Human Language Technology Research Institute

Abstractive Summarization: A Survey of the State of the Art

Hui Lin and Vincent Ng

Human Language Technology Research Institute

University of Texas at Dallas

AAAI 2019

Automatic Text Summarization

- Produce a summary of either
  - one document (single-document summarization)
  - or a set of documents (multi-document summarization)

Why Automati Text Summarization?

- Alleviate user information overload

How?

- **Extractive Summarization**
  - Select a subset of sentences in the source document(s) for inclusion in the summary
- **Abstractive Summarization**
  - Generate sentences that may contain words/phrases not present in the source document(s)

Abstractive Summarization: Example

Source

The Sri Lanka government on Wednesday announced the closure of government schools with immediate effect as a military campaign against Tamil separationists escalated in the north of the country.

Summary

Sri Lanka closes schools as the war escalates.

Plan for the Talk

- Evaluation methods
- Datasets
- Approaches
- The state of the art

Evaluation Methods

- Manual evaluation
  - Human judges rate a summary along multiple dimensions of quality, such as content, grammaticality, and coherence

- Automatic evaluation
  - BLEU (Papineni et al., 2002)
  - METEOR (Denkowski and Lavie, 2014)
  - Pyramid (Nenkova and Passoneau, 2004)
  - ROUGE (Lin and Hovy, 2003)

ROUGE

- Variants: ROUGE-N, ROUGE-SU, ROUGE-L, …
  - All ROUGE variants compute the degree of lexical overlap between a reference summary and a candidate summary
    - More overlap → higher ROUGE score → better summary

ROUGE

- Variants: ROUGE-N, ROUGE-SU, ROUGE-L, …
  - All ROUGE variants compute the degree of lexical overlap between a reference summary and a candidate summary
    - More overlap → higher ROUGE score → better summary

- Adequate for abstractive summarization?

- Designing appropriate evaluation metrics is challenging

Datasets

- **DUC** (Document Understanding Conference, 2000-2007)
  - **Pros**: popularly used to evaluate abstractive summaries
  - **Cons**: corpora are relatively small

Datasets

- **DUC** (Document Understanding Conference, 2000-2007)
  - **Pros**: popularly used to evaluate abstractive summaries
  - **Cons**: corpora are relatively small
- **Annotated English Gigaword** (Rush et al., 2015)
  - **Pros**: considerably larger than DUC (10 million documents)
  - **Cons**: input is composed of a single sentence (first sentence of article), summary is not generated by human (just the headline)

Datasets

- **DUC** (Document Understanding Conference, 2000-2007)
  - **Pros**: popularly used to evaluate abstractive summaries
  - **Cons**: corpora are relatively small
- **Annotated English Gigaword** (Rush et al., 2015)
  - **Pros**: considerably larger than DUC (10 million documents)
  - **Cons**: input is composed of a single sentence (first sentence of article), summary is not generated by human (just the headline)
- **CNN/Daily Mail** (Nallapati et al., 2016)
  - Human summary for each story in the corpus
  - **Pros**: large (286K for training, 13K for validation, 11K for test), each story is longer than those in DUC and Gigaword (781 tokens on average) and contains multiple sentences

Approaches to Abstractive Summarization

- Classical approaches
- Neural approaches

Neural Approaches

- Originally developed for neural machine translation

- Key advantage:
  - Provide an **end-to-end** approach
    - Learning how to (1) **abstract** from the source document and (2) **generate** words to form an abstractive summary in one shot
    - … unlike classical approaches where the key steps are performed in a pipeline fashion
      - Many (potentially suboptimal) heuristic decisions are involved
      - Errors may propagate from one step to the next

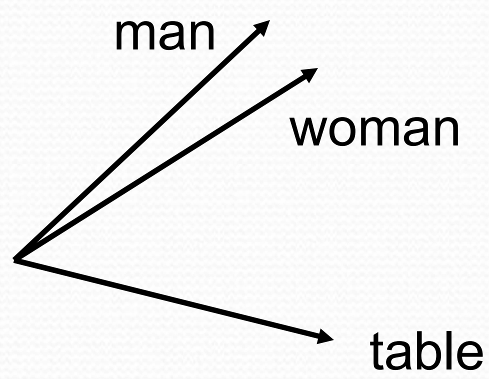
The Encoder-Decoder Framework

source
document

The Encoder-Decoder Framework

source document

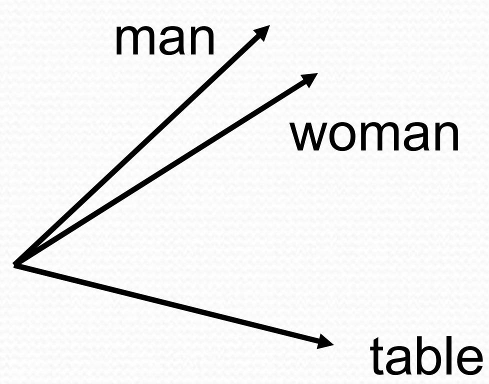
- Each word in the source document is represented as a low-dimensional vector



The Encoder-Decoder Framework

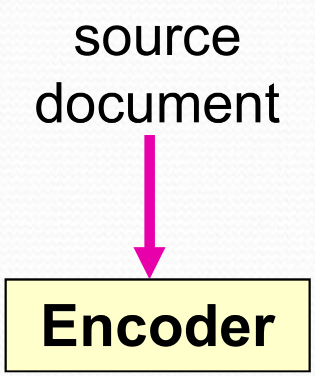
source document

- Each word in the source document is represented as a low-dimensional vector



Word vectors better capture word meaning

The Encoder-Decoder Framework



- Each word in the source document is represented as a low-dimensional vector

- Encodes the source document into an internal representation, often a fixed-length vector known as the context vector

The Encoder-Decoder Framework

source document

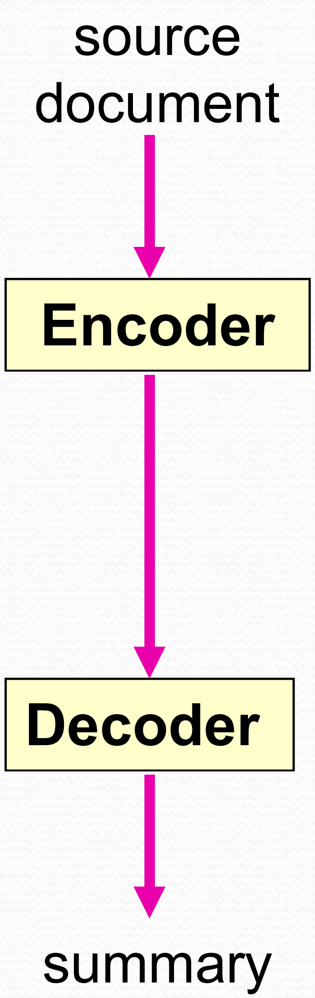
Encoder

Decoder

summary

- Each word in the source document is represented as a low-dimensional vector
- Encodes the source document into an internal representation, often a fixed-length vector known as the context vector
- Outputs a summary by generating a word in each timestep

The Encoder-Decoder Framework

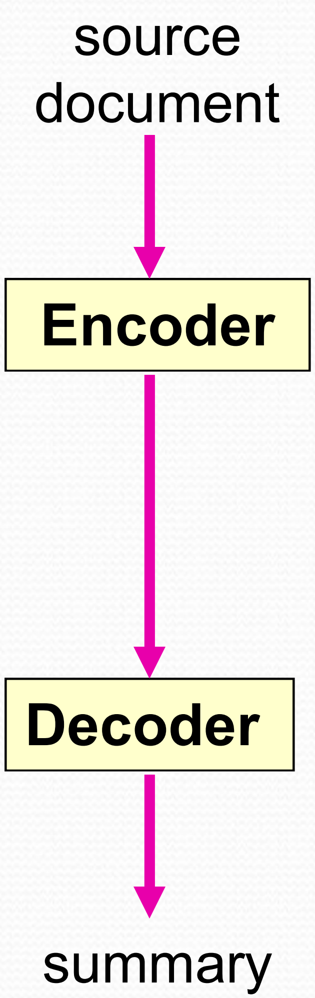


- Each word in the source document is represented as a low-dimensional vector

- Encodes the source document into an internal representation, often a fixed-length vector known as the context vector

- In each timestep:
  - takes as input (1) the context vector and (2) the words generated so far as a summary
  - generates a distribution over the vocabulary
    - Outputs most probable word or keeps k-best paths

Implementing the Framework



- CNNs (convolutional neural nets)
- LSTMs: better at encoding long sequences
- GRUs: fewer parameters, faster to train
- Encoding long documents remains challenging

- RNNs (recurrent neural nets)

Implementing the Framework

source document

Encoder

- CNNs (convolutional neural nets)
- LSTMs: better at encoding long sequences
- GRUs: fewer parameters, faster to train
- Encoding long documents remains challenging

Decoder

- RNNs (recurrent neural nets)

summary

Joint training

The Encoder-Decoder Framework

- Recent work has focused on improving this framework

Improvement 1: Attention

- identifies the important words in a document
  - **Intuition**: they are more likely to appear in a summary

- **Idea**: learn a weight for each word indicating its importance
  - More important words are given higher weights

Improvement 2: Distraction/Coverage

- **Motivation**: attention may cause some content to be overly focused, leading to redundancy in the resulting summary

- Distraction avoids focusing on the same content by reducing probability of repeated content or weight associated with it

Improvement 3: Pointers and Copying

- **Motivation**: neural models are poor at generating rare words and out-of-vocabulary (OOV) words, but some of these words could be important to have in a summary

- Pointer networks and copying mechanisms copy a word or a text segment directly from the source to the summary
  - can be viewed as an extension of attention to rare or OOV words that are important

Improvement 4: Reinforcement Learning

- **Motivation**: the encoder-decoder framework has 2 weaknesses
  - Network minimizes **maximum-likelihood loss**, but this is not equivalent to optimizing the **evaluation metric** (e.g., ROUGE)
  - Decoder has an **exposure bias**
    - Training: predict next word assuming previous words are all correct
    - Application: predict next word using previously generated words

- Reinforcement Learning (RL) addresses both issues
  - Can be used to optimize objectives that are not differentiable
  - Doesn't need gold summaries for training [no exposure bias]

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

The State of the Art (CNN/Daily Mail)

System	ROUGE-1	ROUGE-2	ROUGE-L
words-lvt2k-temp-att (2016)	35.5	13.3	32.7
pointer-generator (2017)	36.4	15.7	33.4
pointer-generator+coverage (2017)	39.5	17.3	36.4
MLE (2017)	38.3	14.8	35.5
RL (2017)	41.2	15.8	**39.1**
DCA MLE+SEM+RL (2018)	**41.7**	**19.5**	37.9
SummaRuNNer (2017)	39.6	16.2	35.3
lead-3 (2017)	40.3	17.7	36.8
REFRESH (2018)	40.0	18.2	36.6

Future Directions

- Text simplification
  - Encoding long sentences remains a challenge
  - Apply text simplification to simplify/shorten long sentences?

Future Directions

- Text simplification
- Phrase-based models
  - Use phrase-based (rather than word-based) encoders and decoders to better capture text semantics

Future Directions

- Text simplification
- Phrase-based models
- Multi-document abstractive summarization

Future Directions

- Text simplification
- Phrase-based models
- Multi-document abstractive summarization
- Evaluation on different text types
  - Most work was evaluated on news articles because (1) they are well-organized and (2) training data is abundant
  - Perform evaluations on meetings and conversations