EVALUATION AND THEORY

SEMINAR NEURAL TEXT SUMMARIZATION

JULIUS STEEN
(WITH SOME SLIDES FROM KATJA MARKERT)

30TH OCTOBER 2019
1 Organisation

2 ROUGE

3 ROUGE and Summary Length

4 Corpora

5 A theoretic approach to summarization
Organisation
Please prefix mails with \textit{NTS} from now on.

- Do not forget to hand in questions before the seminar.
- Paper assignment by tomorrow.
ROUGE
Defining what is a good summary is difficult

Intuitively summaries should
► Cover important material
► Not be redundant
► Be readable

Usually: Comparison to one or more human reference summaries
- ROUGE: Recall-Oriented Understudy for Gisting Evaluation [Lin, 2004]
- Currently most widespread evaluation measure for summarization
- Based on overlap between reference summaries $S_{ref}$ and the system summary $s_{sys}$
ROUGE-N

- ROUGE-N: Ngram-based metric
- Originally only recall:

\[ \text{ROUGE}_n^{(\text{Rec})} = \frac{\sum_{s_{\text{ref}} \in S_{\text{ref}}} \sum_{g \in \text{grams}(s_{\text{ref}}, n)} \text{count}_{s_{\text{ref}}}(s_{\text{sys}}, g)}{\sum_{s_{\text{ref}} \in S_{\text{ref}}} \sum_{g \in \text{grams}(s_{\text{ref}}, n)} \text{count}(s_{\text{ref}}, g)} \]

- Later also precision:

\[ \text{ROUGE}_n^{(\text{Prec})} = \frac{1}{|S_{\text{ref}}|} \sum_{s_{\text{ref}} \in S_{\text{ref}}} \frac{\sum_{g \in \text{grams}(s_{\text{ref}}, n)} \text{count}_{s_{\text{ref}}}(s_{\text{sys}}, g)}{\sum_{g \in \text{grams}(s_{\text{sys}}, n)} \text{count}(s_{\text{sys}}, g)} \]

- We can compute ROUGE-F1 accordingly
ROUGE-L

- ROUGE-N does not handle non-consecutive matches
- ROUGE-L tries to improve this by computing on longest common subsequences of input
- The following works for single sentence summaries
- Recall:

\[
\text{ROUGE}_{\text{LCS}}^{(\text{Rec})} = \frac{\text{LCS}(s_{\text{ref}}, s_{\text{sys}})}{|s_{\text{ref}}|}
\]

- Precision:

\[
\text{ROUGE}_{\text{LCS}}^{(\text{Prec})} = \frac{\text{LCS}(s_{\text{ref}}, s_{\text{sys}})}{|s_{\text{sys}}|}
\]

- F-Score:

\[
\text{ROUGE}_{\text{LCS}}^{(F)} = \frac{(1 + \beta^2) \text{ROUGE}_{\text{LCS}}^{(\text{Prec})} \text{ROUGE}_{\text{LCS}}^{(\text{Rec})}}{\text{ROUGE}_{\text{LCS}}^{(\text{Rec})} + \beta^2 \text{ROUGE}_{\text{LCS}}^{(\text{Prec})}}
\]

- DUC only considers recall
Previous definition works for a single sentence reference and summary

For multiple references compute union LCS for each sentence

▶ Which proportion of reference sentence \( r_i \) is covered by subsequences of system sentences?

Modified recall:

\[
\text{ROUGE}_{LCS}^{(Rec)} = \frac{\sum_{r_i \in S_{ref}} \text{LCS} \cup (r_i, S_{sys})}{|S_{ref}|}
\]
ROUGE-W

- ROUGE-L does not consider distance between correct tokens.
- Given reference \( w_1, w_2, w_3, w_4, w_5 \) the sequence \( w_1, w_2, w_3, w_7, w_8 \) is intuitively better than \( w_1, w_7, w_2, w_8, w_3 \), but both receive same score.
- ROUGE-W penalizes long gaps in sequence by giving more weight to long matches.
- Weighting function: \( f(k) = k^\alpha \).
- Recall:

\[
ROUGE_{WLCS}^{(Rec)} = f^{-1} \left( \frac{WLCS(s_{sys}, s_{ref})}{f(|s_{ref}|)} \right)
\]
**ROUGE-W: Example**

Given $\alpha = 2$

Weights:

<table>
<thead>
<tr>
<th>R/S</th>
<th>#</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

Lengths:

<table>
<thead>
<tr>
<th>R/S</th>
<th>#</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$$\text{ROUGE}_{\text{WLCS}}^{(Rec)} = \sqrt{\frac{16}{7^2}} = 0.57$$
- ROUGE based on Skip-Bigrams
- Skip-Bigram: Two words in the correct order in a sequence
- Recall:
  \[
  \text{ROUGE}^{(\text{Rec})}_S = \frac{\text{skip}_2(s_{\text{ref}}, s_{\text{sys}})}{C(|s_{\text{ref}}|, 2)}
  \]
- Precision:
  \[
  \text{ROUGE}^{(\text{Prec})}_S = \frac{\text{skip}_2(s_{\text{ref}}, s_{\text{sys}})}{C(|s_{\text{sys}}|, 2)}
  \]
- F-Score accordingly
- ROUGE-S is overly punitive for out-of-order matches
- Add unigrams to skip-gram set
A SMALL EXERCISE

Reference 1: The girl liked the dog.

Reference 2: The tall girl liked the beautiful dog.

Summary: The girl that the beautiful dog liked.

Compute ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-S2
ROUGE AND SUMMARY LENGTH
Traditional summary tasks had length constraint

Current setup: Either no explicit length constraint, or constraint in number of sentences.

This makes ROUGE-recall is very easy to "game"

Current practice: Compute ROUGE-F1 score [Nallapati et al., 2016]
[Sun et al., 2019] observe that ROUGE-F1 is not appropriate for comparing summaries with different length.

- ROUGE-scores of systems vary with summary length.
- Longer summaries are not penalized by ROUGE.
Comparison of four summarization strategies

- **Lead:** Take sentences from the beginning of the sentence up to length limit
- **Random:** Randomly select sentences up to length limit
- **TextRank:** Use graph centrality measure to score sentences
- **Pointer Generator:** Neural system, see session in two weeks
**Figure:** ROUGE-scores over different summary lengths (source: [Sun et al., 2019])
[Sun et al., 2019] propose normalizing ROUGE score
- Divide ROUGE by score of the random summarizer at summary length
- Intuition: The shorter the random summary, the easier to improve over its scores.
- [Sun et al., 2019] conduct evaluation with human annotators
- Generally, humans seem to prefer longer summaries
- Possible related to less content being cut from the summary
CORPORATA
- Originally constructed for QA [Hermann et al., 2015]
- Contains articles scraped from CNN and DailyMail Websites
- Summarization targets: Article ”Highlights”
  ▶ Short key points at the beginning of an article
  ▶ Average around three sentences
- For summarization: Concatenate highlights
‘It’s going to be mega-tough’: How Boris Johnson must win FIFTY Labour seats to offset losses to SNP Remainers in Scotland and the Lib Dems in the South

- Boris Johnson promised voters a new parliament for Christmas last night as he secured a General Election
- Comes after MPs backed Government Bill for a poll on Thursday, December 12, after weeks of dither and delay
- Jeremy Corbyn said the Labour Party would kick out ‘reckless' Conservatives and deliver a socialist Britain
- PM told MPs the election would deliver Brexit after months of 'unrelenting parliamentary obstructionism'
- He later addressed backbenchers, giving what one claimed was a 'King Henry V to Agincourt-type speech'

**Figure:** Example of an article with highlights
The following corpora are older, before neural methods
- Smaller, usually not used for training
- However, some appear as additional evaluation corpora
From 2001 to 2007: Datasets still in use:

- DUC 2002: News, MDS, SDS, abstractive and extractive,
- DUC 2003: News, SDS 10 word abstracts
- DUC 2003: News, MDS with topics, 100 word abstracts
- DUC 2004: News, two tasks similar to 2003, additional task in arabic and with question to focus summarization
- DUC 2005: News, MDS, queries, 250 word abstracts

On our servers: /resources/corpora/monolingual/annotated/DUC200{2|3|4|5}
DUC successor

- TAC 2008, 2009: Update Tasks
- TAC 2008: Opinion Task on blogs

See: https://www.nist.gov/tac/data/index.html
A THEORETIC APPROACH TO SUMMARIZATION
Gather dataset from human annotators

Construct model based on empirical observations/assumptions

[Peyrard, 2019] outlines a more principled approach based on information theory
To be able to talk about summarization in information theoretic terms, we need to model the ”information” in a text and a summary

- [Peyrard, 2019] uses *semantic units*
- Represent a text $X$ as distribution $\mathbb{P}_X$ over semantic units $\Omega$
- Interpretation 1: Frequency of semantic information in text
- Interpretation 2: $\mathbb{P}_X(\omega_i) \rightarrow$ Likelihood of $X$ entailing $\omega_i$
- Interpretation 3: $\mathbb{P}_X(\omega_i) \rightarrow$ Contribution of $\omega_i$ to meaning $X$
Summaries should condense information

Thus, avoid redundancy

Given a summary $S$, this can be represented as the entropy of $P_S$

$$H(S) = - \sum_{\omega_i} P_S(\omega_i) \log P_S(\omega_i)$$

We can define Redundancy as the negative entropy:

$$Red(S) = -H(S)$$
A summary $S$ of a document $D$ should reflect the content of $D$

- We can model this by minimizing the cross-entropy between $P_S$ and $P_D$

$$CE(S, D) = - \sum_{\omega_i} P_S(\omega_i) \log P_D(\omega_i)$$

- We can define Relevance as:

$$Rel(S, D) = -CE(S, D)$$
The Kullback-Leibler Divergence $KL(p\|q)$ measures how much information we lose by using $q$ to approximate $p$.

For our document summary pair $D, S$, the KL divergence is:

$$KL(S\|D) = CE(S, D) - H(S)$$

Using our definitions for redundancy and relevance, we get:

$$KL(S\|D) = -Rel(S, D) + Red(S)$$
$$-KL(S\|D) = Rel(S, D) - Red(S)$$

Minimizing $KL(S\|D)$ minimizes redundancy while maximizing relevance.
Many summarizers focus on identifying relevant information from $D$.

Intuitively, a summary should contain information that the reader does not have before.

We can model this as a third distribution $P_K$ over the assumed background knowledge $K$.

The relation between $K$ and $S$ is opposite of that between $D$ and $S$:

- $S$ should not contain information that is not present in $D$.
- $S$ should contain as much information that is not in $K$ as possible.
The relation between $K$ and $S$ is opposite of that between $D$ and $S$

- $S$ should not contain information that is not present in $D$
- $S$ should contain as much information that is not in $K$ as possible

As for relevance, we can model this using cross-entropy:

$$\text{Inf}(S, K) = \text{CE}(S, K)$$
So far, we have not considered how *important* the semantic units of the texts are.

However, summarization discards parts of the text.

To formalize this intuition, [Peyrard, 2019] formulate assume that this importance is encoded by a function $f$. 
Given $k_i = P_K(\omega_i)$, $d_i P_D(\omega_i)$, $f(P_D(\omega_i), P_K(\omega_i))$ encodes importance of $\omega_i$

[Peyrard, 2019] formulate four requirements for $f$:

- **Informativeness:**
  $\forall i \neq j$, if $d_i = d_j$ and $k_i > k_j$ then $f(d_i, k_i) < f(d_j, d_j)$

- **Relevance:**
  $\forall i \neq j$, if $d_i > d_j$ and $k_i = k_j$ then $f(d_i, k_i) > f(d_j, d_j)$

- **Two technical constraints:** Additivity and Normalization
Based on the requirements, this results in importance distributions of the following form (proof see [Peyrard, 2019]):

\[
P_{D,k}(\omega_i) = \frac{1}{C} \cdot \frac{d_i^\alpha}{k_i^\beta}
\]

With C as a normalization constant and \( \alpha \) to weight of relevance and \( \beta \) of informativeness.
The Summary Scoring Function

- Remember the previous combination of relevance and redundancy:

\[ -KL(S\|D) = Rel(S, D) - Red(S) \]

- Replacing the document distribution with the importance distribution, we arrive at the summary scoring function \( \Theta \):

\[ \Theta(S, D, K) = -KL(\mathbb{P}_S||\mathbb{P}_{D_K}) \]
We can decompose $\Theta$ into our criteria: Relevance, Redundancy, Informativeness

\[
\Theta(S, D, K) = -KL(P_S || P_{D_K})
\]

\[
-KL(P_S || P_{D_K}) = -CE(P_S, P_{D_K}) + H(P_S)
\]

\[
= \sum_{\omega_i} P_S(\omega_i) \log P_{D_K}(\omega_i) - \text{Red}(S)
\]

\[
\equiv \sum_{\omega_i} P_S(\omega_i)(\alpha \log P_D(\omega_i) - \beta \log P_K(\omega_i)) - \text{Red}(S)
\]

\[
= \alpha Rel(S, D) + \beta Inf(S, K) - \text{Red}(S)
\]
Study conducted by [Peyrard, 2019] shows that humans value summaries that were generated based on the summary scoring formula.

- Mentioned criteria were often used in previous research on summarization.
- Use this as (one possible) mental model to understand what parts of summarization models are doing.


**A simple theoretical model of importance for summarization.**  

**How to compare summarizers without target length? Pitfalls, solutions and re-examination of the neural summarization literature.**  