Lecture 12:
Summarization and Text Generation

Ankur Parikh – Google
Thanks to Slav Petrov for some of the slides
Outline

- Summarization

- Case Study on Natural Language Generation [Wiseman et al. 2017]

- Conclusion
He Has Millions and a New Job at Yahoo. Soon, He’ll Be 18.
‘Genres’ of Summary?

- Indicative vs. informative
  ...used for quick categorization vs. content processing.

- Extract vs. abstract
  ...lists fragments of text vs. re-phrases content coherently.

- Generic vs. query-oriented
  ...provides author’s view vs. reflects user’s interest.

- Background vs. just-the-news
  ...assumes reader’s prior knowledge is poor vs. up-to-date.

- Single-document vs. multi-document source
  ...based on one text vs. fuses together many texts.
Cyrus the Great

From Wikipedia, the free encyclopedia

Cyrus II of Persia (Old Persian: مَلک‌زoroآس Kūrūs;[6] New Persian: کوروش Kuruš; Hebrew: כורש, Modern Kūrēš, Tiberian Kūrēš; c. 600 – 530 BC),[6] commonly known as Cyrus the Great[7] and also called Cyrus the Elder by the Greeks, was the founder of the Achaemenid Empire.[8] Under his rule, the empire embraced all the previous civilized states of the ancient Near East,[8] expanded vastly and eventually conquered most of Southwest Asia and much of Central Asia and the Caucasus. From the Mediterranean Sea and Hellespont in the west to the Indus River in the east, Cyrus the Great created the largest empire the world had yet seen.[8] Under his successors, the empire eventually stretched at its maximum extent from parts of the Balkans (Bulgaria-Paeonia and Thrace-Macedonia) and Eastern Europe proper in the west, to the Indus Valley in the east. His regal titles in full were The Great King, King of Persia, King of Anshan, King of Media, King of Babylon, King of Sumer and Akkad, and King of the Four Corners of the World.

The reign of Cyrus the Great lasted c. 30 years. Cyrus built his empire by conquering first the Median Empire, then the Lydian Empire, and eventually the Neo-Babylonian Empire. Either before or after Babylon, he led an expedition into central Asia, which resulted in major campaigns that were described as having brought "into subjection every nation without exception".[10] Cyrus did not venture into Egypt, as he himself died in battle, fighting the Massagetae along the Syr Darya in December 530 BC.[11][12] He was succeeded by his son, Cambyses II, who managed to add to the empire by conquering Egypt, Nubia, and Cyrenaica during his short rule.

Cyrus the Great respected the customs and religions of the lands he conquered.[13] This became a very successful model for centralized administration and establishing a government working to the advantage and profit of its subjects.[8] In fact, the administration of the empire through satraps and the vital principle of forming a government at Pasargadae were the works of Cyrus.[14] What is sometimes referred to as the Edict of Restoration (actually two edicts) described in the Bible as being made by Cyrus the Great left a lasting legacy on the Jewish religion, where, because of his policies in Babylonia, he is referred to by the Jewish Bible as messiah (lit. "His anointed one") (Isaiah 45:1),[15] and is the only non-Jew figure in the Bible to be called so.[16]
However, Doesn’t Always Work

China’s Next Potential Boom Spot: The Places People Overlook

By MICHAEL SCHUMAN  DEC. 1, 2017

LIANGDUO, China — One crisp October morning, Han Youjun got into his silver delivery van and left this small town in eastern China. Within minutes, his van brimming with boxes of every size and shape, he was rumbling through rice paddies, down narrow village lanes and past modest farmhouses, deeper and deeper into China’s vast hinterland.

In the past, delivery drivers like Mr. Han would have had little reason to travel so far. China’s boom over the past four decades made its crowded metropolises wealthy. Much of the rest of the country, especially farming communities like those surrounding Liangduo, in the eastern province of Jiangsu, remained relatively poor.

But more and more, the benefits of China’s economic miracle are penetrating into smaller cities and countryside hamlets — as Mr. Han, a 32-year-old deliveryman for JD.com, an online retailer, knows all too well. The 70 packages crammed into his van that day were double the amount he usually hauled only 18 months earlier.

“The workdays have been getting longer,” he said.

China needs spenders in those places. The government is trying to shift the country’s growth engine away from its traditional dependence on factories and building things. Those old growth sources are no longer dependable and require more and more costly debt.

Thanks to China’s digital revolution, advances in farming and billions of dollars spent on thousands of miles of new highways and railways, Chinese
Figure 5: Counts on a 1000-document sample of how frequently both a document prefix baseline and a ROUGE oracle summary contain sentences at various indices in the document. There is a long tail of useful sentences later in the document, as seen by the fact that the oracle sentence counts drop off relatively slowly. Smart selection of content therefore has room to improve over taking a prefix of the document.
Given many documents about Cyrus the Great, produce a summary.

Cyrus the Great

Cyrus the Great, also called Cyrus II (born 590–580 BCE, Media, or Persia [now in Iran]—died c. 529 BCE, Asia), conqueror who founded the Achaemenid empire, centred on Persia and comprising the Near East from the Aegean Sea eastward to the Indus River. He is also remembered in the Cyrus legend—first recorded by Herodotus—Cyrus’ father Cambyses II was the son of Cambyses I, the king of the Persian kingdom called Anshan. During Cambyses’ reign, he conquered Babylon, and after conquering Elam, he did so to celebrate his victory over the Medes, he founded a government for his new kingdom, incorporating both Medes. He was the son of Cambyses I, the king of the Persian kingdom called Anshan. During Cambyses’ reign, he conquered Babylon, the Persians were vassals of the Median leader Astyages.

Expressions like “king of the Persian kingdom” and “the Median kingdom” are a bit misleading. The Medes and the Persians were coalitions of Iranian nomad tribes; in the fifth century, this was still remembered and the Greek researcher Herodotus of Halicarnassus wrote:

The achievement of Cyrus [...] was to unite under his rule the peoples of Media - Bena, Paratacon, Sardates, Arsacids, Buxi, Magi. The Persian nation contains a number of tribes [...]: the Parthas, Murray, and Mapu, upon which all the other tribes are dependent. Of these, the Persians are the most distinguished; they contain the clan of the Achaemenids from which spring the Persian kings. Other tribes are the Parthian, Derushan, German, all of which are attached to the soil, the remainder - the Dari, Mard, Dacipet, Sipari, being nomadic.
Evaluation

- As with all generation tasks, real evaluation requires a human annotator.

- But need an automatic metric for development
  - One typical choice is ROUGE [Lin, 2004]
ROUGE-1 [Lin, 2004]

- Unigram overlap between generated text and reference.

S(ummary)  

Cyrus defeated the empire of the Greeks

G(old)  

Cyrus the Great founded the Persian empire

Recall = 3 / 7
ROUGE vs BLEU

- ROUGE-1 is a recall focused metric,
  - where the length of the generated text is constrained (typically to the length of the gold summary)

- BLEU is a precision focused metric
  - where there is a brevity penalty to encourage the input to be longer

- Why isn’t ROUGE widely used in machine translation?
ROUGE-N and Rouge-L [Lin, 2004]

- ROUGE-N - Measure N-gram recall (as opposed to just unigrams)

- ROUGE-L - ROUGE based on LCS (longest common subsequence statistics), does not require specifying a predetermined n-gram length
Rouge-L [Lin, 2004]

S  Cyrus defeated the empire of the Greeks

G  Cyrus the Great founded the Persian empire.

Recall  \[ R_{lcs} = \frac{\text{LCS}(S, G)}{|G|} \]

Precision  \[ P_{lcs} = \frac{\text{LCS}(S, G)}{|S|} \]

Rouge-L  \[ F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \]
LCS problem: Given two sequences $A$ and $B$, find the length of longest subsequence present in both of them.

$X$ if and only if a subsequence of $A$ if and only if all tokens in $X$ appear in the same order in $X$ as they do in $A$ (They do not need to be contiguous)

$A$  Cyrus the Great founded the Persian empire

Some subsequences of $A$

Cyrus the Great
Cyrus the Great founded
Cyrus founded
Great founded Persian
Shortcomings of ROUGE

- ROUGE-1/2 contain practically no measure of coherence (ROUGE-L does to some extent)

According to Herodotus, Cyrus died fighting the Massagetae and his son Cambyses assumed the throne. Cyrus the Great founded the Persian empire and under his rule it became the largest empire the world had seen.

Cyrus the Great founded the Persian empire and under his rule it became the largest empire the world had seen. According to Herodotus, Cyrus died fighting the Massagetae and his son Cambyses assumed the throne.
Shortcomings of ROUGE

- ROUGE contains practically no measure of coherence.

According to Herodotus, Cyrus died fighting the Massagetae and his son Cambyses assumed the throne. Cyrus the Great founded the Persian empire and under his rule it became the largest empire the world had seen.

Cyrus the Great founded the Persian empire and under his rule it became the largest empire the world had seen. According to Herodotus, Cyrus died fighting the Massagetae and his son Cambyses assumed the throne.

more coherent but ROUGE score probably won’t change
Shortcomings of ROUGE

- ROUGE doesn’t really capture precision of facts.

According to Cambyses, Cyrus died fighting the Greeks and his son Herodotus assumed the throne.

According to Herodotus, Cyrus died fighting the Massagetae and his son Cambyses assumed the throne.

ROUGE-1 will be unaffected, ROUGE-2 will only penalize a few bigrams.
Challenges in Summarization

- Summarization is an incredibly challenging problem in NLP.
- First X sentences baseline gets most of the easy cases.
- Evaluation can be more subjective than other tasks.
  - (Even task itself is more subjective)
- Automatic metrics (e.g. ROUGE) are not that great.
One way to approach sentence summarization is simply to pick a subset of the sentences.

Generally involves trading off between two criteria (subject to a length constraint):

- **Relevance** (Selected sentences should be important)
- **Diversity** (Selected sentences should be different from each other)
Maximum Marginal Relevance
[Carbonell and Goldstein 1998]

- Greedily pick sentences based on the following MMR objective:

\[
D = \text{document} \\
\ s_i = \text{sentence in } D \\
A = \text{generated summary}
\]

\[
\text{MMR} = \arg \max_{s_i \in D \setminus A} \left( \lambda \times \text{Sim}_1(s_i, q) - (1 - \lambda) \times \max_{s_j \in A} \text{Sim}_2(s_i, s_j) \right)
\]

query-focused setting: retrieve sentences that are relevant to a query \( q \)
Graph-Based Approach: TextRank

[Mihalcea and Tarau, 2005]

- Construct a graph where vertices are sentences and edges are similarities

- Compute a score for each vertex using PageRank

\[
\text{score}(s_i) = (1 - d) + d \times \sum_{j \in \text{In}(s_i)} \frac{1}{|\text{Out}(s_j)|} \text{score}(s_j)
\]

- Take top vertices (sentences) as summary.
Graph-Based Approach: TextRank
[Mihalcea and Tarau, 2005]

3: BC–HurricaneGilbert, 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.

TextRank extractive summary
Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil De-
fense alerted its heavily populated south coast to prepare for high winds, heavy rains
and high seas. 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. 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. Strong
winds associated with Gilbert brought coastal flooding, strong southeast winds and up
to 12 feet to Puerto Rico’s south coast.

Manual abstract I
Hurricane Gilbert is moving toward the Dominican Republic, where the residents of
the south coast, especially the Barahona Province, have been alerted to prepare for
heavy rains and high wind and seas. Tropical storm Gilbert formed in the eastern
Caribbean and became a hurricane on Saturday night. By 2 a.m. Sunday it was about
200 miles southeast of Santo Domingo and moving westward at 15 mph with winds
of 75 mph. Flooding is expected in Puerto Rico and in the Virgin Islands. The second
hurricane of the season, Florence, is now over the southern United States and down-
graded to a tropical storm.

Manual abstract II
Tropical storm Gilbert in the eastern Caribbean strengthened into a hurricane Saturday
night. The National Hurricane Center in Miami reported its position at 2 a.m. Sunday to
be about 140 miles south of Puerto Rico and 200 miles southeast of Santo Domingo. It
is moving westward at 15 mph with a broad area of cloudiness and heavy weather with
sustained winds of 75 mph gusting to 92 mph. The Dominican Republic’s Civil Defense
alerted that country’s heavily populated south coast and the National Weather Service
in San Juan, Puerto Rico issued a flash watch for Puerto Rico and the Virgin Islands until
at least 6 p.m. Sunday.
How far can we get with word frequency alone?

Motivation:

Table 1: Percentage of the top $n$ frequency words from the input documents that appear in the four human models and in a state-of-the-art automatic summarizer (average across 30 input sets)
Using Word Frequency Only
SumBasic [Nenkova and Vanderwende, 2005]

\( w_i = \text{word} \)

\( s_i = \text{sentence} \)

1. \( \text{Value}(w_i) = P_D(w_i) \)
2. \( \text{Value}(s_i) = \text{sum of its word values} \)
3. Choose \( s_i \) with largest value that also contains highest scoring word
4. Adjust \( P_D(w) \)
5. Repeat until length constraint
Using Word Frequency Only
SumBasic [Nenkova and Vanderwende, 2005]

\[ w_i = \text{word} \]
\[ s_i = \text{sentence} \]

(1) Value\( (w_i) = P_D(w_i) \)
(2) Value\( (s_i) = \text{sum of its word values} \)
(3) Choose \( s_i \) with largest value that also contains highest scoring word
(4) Adjust \( P_D(w) \)
(5) Repeat until length constraint

\( P_D(w_i) \) is initialized to unigram frequency in source document

In step 4: for all \( w_i \) in \( s_i \)

\[ P_D(w_i)^{\text{new}} = P_D(w_i)^{\text{old}} \times P_D(w_i)^{\text{old}} \]
Using Word Frequency Only
SumBasic [Nenkova and Vanderwende, 2005]

(1) Value\(w_i\) = \(P_D(w_i)\)
(2) Value\(s_i\) = sum of its word values
(3) Choose \(s_i\) with largest value that also contains highest scoring word
(4) Adjust \(P_D(w)\)
(5) Repeat until length constraint

relevance

diversity
Using Word Frequency Only
SumBasic [Nenkova and Vanderwende, 2005]

**SumBasic SUMMARY**
Former Chilean dictator Gen. Augusto Pinochet has been arrested by British police on a Spanish extradition warrant, despite protests from Chile that he is entitled to diplomatic immunity. Human rights and international law experts expressed enthusiastic support Saturday for the British arrest, and said it could have wide implications. Both President Eduardo Frei of Chile and President Eduardo Menem of Argentina have resisted the Spanish legal motions, arguing that they infringe on their nations’ sovereignty. Baltasar Garzon, one of two Spanish magistrates handling probes into human rights violations in Chile and Argentina, filed a request to question Pinochet on Wednesday.

**HUMAN SUMMARY**
Augusto Pinochet, former Chilean dictator, was arrested in London on 14 October 1998. Pinochet, 82, was recovering from surgery. The arrest was in response to an extradition warrant served by Baltasar Garzon, a maverick Spanish judge. Pinochet was charged with murdering thousands, including many Spaniards. Pinochet is awaiting a hearing, his fate in the balance. Chile protested, insisting that Pinochet had diplomatic immunity. Britain disagreed. The reaction of the Chilean people was mixed. American scholars applauded the arrest, saying that it set a precedent for other terrorist dictators. Castro criticized the arrest, and called it unprecedented international meddling.

Figure 1: Summaries produced for the same input by SumBasic and by a human.
Global Inference for Document Summarization

[McDonald 2007]

Instead of being greedy, try to solve global (NP-hard) objective.

\[ D = \{t_1, \ldots, t_n\} \]

\[ \hat{A} = \arg \max_{A \subseteq D} \sum_{t_i \in A} \text{Rel}(t_i) - \sum_{t_i, t_j \in A, i < j} \text{Red}(t_i, t_j) \]

such that \( \sum_{t_i \in A} \text{len}(t_i) \leq K \)
Global Inference for Document Summarization

[McDonald 2007]

\[ D = \{ t_1, \ldots, t_n \} \]

\[ \hat{A} = \arg\max_{A \subseteq D} \sum_{t_i \in A} \text{Rel}(t_i) - \sum_{t_i, t_j \in A, i<j} \text{Red}(t_i, t_j) \]

such that \[ \sum_{t_i \in A} \text{len}(t_i) \leq K \]
Can reformulate as integer linear program:

$$\text{maximize}_\alpha \sum_i \alpha_i \text{Rel}(t_i) - \sum_{i<j} \alpha_{ij} \text{Red}(t_i, t_j)$$

such that

1. $\alpha_i, \alpha_j \in \{0, 1\}$
2. $\sum_i \alpha_i \text{len}(t_i) \leq K$
3. $\alpha_{ij} - \alpha_i \leq 0$
4. $\alpha_{ij} - \alpha_j \leq 0$
5. $\alpha_i + \alpha_j - \alpha_{ij} \leq 1$

In his experiments, textual sentences are chosen to be sentences. Note that number of variables is quadratic in the number of sentences!
Global Inference for Document Summarization
[McDonald 2007]

Table 1. (a) Results for generic summarization experiments using DUC 2002 data set. Each cell contains the ROUGE-1 and 2 scores (R1 / R2). (b) Results for query-focused summarization experiments using DUC 2005 data set. Original: Using original inference formulation from Equation 1. Alternate: Using alternate inference formulation from Section 3.4.
A Scalable Global Model for Summarization
[Gillick and Favre 2009]

- Aims to solve two problems:
  - Reduce the number of variables in the ILP so that it can be solved efficiently.
  - Model redundancy more globally than via pairwise constraints.

- Consider following example:

  (1) The cat is in the kitchen.
  (2) The cat drinks the milk.
  (3) The cat drinks the milk in the kitchen.

Figure 1: Example of sentences redundant as a group. Their redundancy is only partially captured by sentence-level pairwise measurement.
A Scalable Global Model for Summarization
[ Gillick and Favre 2009]

Define sub-sentence units called *concepts* that are associated with *weights*

| S1   | The health care bill is a major test for the Obama administration. |
| S2   | Universal health care is a divisive issue.                        |
| S3   | President Obama remained calm.                                   |
| S4   | Obama addressed the House on Tuesday.                            |

<table>
<thead>
<tr>
<th>concept</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Concepts** could be *n*-grams, named entities, syntactic subtrees etc. Gillick and Favre choose bigrams.

**Weights** are fixed (not learned). Gillick and Favre choose frequency count.
A Scalable Global Model for Summarization
[Gillick and Favre 2009]

Define sub-sentence units called *concepts* that are associated with *weights*

<table>
<thead>
<tr>
<th>S1</th>
<th>The health care bill is a major test for the Obama administration.</th>
<th>concept</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2</td>
<td>Universal health care is a divisive issue.</td>
<td>Obama</td>
<td>3</td>
</tr>
<tr>
<td>S3</td>
<td>President Obama remained calm.</td>
<td>health care</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>Obama addressed the House on Tuesday.</td>
<td>House</td>
<td></td>
</tr>
</tbody>
</table>

*Obama* is highlighted in each sentence.
A Scalable Global Model for Summarization
[ Gillick and Favre 2009]

Define sub-sentence units called *concepts* that are associated with *weights*

- **S1**: The [health care] bill is a major test for the Obama administration.
- **S2**: Universal [health care] is a divisive issue.
- **S3**: President Obama remained calm.
- **S4**: Obama addressed the House on Tuesday.

<table>
<thead>
<tr>
<th>concept</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obama</td>
<td>3</td>
</tr>
<tr>
<td>health care</td>
<td>2</td>
</tr>
<tr>
<td>House</td>
<td></td>
</tr>
</tbody>
</table>
Define sub-sentence units called *concepts* that are associated with *weights*.

- **S1**: The health care bill is a major test for the Obama administration.
- **S2**: Universal health care is a divisive issue.
- **S3**: President Obama remained calm.
- **S4**: Obama addressed the House on Tuesday.

<table>
<thead>
<tr>
<th>concept</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obama</td>
<td>3</td>
</tr>
<tr>
<td>health care</td>
<td>2</td>
</tr>
<tr>
<td>House</td>
<td>1</td>
</tr>
</tbody>
</table>
A Scalable Global Model for Summarization
[ Gillick and Favre 2009]

c_i = concept is in summary
s_j = sentence is in summary

\[
\text{maximize}_{c_i, s_i} \sum_i w_i c_i
\]
such that

1. \( c_i \in \{0, 1\} \ \forall i \)
2. \( s_j \in \{0, 1\} \ \forall j \)
3. \( s_j \mathbb{I}[c_i \in s_j] \leq c_j \ \forall i, j \)
4. \( \sum_j s_j \mathbb{I}[c_i \in s_j] \geq c_i \)
A Scalable Global Model for Summarization
[ Gillick and Favre 2009]

<table>
<thead>
<tr>
<th>System</th>
<th>ROUGE-2</th>
<th>Pyramid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.058</td>
<td>0.186</td>
</tr>
<tr>
<td>McDonald</td>
<td>0.072</td>
<td>0.295</td>
</tr>
<tr>
<td>Concepts</td>
<td>0.110</td>
<td>0.345</td>
</tr>
</tbody>
</table>

Table 2: Scores for both systems and a baseline on TAC 2008 data (Set A) for ROUGE-2 and Pyramid evaluations.

Figure 2: A comparison of ILP run-times (on an AMD 1.8Ghz desktop machine) of McDonald’s sentence-based formulation and our concept-based formulation with an increasing number of input sentences.
Results [Gillick and Favre, 2009]

- Overall Quality
  - Rating scale: 1-10
  - Humans in [8.3, 9.3]

- Pyramid
  - Rating scale: 0-1
  - Humans in [0.62, 0.77]

- Linguistic Quality
  - Rating scale: 1-10
  - Humans in [8.5, 9.3]

- ROUGE-2
  - Rating scale: 0-1
  - Humans in [0.11, 0.15]

52 submissions
27 teams
44 topics
- 10 input docs
- 100 word summaries

Gillick & Favre
Submodularity in Summarization
[ Lin and Bilmes 2011]

- A really cool machine learning paradigm that very naturally fits summarization.

- Idea is to define an objective that is monotone submodular.

- It is NP-hard to find the globally optimal solution, but a greedy approximation is guaranteed to give a constant factor approximation

\[
\mathcal{F}(\hat{A}) \geq (1 - 1/e) \mathcal{F}(A_{opt}) \approx 0.632 \mathcal{F}(A_{opt})
\]
Submodularity

Property of diminishing returns.

Let $\Omega$ be a set and $\mathcal{F}$ be a function that maps subsets $X \subseteq \Omega$ to real values in $\mathbb{R}$. Then $\mathcal{F}$ is submodular if for every set $X, Y \subseteq \Omega$ with $X \subseteq Y \subseteq \Omega \setminus v$

$$\mathcal{F}(X + v) - \mathcal{F}(X) \geq \mathcal{F}(Y + v) - \mathcal{F}(Y)$$

$v$ relatively adds more to $X$ than it does to $Y$
Monotone Submodularity

Furthermore, \( \mathcal{F} \) is monotone submodular if for all \( X \subseteq Y \subseteq \Omega \), \( \mathcal{F}(Y) \geq \mathcal{F}(X) \)

adding more sentences to the summary shouldn’t decrease the objective :)

MMR is not Monotone Submodular

$$\text{MMR} = \arg\max_{s_i \in D \setminus A} \left( \lambda \times \text{Sim}_1(s_i, q) - (1 - \lambda) \times \max_{s_j \in A} \text{Sim}_2(s_i, s_j) \right)$$

- MMR is submodular, but it is not monotone.
- Why?
Submodularity in Summarization
[ Lin and Bilmes 2011]

- Instead of penalizing redundancy, reward diversity

\[
\mathcal{F}(A) := L(A) + \lambda R(A)
\]

relevance \hspace{2cm} diversity
Assume sentences are clustered into $K$ clusters: $C_1, \ldots, C_k$

$$F(A) := L(A) + \lambda R(A)$$

$$R(A) = \sum_{k=1}^{K} \left( \sum_{j \in C_k \cap A} \sqrt{\sum_{s_j}} \right)$$

sqrt function lowers reward for picking too many sentences from the same cluster

relevance score for sentence $j$

$A =$ generated summary
Table 1: ROUGE-1 recall (R) and F-measure (F) results (%) on DUC-04. DUC-03 was used as development set.

<table>
<thead>
<tr>
<th>DUC-04</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sum_{i \in V} \sum_{j \in S} w_{i,j})</td>
<td>33.59</td>
<td>32.44</td>
</tr>
<tr>
<td>(L_1(S))</td>
<td>39.03</td>
<td>38.65</td>
</tr>
<tr>
<td>(R_1(S))</td>
<td>38.23</td>
<td>37.81</td>
</tr>
<tr>
<td>(L_1(S) + \lambda R_1(S))</td>
<td><strong>39.35</strong></td>
<td><strong>38.90</strong></td>
</tr>
<tr>
<td>Takamura and Okumura (2009)</td>
<td>38.50</td>
<td>-</td>
</tr>
<tr>
<td>Wang et al. (2009)</td>
<td>39.07</td>
<td>-</td>
</tr>
<tr>
<td>Lin and Bilmes (2010)</td>
<td>-</td>
<td>38.39</td>
</tr>
<tr>
<td>Best system in DUC-04 (peer 65)</td>
<td>38.28</td>
<td>37.94</td>
</tr>
</tbody>
</table>
Abstractive Summarization with Neural Nets

- Instead of selecting sentences, generate the summary with a neural net.

- Can use Encoder/Decoder models e.g.
  - Rush et al. 2015
  - Paulus et al. 2017
Abstractive Summarization with Neural Nets

Figure 1: Example output of the attention-based summarization (ABS) system. The heatmap represents a soft alignment between the input (right) and the generated summary (top). The columns represent the distribution over the input after generating each word.
Copying Mechanisms [Gu et al. 2016, See et al. 2017]

- Allow the neural network the option of directly copying words from the source text.

Figure 2: Baseline sequence-to-sequence model with attention. The model may attend to relevant words in the source text to generate novel words, e.g., to produce the novel word *beat* in the abstractive summary *Germany beat Argentina 2-0* the model may attend to the words *victorious* and *win* in the source text.
Copy the Mechanisms [Gu et al. 2016, See et al. 2017]

- Allow the neural network the option of directly copying words from the source text.
The 120th step. The search algorithm is self-stopping and almost never reaches the article can ing and testing, but we also found that truncating kens at test time. Summary to 100 tokens for training and 120 to article to 400 tokens and limit the length of the implement early stopping.

We use loss on the validation set to gradient norm of 2, but do not use any form of regularization. We use SProp). We use gradient clipping with a maximum Descent, Adadelta, Momentum, Adam and RM was found to work best of Stochastic Gradient train the word embeddings – they are learned average adds 512 extra parameters (parameters, the pointer-generator adds 1153 extra parameters to the network: for the models with vocabulary size of 150k. For the baseline model, we also try a larger vocab size 2016, See et al. 2017). Duchi et al. 2011, w b, and w x h ⇤, w s, w x 8, and b 8 in equation 15), we do not pre- in equation 2016, w s, w x See et al. 2017, w s, w x Nallapati et al. 2017, w s, w x Nallapati et al. 2011, w s, w x Nallapati et al. 2016, w s, w x 3, and w x 3; this re- l = 13, and l = 13.

<table>
<thead>
<tr>
<th></th>
<th>ROUGE</th>
<th>METEOR</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>L</td>
<td>exact match</td>
</tr>
<tr>
<td>abstractive model</td>
<td>35.46</td>
<td>13.30</td>
<td>32.65</td>
<td>-</td>
</tr>
<tr>
<td>(Nallapati et al., 2016)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq-to-seq + attn baseline (150k vocab)</td>
<td>30.49</td>
<td>11.17</td>
<td>28.08</td>
<td>11.65</td>
</tr>
<tr>
<td>seq-to-seq + attn baseline (50k vocab)</td>
<td>31.33</td>
<td>11.81</td>
<td>28.83</td>
<td>12.03</td>
</tr>
<tr>
<td>pointer-generator</td>
<td>36.44</td>
<td>15.66</td>
<td>33.42</td>
<td>15.35</td>
</tr>
<tr>
<td>pointer-generator + coverage</td>
<td>39.53</td>
<td>17.28</td>
<td><strong>36.38</strong></td>
<td>17.32</td>
</tr>
<tr>
<td>lead-3 baseline (ours)</td>
<td>40.34</td>
<td>17.70</td>
<td>36.57</td>
<td>20.48</td>
</tr>
<tr>
<td>lead-3 baseline</td>
<td>39.2</td>
<td>15.7</td>
<td>35.5</td>
<td>-</td>
</tr>
<tr>
<td>(Nallapati et al., 2017)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>extractive model</td>
<td>39.6</td>
<td>16.2</td>
<td>35.3</td>
<td>-</td>
</tr>
<tr>
<td>(Nallapati et al., 2017)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: ROUGE F₁ and METEOR scores on the test set. Models and baselines in the top half are abstractive, while those in the bottom half are extractive. Those marked with * were trained and evaluated on the anonymized dataset, and so are not strictly comparable to our results on the original text. All our ROUGE scores have a 95% confidence interval of at most ±0.25 as reported by the official ROUGE script. The METEOR improvement from the 50k baseline to the pointer-generator model, and from the pointer-generator to the pointer-generator+coverage model, were both found to be statistically significant using an approximate randomization test with \( p < 0.01 \).
End to end neural systems are good at fluency, poor at content selection.

Two step process:
- Select certain words in the source.
- Limit the copy mechanism to only copy those words.
Bottom Up Abstractive Summarization
[Gehrmann et al. 2018]

- End to end neural systems are good at fluency, poor at content selection.

- Two step process:
  - Select certain words in the source.
  - Limit the copy mechanism to only copy those words.

Figure 2: Overview of the selection and generation processes described throughout Section 4.
### Bottom Up Abstractive Summarization

[Gebrmann et al. 2018]

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pointer-Generator (See et al., 2017)</td>
<td>36.44</td>
<td>15.66</td>
<td>33.42</td>
</tr>
<tr>
<td>Pointer-Generator + Coverage (See et al., 2017)</td>
<td>39.53</td>
<td>17.28</td>
<td>36.38</td>
</tr>
<tr>
<td>ML + Intra-Attention (Paulus et al., 2017)</td>
<td>38.30</td>
<td>14.81</td>
<td>35.49</td>
</tr>
<tr>
<td>ML + RL (Paulus et al., 2017)</td>
<td>39.87</td>
<td>15.82</td>
<td>36.90</td>
</tr>
<tr>
<td>Saliency + Entailment reward (Pasunuru and Bansal, 2018)</td>
<td>40.43</td>
<td>18.00</td>
<td>37.10</td>
</tr>
<tr>
<td>Key information guide network (Li et al., 2018a)</td>
<td>38.95</td>
<td>17.12</td>
<td>35.68</td>
</tr>
<tr>
<td>Inconsistency loss (Hsu et al., 2018)</td>
<td>40.68</td>
<td>17.97</td>
<td>37.13</td>
</tr>
<tr>
<td>Sentence Rewriting (Chen and Bansal, 2018)</td>
<td>40.88</td>
<td>17.80</td>
<td><strong>38.54</strong></td>
</tr>
<tr>
<td>Pointer-Generator (our implementation)</td>
<td>36.25</td>
<td>16.17</td>
<td>33.41</td>
</tr>
<tr>
<td>Pointer-Generator + Coverage Penalty</td>
<td>39.12</td>
<td>17.35</td>
<td>36.12</td>
</tr>
<tr>
<td>CopyTransformer + Coverage Penalty</td>
<td>39.25</td>
<td>17.54</td>
<td>36.45</td>
</tr>
<tr>
<td>Pointer-Generator + Mask Only</td>
<td>37.70</td>
<td>15.63</td>
<td>35.49</td>
</tr>
<tr>
<td>Pointer-Generator + Multi-Task</td>
<td>37.67</td>
<td>15.59</td>
<td>35.47</td>
</tr>
<tr>
<td>Pointer-Generator + DiffMask</td>
<td>38.45</td>
<td>16.88</td>
<td>35.81</td>
</tr>
<tr>
<td>Bottom-Up Summarization</td>
<td><strong>41.22</strong></td>
<td><strong>18.68</strong></td>
<td>38.34</td>
</tr>
<tr>
<td>Bottom-Up Summarization (CopyTransformer)</td>
<td>40.96</td>
<td>18.38</td>
<td>38.16</td>
</tr>
</tbody>
</table>

Table 1: Results of abstractive summarizers on the CNN-DM dataset.² The first section shows encoder-decoder abstractive baselines trained with cross-entropy. The second section describes reinforcement-learning based approaches. The third section presents our baselines and the attention masking methods described in this work.
Outline

- Summarization

- Case Study on Natural Language Generation [Wiseman et al. 2017]

- Conclusion