaries of email threads
 - action items from a meeting
 - simplifying text by compressing sentences

Categories

- **Input**

- Single-Document Summarization (SDS)
- Multiple-Document Summarization (MDS)

- **Output**

- Extractive
- Abstractive

- **Focus**

- Generic
- Query-focused Summarization

- **Machine learning methods**

- Supervised
- Unsupervised

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

Single-document Summarization

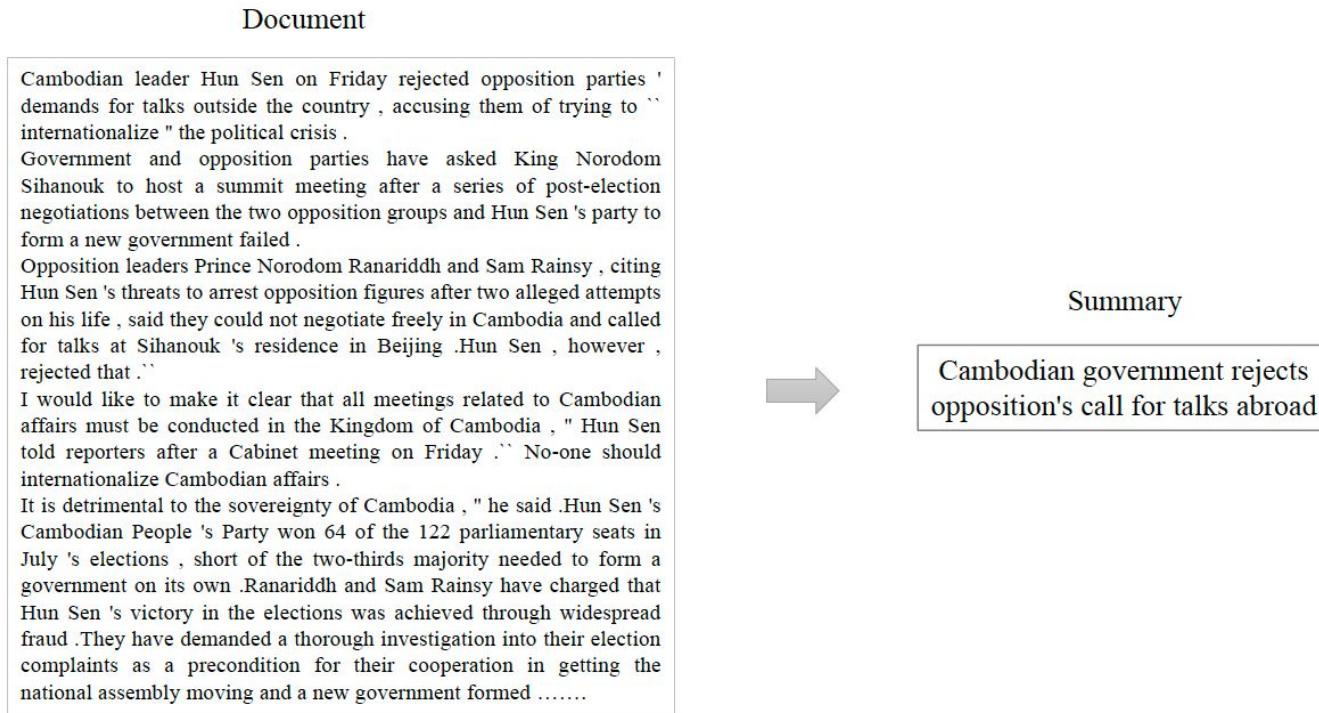


Figure 1: Single-document summarization.

Multiple-document Summarization

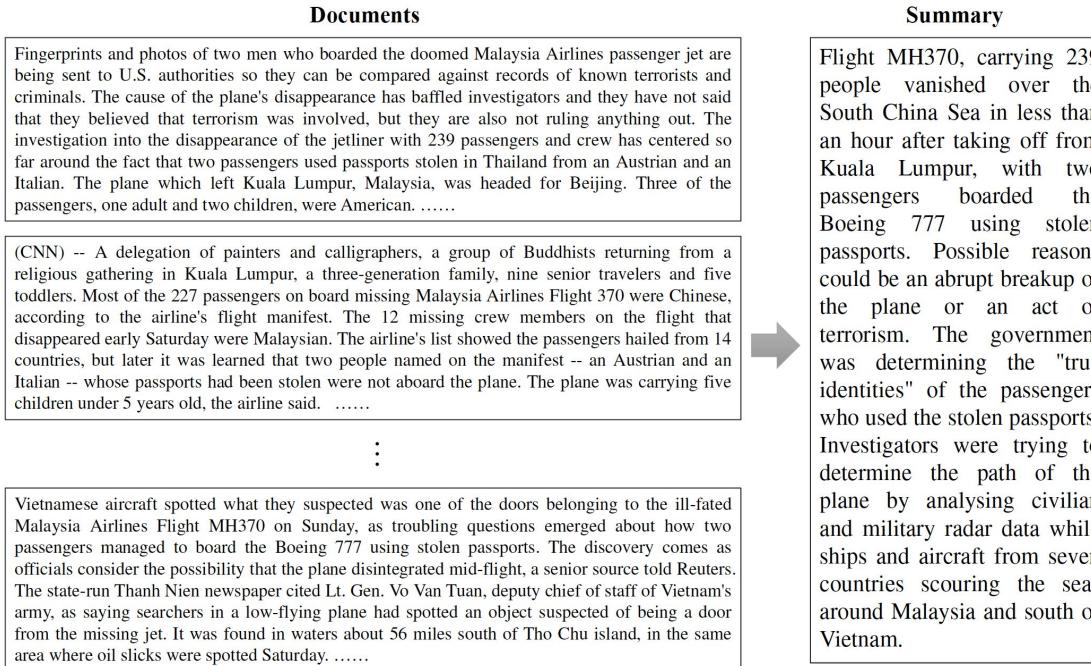


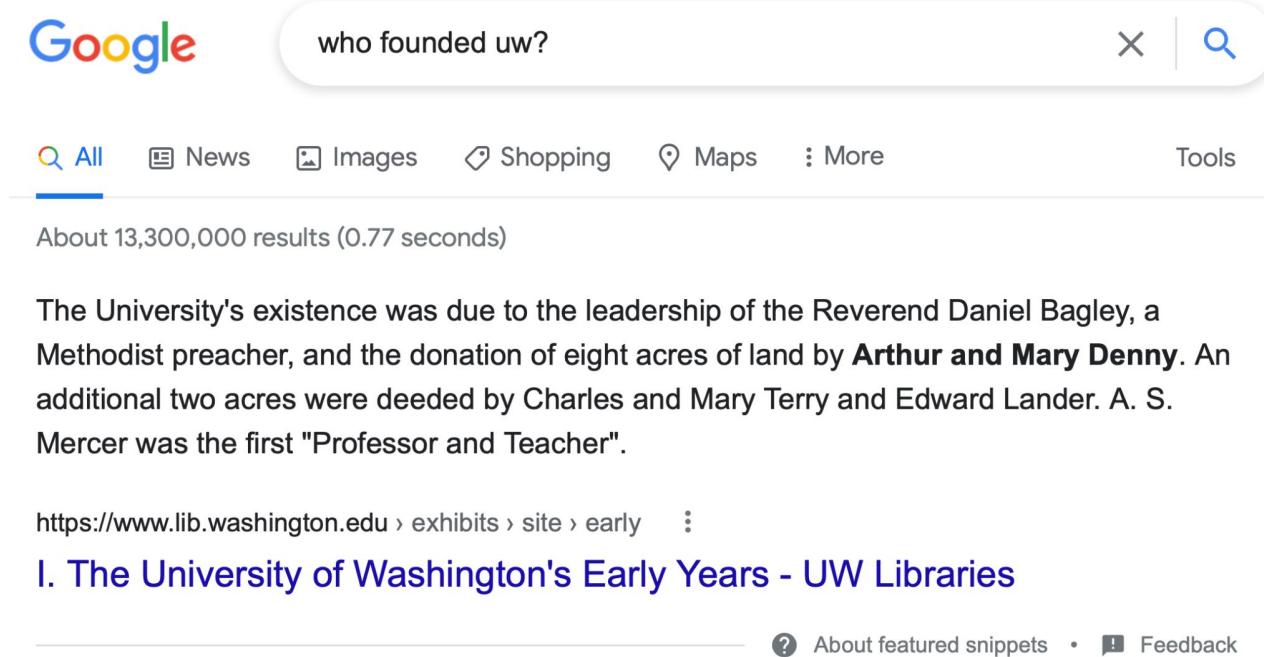
Figure 2: Multi-document summarization for the topic “Malaysia Airlines Disappearance”.

Generic & Query-focused 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 for Question Answering: Snippets

- Create snippets summarizing a web page for a query



The screenshot shows a Google search results page. The search bar contains the query "who founded uw?". Below the search bar, there are navigation links for All, News, Images, Shopping, Maps, More, and Tools. A message indicates "About 13,300,000 results (0.77 seconds)". The main content area displays a snippet of text: "The University's existence was due to the leadership of the Reverend Daniel Bagley, a Methodist preacher, and the donation of eight acres of land by **Arthur and Mary Denny**. An additional two acres were deeded by Charles and Mary Terry and Edward Lander. A. S. Mercer was the first "Professor and Teacher".". Below this snippet, the URL "https://www.lib.washington.edu › exhibits › site › early" and three vertical dots are shown. A large blue link labeled "I. The University of Washington's Early Years - UW Libraries" is prominently displayed.

Extractive & 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

History of Summarization

- Since 1950s:
 - Concept Weight (Luhn, 1958), Centroid (Radev et al., 2004), LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004), Sparse Coding (He et al., 2012; Li et al., 2015) Feature+Regression (Min et al., 2012; Wang et al., 2013)
- Most of the summarization methods are extractive. (pre deep learning)
- Abstractive summarization is full of challenges.
 - Some indirect methods employ sentence fusing (Barzilay and McKeown, 2005) or phrase merging (Bing et al., 2015).
 - The indirect strategies will do harm to the linguistic quality of the constructed sentences.

How to detect salient words/sentences? (= Salience Detection)

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

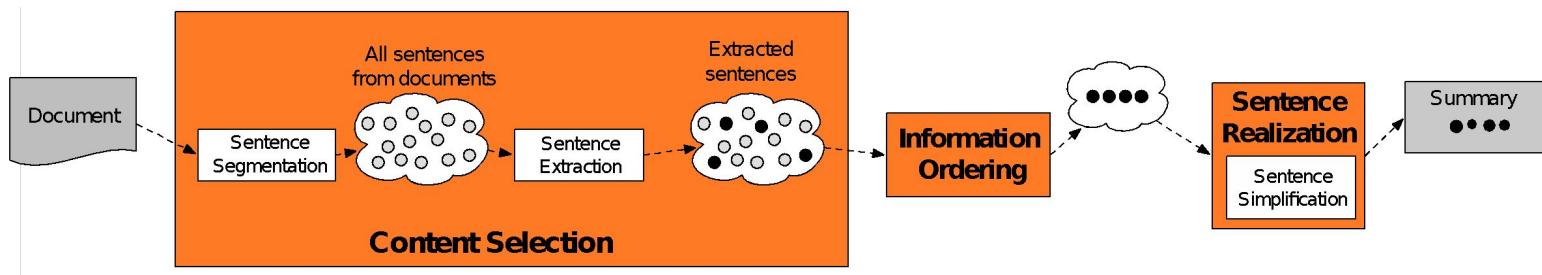
The screenshot shows a Google search results page. The search query "what is die brücke?" is entered in the search bar. The search button is labeled "Search". Below the search bar, it says "About 5,910,000 results (0.28 seconds)". On the left, there are filters: "Everything", "Images", and "Maps". The main result is a link to the Wikipedia page for "Die Brücke - Wikipedia, the free encyclopedia" at en.wikipedia.org/wiki/Die_Brücke. The snippet from the Wikipedia page 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 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." There are also other links and snippets below the main result.

Snippets: query-focused summaries

Was cast-metal movable type invented in korea?

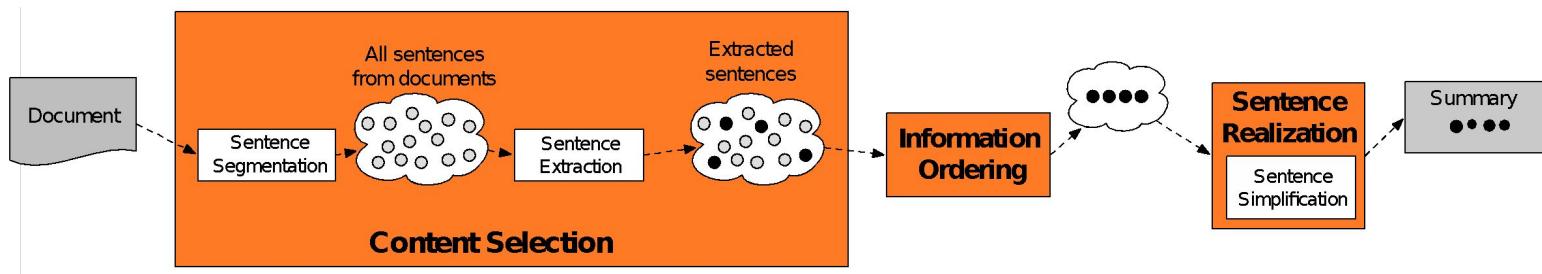
Summarization: Three Stages

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



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



Unsupervised content selection

- Intuition dating back to Luhn (1958):
 - Choose sentences that have **salient** or **informative** words
- Two approaches to defining salient words
 1. **tf-idf:** weigh each word w_i in document j by tf-idf
(tf: term-frequency, idf: inverse-document-frequency)
$$weight(w_i) = tf_{ij} \times idf_i$$
 2. **topic signature:** choose a smaller set of salient words
 - Use log-likelihood ratio (LLR) Dunning (1993), Lin and Hovy (2000)

Unsupervised content selection

- Topic signature: choose a smaller set of salient words

| Topic 10 Signature Terms of Topic 258 — Computer Security | | | | | |
|---|-----------------|--------------------|-----------------|-------------------------------|-----------------|
| Unigram | $-2\log\lambda$ | Bigram | $-2\log\lambda$ | Trigram | $-2\log\lambda$ |
| computer | 1159.351 | computer security | 213.331 | jet propulsion laboratory | 98.854 |
| virus | 927.674 | graduate student | 178.588 | robert t. mo | 98.854 |
| hacker | 887.377 | computer system | 146.328 | cornell university graduate | 79.081 |
| morris | 666.392 | research center | 132.413 | lawrence berkeley laboratory | 79.081 |
| cornell | 385.684 | computer virus | 126.033 | nasa jet propulsion | 79.081 |
| university | 305.958 | cornell university | 108.741 | university graduate student | 79.081 |
| system | 290.347 | nuclear weapon | 107.283 | lawrence livermore national | 69.195 |
| laboratory | 287.521 | military computer | 106.522 | livermore national laboratory | 69.195 |
| lab | 225.516 | virus program | 106.522 | computer security expert | 66.196 |
| mcclary | 128.515 | west german | 82.210 | security center bethesda | 49.423 |

Topic signature-based content selection w/ queries

Conroy, Schlesinger, and O'Leary 2006

- choose words that are informative either
 - by log-likelihood ratio (LLR)
 - or by appearing in the query

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

(could learn more complex weights)

- Weigh a sentence (or window) by weight of its words:

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

Graph-based Ranking Algorithms

- unsupervised sentence extraction

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

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.

Rada Mihalcea, ACL 2004

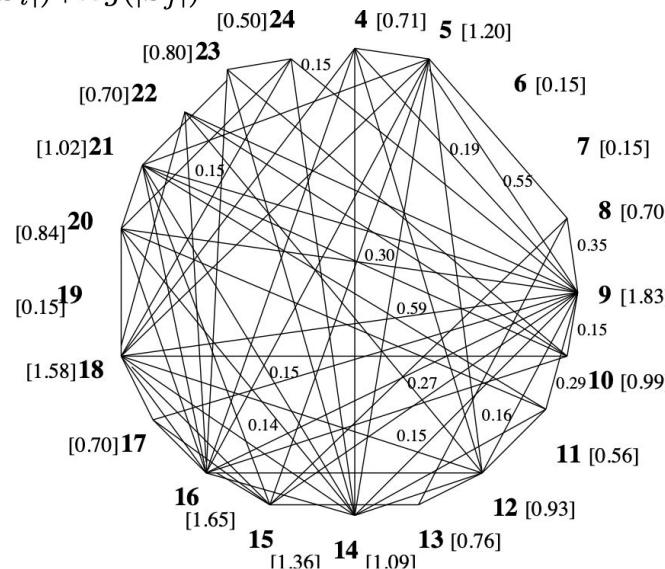


Figure 1: Sample graph build for sentence extraction from a newspaper article.

CNN/DM dataset

STORY HIGHLIGHTS

Trump will head to Texas on Tuesday

The White House has yet to say where Trump will travel

Washington (CNN) — President Donald Trump struck a unifying tone Monday as he addressed the devastation in Texas wrought by Hurricane Harvey at the top of a joint news conference with Finland's president.

"We see neighbor helping neighbor, friend helping friend and stranger helping stranger," Trump said. "We are one American family. We hurt together, we struggle together and believe me, we endure together."

Trump extended his "thoughts and prayers" to those affected by the hurricane and catastrophic flooding that ensued in Texas, and also promised Louisiana residents that the federal government is prepared to help as the tropical storm makes its way toward that state.

"To the people of Texas and Louisiana, we are 100% with you," Trump said from the East Room of the White House.

Supervised content selection

- Given:
 - a labeled training set of good summaries for each document
- Align:
 - the sentences in the document with sentences in the summary
- Extract features
 - position (first sentence?)
 - length of sentence
 - word informativeness
 - cue phrases
- Train
 - a binary classifier
(put sentence in summary? yes or no)
- Problems:
 - alignment difficult → performance not better than unsupervised algorithms
- So in practice:
 - Unsupervised content selection is more common

Evaluating Summaries: ROUGE

ROUGE

(Recall Oriented Understudy for Gisting Evaluation)

Lin and Hovy 2003

- A metric for automatically evaluating summaries
 - Based on BLEU (a metric used for machine translation)
 - Not as good as human evaluation (“Did this answer the user’s question?”)
 - But much more convenient
- 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?

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

A ROUGE example:

Q: "What is water spinach?"

- System output: "Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia."
- Human Summaries (Gold)

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.

- ROUGE-2 =

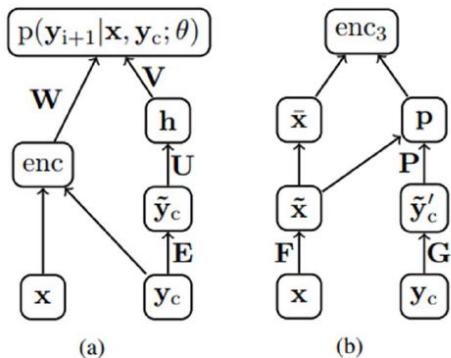
$$\frac{3 + 3 + 6}{10 + 9 + 9} = 12/28 = .43$$

Neural Text Summarization

A neural attention model for abstractive sentence summarization

Rush et al., EMNLP 2015

- Inspired by attention-based seq2seq models (Bahdanau, 2014)



$$\begin{aligned}
 enc_3(\mathbf{x}, \mathbf{y}_c) &= \mathbf{p}^\top \bar{\mathbf{x}}, \\
 \mathbf{p} &\propto \exp(\bar{\mathbf{x}} \mathbf{P} \tilde{\mathbf{y}}'_c), \\
 \bar{\mathbf{x}} &= [\mathbf{F}\mathbf{x}_1, \dots, \mathbf{F}\mathbf{x}_M], \\
 \tilde{\mathbf{y}}'_c &= [\mathbf{G}\mathbf{y}_{i-Q+1}, \dots, \mathbf{G}\mathbf{y}_i], \\
 \forall i \quad \bar{\mathbf{x}}_i &= \sum_{q=i-Q}^{i+Q} \tilde{\mathbf{x}}_q / Q.
 \end{aligned}$$

Figure 5: (a) NNLM decoder with additional encoder element. (b) Attention based encoder.

A neural attention model for abstractive sentence summarization

Rush et al., EMNLP 2015

- Inspired by attention-based seq2seq models (Bahdanau, 2014)

Input $(\mathbf{x}_1, \dots, \mathbf{x}_{18})$. First sentence of article:

russian defense minister ivanov called sunday for the creation of a joint front for combating global terrorism

Output $(\mathbf{y}_1, \dots, \mathbf{y}_8)$. Generated headline:

russia calls for joint front against terrorism \Leftarrow $g(\text{terrorism}, \mathbf{x}, \text{for}, \text{joint}, \text{front}, \text{against})$

Figure 2: Example input sentence and the generated summary. The score of generating \mathbf{y}_{i+1} (terrorism) is based on the context \mathbf{y}_c (for ... against) as well as the input $\mathbf{x}_1 \dots \mathbf{x}_{18}$. Note that the summary generated is abstractive which makes it possible to **generalize** (russian defense minister to russia) and **paraphrase** (for combating to against), in addition to **compressing** (dropping the creation of), see Jing (2002) for a survey of these editing operations.

Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

Nallapati et al., CoNLL 2016

- Implements many tricks (nmt, copy, hierarchical, external knowledge...)

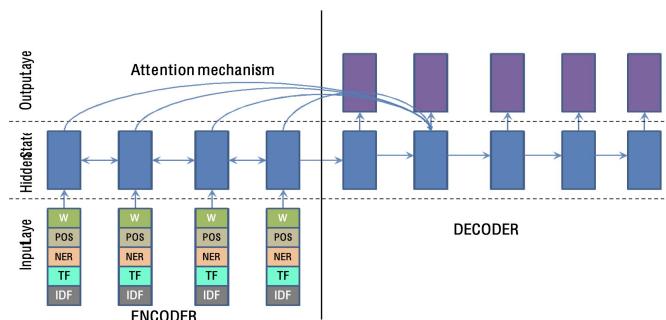


Figure 1: Feature-rich-encoder: We use one embedding

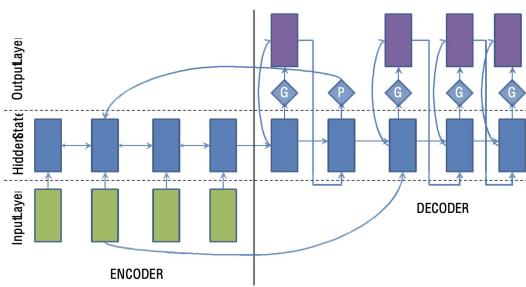


Figure 2: Switching generator/pointer model: When the

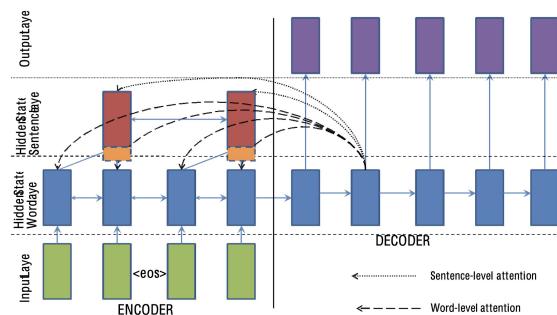


Figure 3: Hierarchical encoder with hierarchical attention:

Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

Nallapati et al., CoNLL 2016

- Implements many tricks (nmt, copy, coverage, hierarchical, external knowledge)
- Hierarchical encoder showed mixed results
- Word-level features slightly helpful
- Copy mechanism leads to the best performance (ptr)

| # | Model name | Rouge-1 | Rouge-2 | Rouge-L | Src. copy rate (%) |
|--|---------------------------------|---------------|--------------|---------------|--------------------|
| Full length F1 on our internal test set | | | | | |
| 1 | words-lvt2k-1sent | 34.97 | 17.17 | 32.70 | 75.85 |
| 2 | words-lvt2k-2sent | 35.73 | 17.38 | 33.25 | 79.54 |
| 3 | words-lvt2k-2sent-hieratt | 36.05 | 18.17 | 33.52 | 78.52 |
| 4 | feats-lvt2k-2sent | 35.90 | 17.57 | 33.38 | 78.92 |
| 5 | feats-lvt2k-2sent-ptr | *36.40 | 17.77 | *33.71 | 78.70 |
| Full length F1 on the test set used by (Rush et al., 2015) | | | | | |
| 6 | ABS+ (Rush et al., 2015) | 29.78 | 11.89 | 26.97 | 91.50 |
| 7 | words-lvt2k-1sent | 32.67 | 15.59 | 30.64 | 74.57 |
| 8 | RAS-Elman (Chopra et al., 2016) | 33.78 | 15.97 | 31.15 | |
| 9 | words-lvt5k-1sent | *35.30 | 16.64 | *32.62 | |

Article: (...) wayne rooney smashes home during manchester united 's 3-1 win over aston villa on saturday. (...)

Summary: manchester united **beat** aston villa 3-1 at old trafford on saturday.

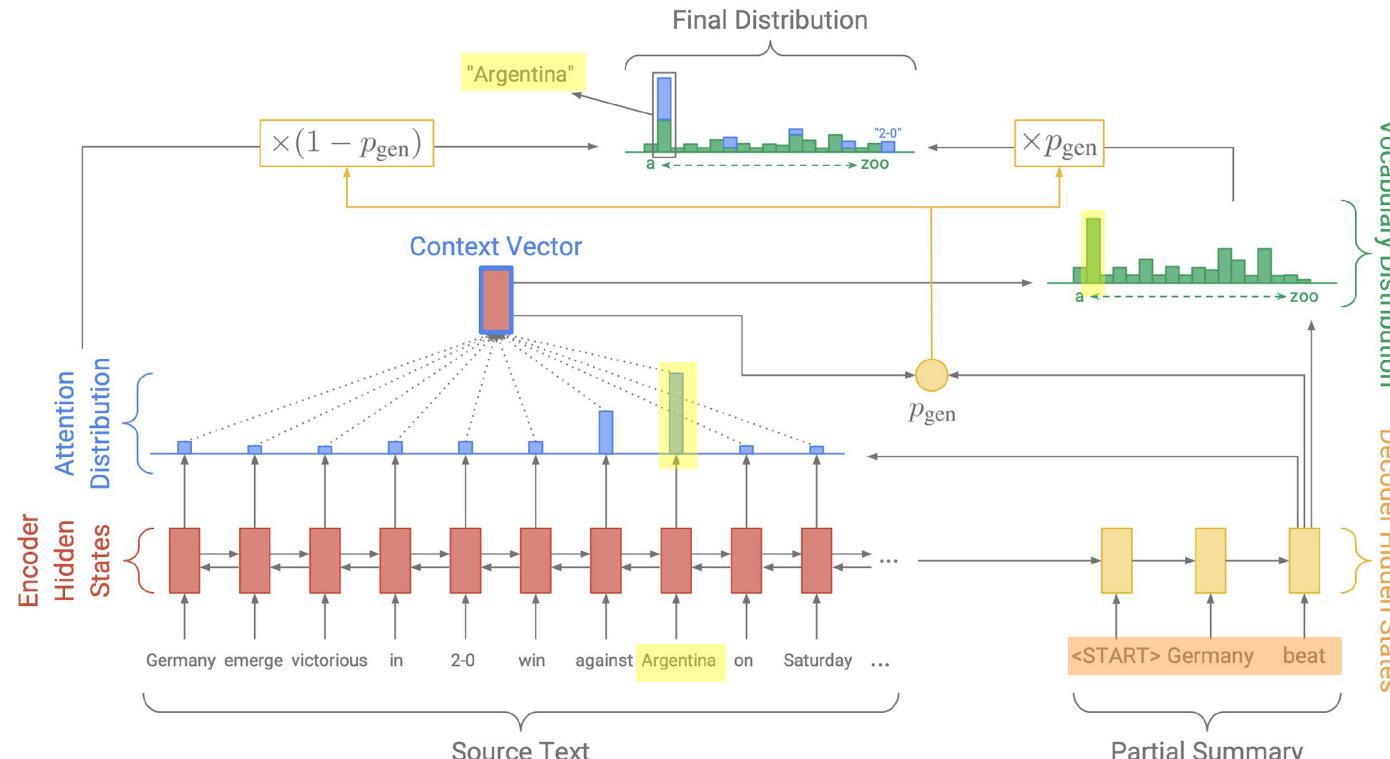
Copy Mechanism

- OOV, Extraction
- "Pointer networks" (Vinyals et al., 2015 NIPS)
- "Pointing the Unknown Words" (Gulcehre et al., ACL 2016)
- " Incorporating Copying Mechanism in Sequence-to-Sequence Learning
" (Gu et al., ACL 2016)
- " Get To The Point: Summarization with Pointer-Generator Networks "
(See et al., ACL 2017)

Pointer Generator Networks

Copy words from the source text

See et al., ACL 2017



Pointer Generator Networks

1. Calculate vocab generation probability

$$P_{\text{vocab}} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b')$$

2. Calculate probability of generation (vs. copy)

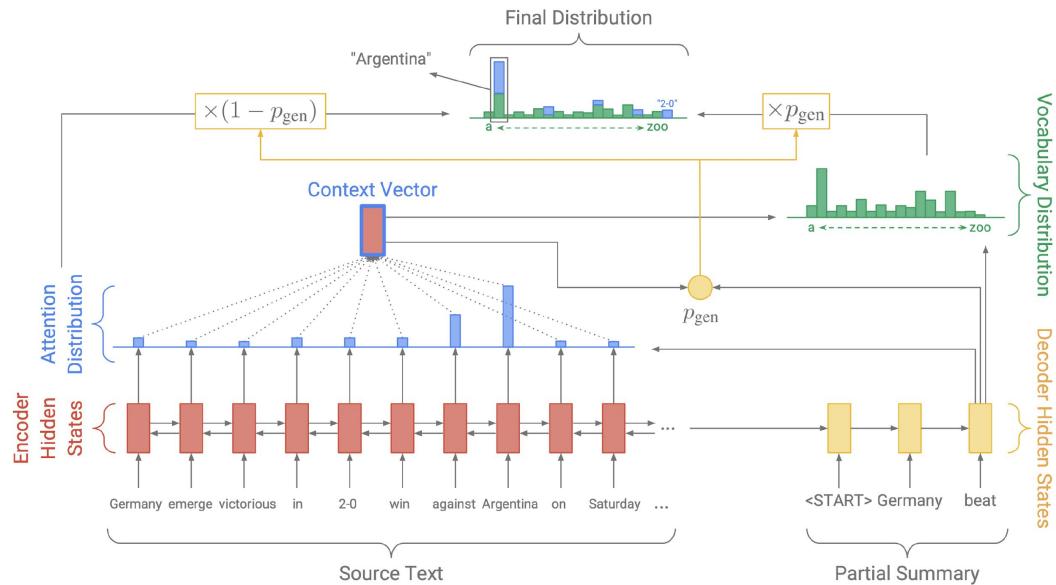
$$p_{\text{gen}} = \sigma(w_h^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}})$$

3. Combine generation and copy probability of words

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

4. Choose the final output

See et al., ACL 2017



Pointer Generator Networks

Article: andy murray (...) is into the semi-finals of the miami open , but not before getting a scare from 21 year-old austrian dominic thiem, who pushed him to 4-4 in the second set before going down 3-6 6-4, 6-1 in an hour and three quarters. (...)

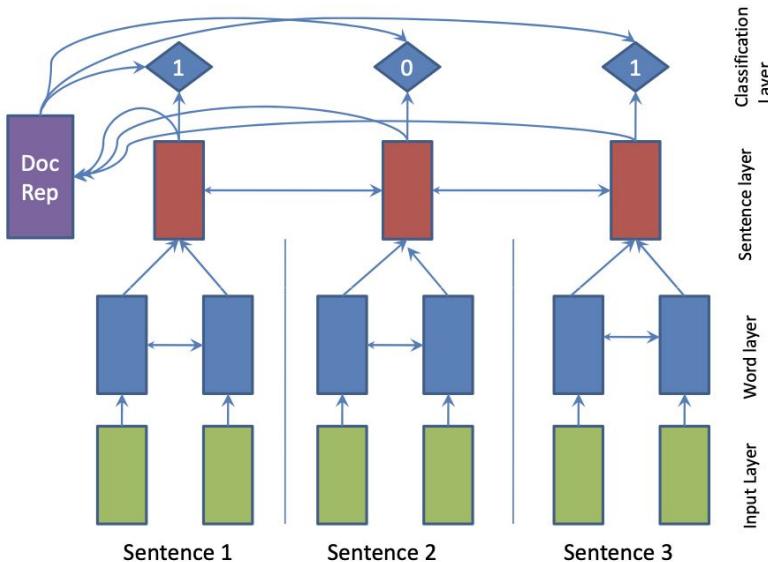
Summary: andy murray **defeated** dominic thiem 3-6 6-4, 6-1 in an hour and three quarters.

Article: (...) wayne rooney smashes home during manchester united 's 3-1 win over aston villa on saturday. (...)

Summary: manchester united **beat** aston villa 3-1 at old trafford on saturday.

Neural Extractive Models

- "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents." (Nallapati et al., AAAI 2017)



Hybrid approach

- "Bottom-Up Abstractive Summarization" (Gehrmann et al., EMNLP 2018)
"a common mistake made by neural copy models is copying very long sequences or even whole sentences."

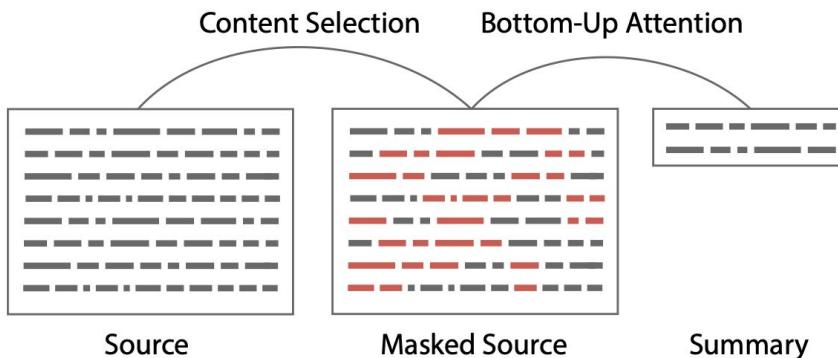
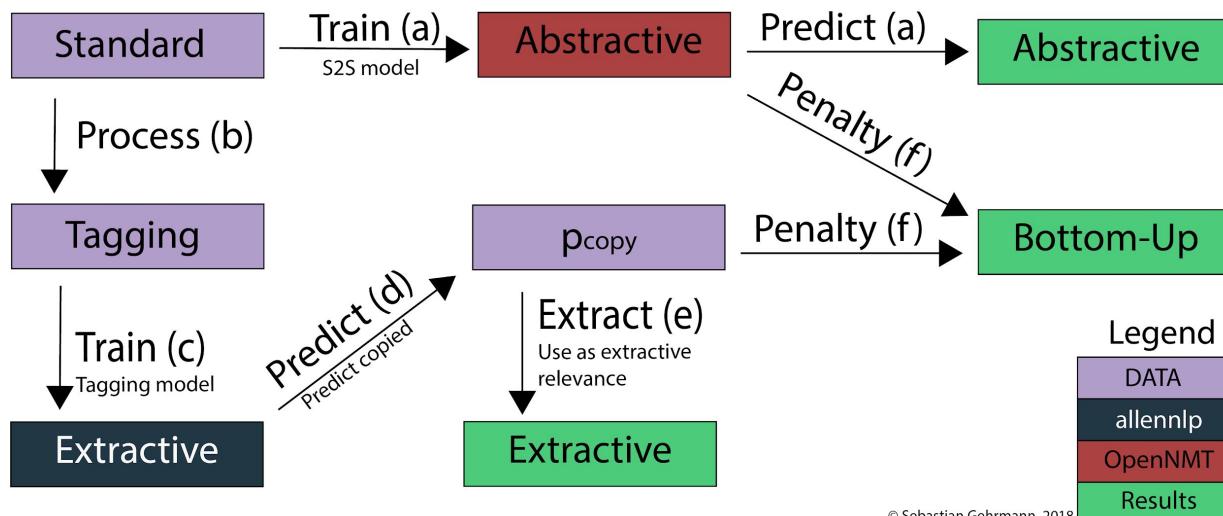


Figure 2: Overview of the selection and generation processes described throughout Section 4.

Hybrid approach

- "Bottom-Up Abstractive Summarization" (Gehrmann et al., EMNLP 2018)

Bottom-Up Attention Models for Extractive Abstractive Summarization



Hybrid approach

- "Bottom-Up Abstractive Summarization" (Gehrmann et al., EMNLP 2018)

| Method | R-1 | R-2 | R-L |
|--|--------------|--------------|--------------|
| ML + RL (Paulus et al., 2017) | 39.87 | 15.82 | 36.90 |
| Saliency + Entailment reward (Pasunuru and Bansal, 2018) | 40.43 | 18.00 | 37.10 |
| Key information guide network (Li et al., 2018a) | 38.95 | 17.12 | 35.68 |
| Inconsistency loss (Hsu et al., 2018) | 40.68 | 17.97 | 37.13 |
| Sentence Rewriting (Chen and Bansal, 2018) | 40.88 | 17.80 | 38.54 |
| Pointer-Generator (our implementation) | 36.25 | 16.17 | 33.41 |
| Pointer-Generator + Coverage Penalty | 39.12 | 17.35 | 36.12 |
| CopyTransformer + Coverage Penalty | 39.25 | 17.54 | 36.45 |
| Pointer-Generator + Mask Only | 37.70 | 15.63 | 35.49 |
| Pointer-Generator + Multi-Task | 37.67 | 15.59 | 35.47 |
| Pointer-Generator + DiffMask | 38.45 | 16.88 | 35.81 |
| Bottom-Up Summarization | 41.22 | 18.68 | 38.34 |
| Bottom-Up Summarization (CopyTransformer) | 40.96 | 18.38 | 38.16 |

Other lines of research

- **Coverage Mechanism**
 - “Modeling Coverage for Neural Machine Translation” (Tu et al., 2016 ACL)
- **Reinforcement Learning**
 - “A deep reinforced model for abstractive summarization.” (Paulus et al., ICLR 2018)
 - “Learning to summarize from human feedback” (Stiennon et al., NIP 2020)
- **Factual Summarization**
 - “Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics” (Pagnoni et al., NAACL 2021)
- **Prompt-based Summarization**
 - “Planning with Learned Entity Prompts for Abstractive Summarization” (Narayan et al., TACL 2021)
 - “Prefix-Tuning: Optimizing Continuous Prompts for Generation” (Li and Liang, ACL 2021)

Conclusion

- **Salience Detection**
 - How to detect important/relevant words or sentences?
- **Remaining Challenges**
 - Long text abstractive summarization
 - Abstractive multi-document summarization