Evaluation and Theory

Seminar Neural Text Summarization

Julius Steen
(with some slides from Katja Markert)

30th October 2019

Organisation

Questions

- Please prefix mails with

  NTS

  from now on
- Do not forget to hand in questions before the seminar
- Paper assignment by tomorrow

ROUGE

What is a good summary?

- 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

- 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:

$$ROUGE_n^{(Rec)} = \frac{\sum_{s_{ref} \in S_{ref}} \sum_{g \in grams(s_{ref}, n)} count_{s_{ref}}(s_{sys}, g)}{\sum_{s_{ref} \in S_{ref}} \sum_{g \in grams(s_{ref}, n)} count(s_{ref}, g)}$$

- Later also precision:

$$ROUGE_n^{(Prec)} = \frac{1}{|S_{ref}|} \sum_{s_{ref} \in S_{ref}} \frac{\sum_{g \in grams(s_{ref}, n)} count_{s_{ref}}(s_{sys}, g)}{\sum_{g \in grams(s_{sys}, n)} count(s_{sys}, g)}$$

- 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

ROUGE-L

- The following works for single sentence summaries
- Recall:

$$ROUGE_{LCS}^{(Rec)} = \frac{LCS(s_{ref}, s_{sys})}{|s_{ref}|}$$

- Precision:

$$ROUGE_{LCS}^{(Prec)} = \frac{LCS(s_{ref}, s_{sys})}{|s_{sys}|}$$

- F-Score:

$$ROUGE_{LCS}^{(F)} = \frac{(1+\beta^2) ROUGE_{LCS}^{(Prec)} ROUGE_{LCS}^{(Rec)}}{ROUGE_{LCS}^{(Rec)} + \beta^2 ROUGE_{LCS}^{(Prec)}}$$

- 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:

$$ROUGE_{LCS}^{(Rec)} = \frac{\sum_{r_i \in s_{ref}} LCS_{\cup}(r_i, s_{sys})}{|s_{ref}|}$$

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:

$$ROUGE_{WLCS}^{(Rec)} = f^{-1}\left(\frac{WLCS(s_{sys}, s_{ref})}{f(|s_{ref}|)}\right)$$

Given $\alpha = 2$

Weights:

R/S	#	A	B	C	D	H	I
#	0	0	0	0	0	0	0
A	0	1	1	1	1	1	1
B	0	1	4	4	4	4	4
C	0	1	4	9	9	9	9
D	0	1	4	9	16	16	16
E	0	1	4	9	16	16	16
F	0	1	4	9	16	16	16
G	0	1	4	9	16	16	16

$$ROUGE_{WLCS}^{(Rec)} = \sqrt{\frac{16}{7^2}} = 0.57$$

Lengths:

R/S	#	A	B	C	D	H	I
#	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0
B	0	0	2	0	0	0	0
C	0	0	0	3	0	0	0
D	0	0	0	0	4	0	0
E	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0

ROUGE-S

- ROUGE based on Skip-Birams
- Skip-Bigram: Two words in the correct order in a sequence
- Recall:

$$ROUGE_S^{(Rec)} = \frac{skip2(s_{ref}, s_{sys})}{C(|s_{ref}|, 2)}$$

- Precision:

$$ROUGE_S^{(Prec)} = \frac{skip2(s_{ref}, s_{sys})}{C(|s_{sys}|, 2)}$$

- F-Score accordingly

ROUGE-SU

- 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

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]

ROUGE-F1 Pitfalls

- [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

Experiments

- 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

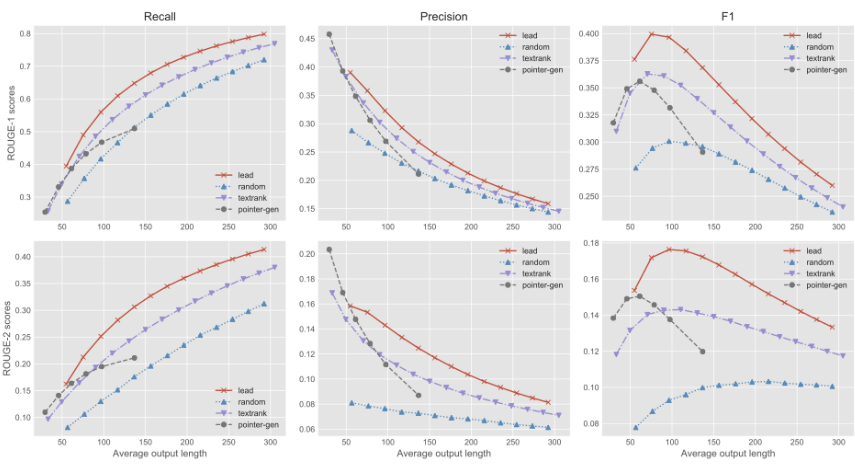

Figure: ROUGE-scores over different summary lengths (source: [Sun et al., 2019]

Suggested Solution

- [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.

ROUGE-F1 and Human Ratings

- [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

Corpora

CNN/Daily Mail

- 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

CNN/Daily Mail - Example

‘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

DUC: Document Understanding Conference

- The following corpora are older, before neural methods
- Smaller, usually not used for training
- However, some appear as additional evaluation corpora

DUC: Document Understanding Conference

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}`

TAC

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

"Usual" approach to summarization research

- Gather dataset from human annotators
- Construct model based on empirical observations/assumptions
- [Peyrard, 2019] outlines a more principled approach based on information theory

Semantic Units

- 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 Ω
- Interpretation 1: Frequency of semantic information in text
- Interpretation 2: $\mathbb{P}_X(\omega_i) \rightarrow$ Likelihood of X entailing ω_i
- Interpretation 3: $\mathbb{P}_X(\omega_i) \rightarrow$ Contribution of ω_i to meaning X

- Summaries should condense information
- Thus, avoid redundancy
- Given a summary S, this can be represented as the entropy of $\mathbb{P}_S$

$$H(S) = -\sum_{\omega_i} \mathbb{P}_S(\omega_i) \log \mathbb{P}_S(\omega_i)$$

- We can define Redundancy as the negative entropy:

$$Red(S) = -H(S)$$

Relevance

- A summary S of a document D should reflect the content of D
- We can model this by minimizing the cross-entropy between $\mathbb{P}_S$ and $\mathbb{P}_D$

$$CE(S, D) = -\sum_{\omega_i} \mathbb{P}_S(\omega_i) \log \mathbb{P}_D(\omega_i)$$

- We can define Relevance as:

$$Rel(S, D) = -CE(S, D)$$

Combining Relevance and Redundancy

- 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

Informativeness

- 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 $\mathbb{P}_K$ 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:

$$Inf(S, K) = CE(S, K)$$

Importance

- 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

The Importance Function

- Given $k_i = \mathbb{P}_K(\omega_i)$, $d_i\mathbb{P}_D(\omega_i)$, $f(\mathbb{P}_D(\omega_i), \mathbb{P}_K(\omega_i))$ encodes importance of ω_i
- [Peyrard, 2019] formulate four requirements for f:
  - Informativeness:

    $$\forall i \neq j, \text{ if } d_i = d_j \text{ and } k_i > k_j \text{then } f(d_i, k_i) < f(d_j, d_j)$$

  - Relevance:

    $$\forall i \neq j, \text{ if } d_i > d_j \text{ and } k_i = k_j \text{then } f(d_i, k_i) > f(d_j, d_j)$$

  - Two technical constrains: Additivity and Normalization

- Based on the requirements, this results in importance distribiutions of the following form (proof see [Peyrard, 2019]:

$$\mathbb{P}_{\frac{D}{K}}(\omega_i) = \frac{1}{C} \cdot \frac{d_i^{\alpha}}{k_i^{\beta}}$$

With C as a normalization constant and α to weight of relevance and β of informativeness

The Summary Scoring Function

- Remember the previous combination of relevance and redundancy:

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

- Replacing the the document distribution with the importance distribution, we arrive at the summary scoring function Θ:

$$\Theta(S, D, K) = -KL(\mathbb{P}_S||\mathbb{P}_{\frac{D}{K}})$$

- We can decompose Θ into our criteria: Relevance, Redundancy, Informativeness

$$
\begin{aligned}
\Theta(S, D, K) &= -KL(\mathbb{P}_S || \mathbb{P}_{\frac{D}{K}}) \\
-KL(\mathbb{P}_S || \mathbb{P}_{\frac{D}{K}}) &= -CE(\mathbb{P}_S, \mathbb{P}_{\frac{D}{K}}) + H(\mathbb{P}_S) \\
&= \sum_{\omega_i} \mathbb{P}_S(\omega_i) \log \mathbb{P}_{\frac{D}{K}}(\omega_i)) - Red(S) \\
&\equiv \sum_{\omega_i} \mathbb{P}_S(\omega_i)(\alpha \log \mathbb{P}_D(\omega_i) - \beta \log \mathbb{P}_K(\omega_i)) - Red(S) \\
&= \alpha Rel(S, D) + \beta Inf(S, K) - Red(S)
\end{aligned}
$$

How to use this?

- 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

References I

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015).
Teaching machines to read and comprehend.
In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Lin, C.-Y. (2004).
ROUGE: A package for automatic evaluation of summaries.
In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., and Xiang, B. (2016).
Abstractive text summarization using sequence-to-sequence RNNs and beyond.
In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

References II

Peyrard, M. (2019).
A simple theoretical model of importance for summarization.
In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Sun, S., Shapira, O., Dagan, I., and Nenkova, A. (2019).
How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature.
In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.