Neural Text
Summarization

Course Organization and Papers
Julius Steen
WS19/20
About me

• **Mail**: steen@cl.uni-heidelberg.de

• **Office Hours**: Wednesday, 13:00-14:00, R. 123a
What you should already know about

• Neural Networks and how to train them
  • Structure of neurons, backpropagation etc.

• Common NLP-architectures/concepts
  • LSTMs
  • CNNs
  • seq2seq
  • attention
Useful, but not required

- Transformers
- Reinforcement-learning
What is Summarization?

• „[A] reductive transformation of source text to summary text through content reduction by selection and/or generalisation on what is important in the source“.¹

• **Input:** Long text(s) with irrelevant and/or redundant information

• **Output:** Concise, non-redundant summary

---

Extractive vs. Abstractive Summarization

- Summarization has two subcategories

  - **Extractive Summarization** only identifies key sentences from input, possible rearranging them.

  - **Abstractive Summarization** generates new text „from scratch“

- Intermediate category: **Compressive Summarization** uses no new words, but may remove/rearrange on the word level
Summarization Tasks

• Sentence Summarization/Headline Generation
  
  • Generate a headline based, e.g. on the initial sentence of a document (Not the focus of this seminar)

• Single Document Summarization (SDS)
  
  • Generate a short summary based on a single input document

• Multi Document Summarization (MDS)
  
  • Generate a concise summary based on multiple documents

• Many others: Query Summarization, Timeline Summarization etc. (not in this seminar)
Impact of Neural Methods

• Before neural summarization
  
  • Focus on extractive methods
  
  • Relatively small, but well curated datasets (DUC)
  
  • Many unsupervised systems, some supervision, focus on global optimization of scoring functions (Integer Linear Programming, Submodular Functions, Determinantal Point Processes, …)
Impact of Neural Methods

- With Neural Summarization
  - Viable abstractive systems
  - Huge, but noisy datasets with unclear summarization schemas (CNN/Dailymail)
  - Initially focused on Sentence Summarization and later SDS, now some work on MDS
Relation to NMT

- Abstractive Text Summarization is similar to and often influenced by Neural Machine Translation (NMT)
  - Translate document in „document language“ to „summary language“
- Same basic seq2seq architecture can be used for abstractive summarization
- However, there are important differences
  - Copying turns out to be very important
  - Input are full documents, or even multiple documents
  - Not all content should be preserved (content selection)
Organization
This Seminar

• Rest of today: Organisation and paper overview

• Next week: Fundamentals
  • Some datasets
  • Evaluation Measures (ROUGE)
  • Possibly some fun summarization theory

• After that: presentations by students/reading groups

• Literature list with schedule on the course page
How to get points

• **Active Participation**
  - No more than one unexcused absence
  - Active Participation in classroom discussion

• **Preparation**
  - Read all papers due to be presented (at most two)
  - Hand in two questions or comments about each paper via mail (steen@cl)
  - Deadline: Each Monday before the seminar, 3pm
  - Part of your participation grade
How to get points

- Additionally, *one* of the following
  - Term paper
  - A small implementation project
  - Second presentation
Presentation

- PS
  - usually one paper
  - 30 minutes

- HS
  - usually two papers
  - 60 minutes

- Discuss the presentation with me before the seminar (in my office hours)
Presentation Grading

- Presentation Content
  - Explain methods and results
  - Point out strengths, weaknesses
  - Compare to what we have seen before in the seminar

- Presentation Style
  - Structure
  - Clarity of the presentation
  - Design of the slides, use of illustration etc.
Term Paper

• Max. 10 (PS) or 14 (HS) pages (standard latex article template)

• Contextualise the contents of one of the papers
  • Compare with others (other approaches, or earlier research on summarization)
  • Find similarities among approaches

• Approaches should be well explained, show that you understood them
Project

- Max 8 pages (both PS and HS)
- Submit (working) code + project report
- Possibilities
  - A clean reimplementation of one of the approaches
  - An exploration of one of your own ideas
  - Corpus analysis
- ...
Submission and Final Grade

- Submission of all final projects and papers by 30th of April via mail as PDF
- If you do a second presentation, you are done by the end of the semester of course
- Final grade is made up of
  - Participation (30%)
  - Presentation (40%)
  - Project, term paper or second presentation (30%)
    - If you do a second presentation, the better one will count 40%
Selecting a Paper

• Papers are tentatively labeled for HS or PS
  • HS papers are generally more difficult, cover a wider area
• If you want to do PS, but are interested in HS: no problem
• If you want to cover PS papers, but want HS points:
  • Write this in your registration mail
  • We can possibly add more background, comparison
• If you want to present a paper not listed here, this might also be possible
The Papers
Nallapati et. al. (2017) (PS)

• Simple classification task for every sentence
  • Should it be in the summary or not?

• This can be framed as a sequence labelling task => RNN

• Derive ground-truth labels from abstractive gold summaries via heuristic

• CE-loss for training
Yasunaga et. al. (2017) (PS)

• Built on a classical two step procedure: salience estimation, followed by selection for MDS

• Salience estimation = regression on ROUGE-scores

• Construct a graph based on sentence similarity, discourse markers and salience

• Use a graph convolutional network over the graph for ROUGE prediction
Pointer Mechanisms (PS)

- Problem in abstractive summarization: how to deal with unknown words?
  - Extending the vocabulary increases parameter count massively
  - We can never cover all words
- Idea: Point to the unknown words
- Two approaches: (See 2017, Nallapati 2016)
Grusky et. al. (2018) (PS)

- Not all Summarization Datasets are equal
  - Important measure: How abstractive are the datasets?
- Introduces new datasets
- New metrics for dataset analysis
Narayan et. al. (2018) (PS)

• Existing methods tend towards extraction
  • Analysis reveals that this is also due to dataset characteristics

• New dataset: XSUM (Extreme Summarization)
  • Very short summaries
  • High abstraction

• Also describes a CNN-based seq2seq model for the problem
Extractive Summarization without Labels (HS)

• Heuristic labels for extractive summarization are only approximations

• Can we directly optimise evaluation metrics (ROUGE)?

• Solution: Reinforcement-learning over sentence labels
Extractive Summarization without Labels

• Narayan et. al. (2018b)
  • Sample complete sentence labelling
  • Compute ROUGE as feedback score
    \[
    \nabla L(\theta) = - \mathbb{E}_{\hat{y} \sim p_\theta}[r(\hat{y}) \nabla p_\theta(\hat{y} | \theta, D)]
    \]

• Zhang et. al. (2018)
  • Trains additional compression model
  • Uses compression model to identify an alignment between extracted sentences and gold summary for feedback
Extractive Summarization without Labels

Source: Narayan et. al. (2018 b)
More Architectures for Abstractive Summarization (HS)

- There are many tweaks to abstractive summarization architectures
- Another way to improve(?) results: reinforcement learning to directly optimise ROUGE
  - Self-critical policy gradient (Rennie et. al. 2016)
Paulus et al. (2018)

In this section, we present our intra-attention model based on the encoder-decoder network (Sutskever et al., 2014). In all our equations, $x = \{x_1, x_2, \ldots, x_n\}$ represents the sequence of input (article) tokens, $y = \{y_1, y_2, \ldots, y_{n_0}\}$ the sequence of output (summary) tokens, and $\mathbf{k}$ denotes the vector concatenation operator.

Our model reads the input sequence with a bi-directional LSTM encoder $\{\text{RNN}_e^{\text{fwd}}, \text{RNN}_e^{\text{bwd}}\}$ computing hidden states $h_e^i = [h_e^{fwd}_i, h_e^{bwd}_i]$ from the embedding vectors of $x_i$. We use a single LSTM decoder $\text{RNN}_d$, computing hidden states $h_d^t$ from the embedding vectors of $y_t$.

Both input and output embeddings are taken from the same matrix $W_{\text{emb}}$. We initialize the decoder hidden state with $h_d^0 = h_e^n$.

### 2.1 Intra-Temporal Attention on Input Sequence

At each decoding step $t$, we use an intra-temporal attention function to attend over specific parts of the encoded input sequence in addition to the decoder's own hidden state and the previously-generated word (Sankaran et al., 2016). This kind of attention prevents the model from attending over the same parts of the input on different decoding steps.

Nallapati et al. (2016) have shown that such an intra-temporal attention can reduce the amount of repetitions when attending over long documents.

We define $e_{ti}$ as the attention score of the hidden input state $h_e^i$ at decoding time step $t$:

$$ e_{ti} = f(h_d^t, h_e^i), $$

where $f$ can be any function returning a scalar $e_{ti}$ from the $h_d^t$ and $h_e^i$ vectors. While some attention models use functions as simple as the dot-product between the two vectors, we choose to use a bilinear function:

$$ f(h_d^t, h_e^i) = h_d^t \mathbf{T} W_{\text{attn}} h_e^i. $$
Celikyilmaz et. al. (2018)
Controllability (PS/HS)

- "Traditional" summarization put a lot of emphasis on length constraints

- Neural methods have difficulty sticking to exact constraints

- We might also want to influence style, or focus of the summary

- How can we integrate this into (abstractive) summarizers?
Controllability (PS)

- Fan et. al. (2018)
  - General approach to control for length and additional summary characteristics
- Liu et. al. (2018)
  - Focus on length control
  - Directly integrated into CNN architecture
Controllability - Global Optimization (HS)

- Control-methods only give hints to the network

- Can we do better? => Global optimization based on Minimum Risk Training (Shen et. al. 2016)
Pretraining for Summarization (HS)

- Pretrained transformer architectures have proven useful for many tasks

- Zhang et. al. (2019b) use BERT to encode and generate summaries

  - Challenge: BERT is bidirectional, how can we decode with that?

- Zhang et. al. (2019a) introduce a hierarchical transformer architecture for extractive summarization
Improving Summary Coherence (PS/HS)

• The commonly used CNN/DM has bullet-point like summaries and lack global coherence

• Gabriel et. al (2019) introduce a new dataset with scientific summarization

  • They also integrate a coherence model into the decoding process

• Improve global coherence
Improving Summary Coherence (PS/HS)

• Wu and Hu (2018) integrate a coherence reward into RL-based extractive summarization

• Sharma et. al. (2019) improve coherence for abstractive summarization

  • They also integrate coreference information into encoding

  • Coherence model is used in conjunction with reinforcement learning
Factual Correctness (PS)

- Abstractive Summarizers can „hallucinate“ information that is not in the summary

- Cao et. al. (2017) observe the following example

  - **Source:** the repatriation of at least #,#,# bosnian moslems was postponed friday after the unhcr pulled out of the first joint scheme to return refugees to their homes in northwest bosnia

  - **seq2seq:** bosnian moslems postponed after unhcr pulled out of bosnia

- They propose an IE-based method to alleviate this
Integrating Knowledge (PS)

- Summarization is often focused on real-world news
- Giving background knowledge might help in creating better summaries
- Amplayo et. al. (2018) integrate KB-information

Source: Amplayo et. al. (2018)
Abstractive MDS (HS)

- Acquiring training data for MDS is difficult
- Lebanoff at. al. (2018) propose adapting the Pointer Generator trained on SDS to MDS
- Recently Fabbri et. al. (2019) have introduced a Multi-Document Corpus and corresponding architecture using maximum marginal relevance to modify attention weights

\[
MMR = \underset{D_i \in R \setminus S}{\text{argmax}} \left[ \lambda \text{Sim}_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} \text{Sim}_2(D_i, D_j) \right]
\]
What now?

- Write a mail to steen@cl… by Sunday (26th) containing…
  - Three papers/sessions that you would like to present, ranked by your preferences
  - If you are interested in a second presentation, two more papers you would like to present
  - At most one date on which you can absolutely not present on (current dates might change)
  - Your name

- For next time: Read the ROUGE-Paper (Lin, 2004) and write two comments/questions