Salience Estimation and Faithful Generation:

Modeling Methods for Text Summarization and Generation.

Chris Kedzie

Columbia University

kedzie@cs.columbia.edu

January 3, 2019

Key Challenges to Summarization

- **Salience Estimation** — determining the most important or essential information in the input

- **Faithful Generation** — guarranteeing that the resulting summary does not misrepresent the input or otherwise hallucinate facts.

Key Challenges to Summarization

- **Salience Estimation** — determining the most important or essential information in the input
  - **Feature Based Models of Salience** (Stream Summarization)
  - **Deep Learning Models of Salience** (Single/Multi-Document Summarization)

- **Faithful Generation** — guarranteeing that the resulting summary does not misrepresent the input or otherwise hallucinate facts.

Key Challenges to Summarization

- **Salience Estimation** — determining the most important or essential information in the input
    - **Feature Based Models of Salience** (Stream Summarization)
    - **Deep Learning Models of Salience** (Single/Multi-Document Summarization)
- **Faithful Generation** — guarranteeing that the resulting summary does not misrepresent the input or otherwise hallucinate facts.
    - **Data-to-Text** (Text Generation)
    - **Text-to-Text** (Abstractive Summarization)

Talk Outline

1. Feature Based Models of Sentence Salience
   - Stream Summarization
   - Learning-to-Search Summarizer
2. Deep Learning Models of Salience
   - Sentence Salience
   - Word Salience
3. Faithful Generation
4. Research Plan and Contributions

Talk Outline

1. Feature Based Models of Sentence Salience
   - Stream Summarization
   - Learning-to-Search Summarizer
2. Deep Learning Models of Salience
   - Sentence Salience
   - Word Salience
3. Faithful Generation
4. Research Plan and Contributions

Feature Based Models of Sentence Salience

Two models for sentence extractive stream summarization:

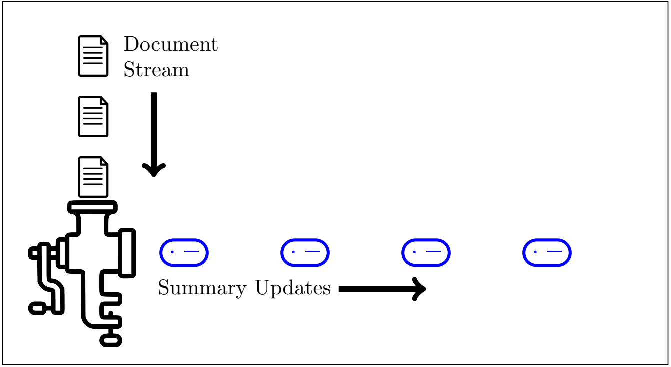
1. **SAP** Sentence salience regression biases an exemplar based clustering algorithm (Salience-biased Affinity Propagation Clustering).
2. **L2S** Sentence salience regression with exploration (learning-to-search) for dynamic summary features.

Feature Based Models of Sentence Salience

Two models for sentence extractive stream summarization:

1. **SAP** ~~Sentence salience regression biases an exemplar based clustering algorithm (Salience-biased Affinity Propagation Clustering).~~
2. **L2S** Sentence salience regression with exploration (learning-to-search) for dynamic summary features.

Stream Summarization



- Data from **TREC Temporal Summarization Track**, 2013-2015
- **Query focused** crisis monitoring scenario
- E.g., summarize a stream of news about Hurricane Sandy

TREC Temporal Summarization Data

- **Stream Corpus** (Document Stream Input)
  - Timestamped collection of news websites.
  - Simulates web news publishing.
- **Event Queries** (Summarizer Focus)
  - Queries correspond to real-life disasters/crises.
  - E.g. "Hurricane Sandy," or "Boston Marathon Bombing"
- **Event Nuggets** (Reference "Summary" atoms)
  - Timestamped text snippets of important event facts.
  - Selected by NIST from Wikipedia event page.

Nuggets for event query "hurricane sandy"	
[10/23 8:20pm]	Sandy strengthened from a tropical depression into a tropical storm
[10/23 8:20pm]	2 pm Oct 23 Sandy moving north-northeast at 4 knots
[10/23 8:53pm]	forecast track uncertain
[10/25 12:20am]	In Jamaica damage was extensive

TREC Temporal Summarization Data

- **Stream Corpus** (Document Stream Input)
  - Timestamped collection of news websites.
  - Simulates web news publishing.
- **Event Queries** (Summarizer Focus)
  - Queries correspond to real-life disasters/crises.
  - E.g. "Hurricane Sandy," or "Boston Marathon Bombing"
- **Event Nuggets** (Reference "Summary" atoms)
  - Timestamped text snippets of important event facts.
  - Selected by NIST from Wikipedia event page.

Nuggets for event query "hurricane sandy"	
[10/23 8:20pm]	Sandy strengthened from a tropical depression into a tropical storm
[10/23 8:20pm]	2 pm Oct 23 Sandy moving north-northeast at 4 knots
[10/23 8:53pm]	forecast track uncertain
[10/25 12:20am]	In Jamaica damage was extensive

TREC Temporal Summarization Data

- **Stream Corpus** (Document Stream Input)
  - Timestapped collection of news websites.
  - Simulates web news publishing.
- **Event Queries** (Summarizer Focus)
  - Queries correspond to real-life disasters/crises.
  - E.g. "Hurricane Sandy," or "Boston Marathon Bombing"
- **Event Nuggets** (Reference "Summary" atoms)
  - Timestamped text snippets of important event facts.
  - Selected by NIST from Wikipedia event page.

Nuggets for event query "hurricane sandy"

[10/23 8:20pm]	Sandy strengthened from a tropical depression into a tropical storm
[10/23 8:20pm]	2 pm Oct 23 Sandy moving north-northeast at 4 knots
[10/23 8:53pm]	forecast track uncertain
[10/25 12:20am]	In Jamaica damage was extensive

TREC Temporal Summarization Data

- **Stream Corpus** (Document Stream Input)
  - Timestamped collection of news websites.
  - Simulates web news publishing.
- **Event Queries** (Summarizer Focus)
  - Queries correspond to real-life disasters/crises.
  - E.g. "Hurricane Sandy," or "Boston Marathon Bombing"
- **Event Nuggets** (Reference "Summary" atoms)
  - Timestamped text snippets of important event facts.
  - Selected by NIST from Wikipedia event page.

Nuggets for event query "hurricane sandy"

[10/23 8:20pm]	Sandy strengthened from a tropical depression into a tropical storm
[10/23 8:20pm]	2 pm Oct 23 Sandy moving north-northeast at 4 knots
[10/23 8:53pm]	forecast track uncertain
[10/25 12:20am]	In Jamaica damage was extensive

Model Features

- Surface Features
- Query Features
- Language Model Scores
- Single Document Summarization Rankings
- Nugget Probability
- Document Frequency
- Update Similarity

Model Features

- Surface Features
- Query Features
- Language Model Scores
- Single Document Summarization Rankings
- **Nugget Probability**
    - N-gram classifier for predicting when a sentence contains a nugget.
- Document Frequency
- Update Similarity

Model Features

- Surface Features
- Query Features
- Language Model Scores
- Single Document Summarization Rankings
- **Nugget Probability**
  - N-gram classifier for predicting when a sentence contains a nugget.
- **Document Frequency**
  - Hour-to-hour relative change in number of documents in the stream.
- Update Similarity

Model Features

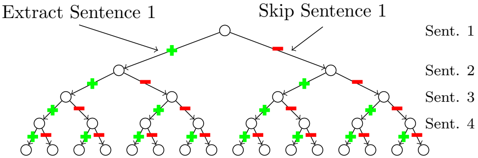
- Surface Features
- Query Features
- Language Model Scores
- Single Document Summarization Rankings
- **Nugget Probability**
    - N-gram classifier for predicting when a sentence contains a nugget.
- **Document Frequency**
    - Hour-to-hour relative change in number of documents in the stream.
- **Update Similarity**
    - Similarity of candidate sentence extract to all previously extracted sentences.

Learning-to-Search Summarization

- Nugget Probability model learned "offline."
  - I.e. doesn't take into account future extraction decisions, or account for compounding error.
  - It's difficult to take into account summary features.

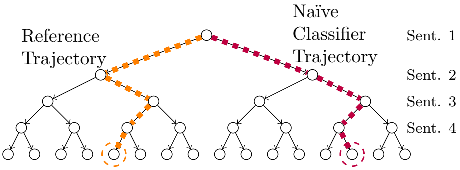
Learning-to-Search Summarization

- Nugget Probability model learned "offline."
  - I.e. doesn't take into account future extraction decisions, or account for compounding error.
  - It's difficult to take into account summary features.



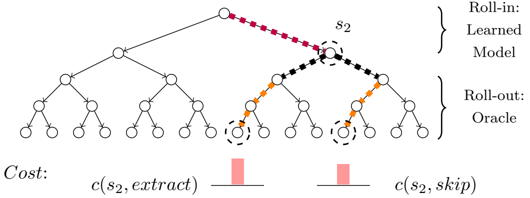
Learning-to-Search Summarization

- Nugget Probability model learned "offline."
  - I.e. doesn't take into account future extraction decisions, or account for compounding error.
  - It's difficult to take into account summary features.



Learning-to-Search Summarization

- Nugget Probability model learned "offline."
  - I.e. doesn't take into account future extraction decisions, or account for compounding error.
  - It's difficult to take into account summary features.
- Learning-to-search (Daumé, Langford, and Marcu (2009)), effectively an exploration/data sampling scheme can correct for this!



Experiments

- Leave-One-Out evaluation
  - 44 TREC 2015 Temporal Summarization events.
  - 5 queries randomly selected for development set.
  - Leave-One-Out Evaluation on remaining 39 events.

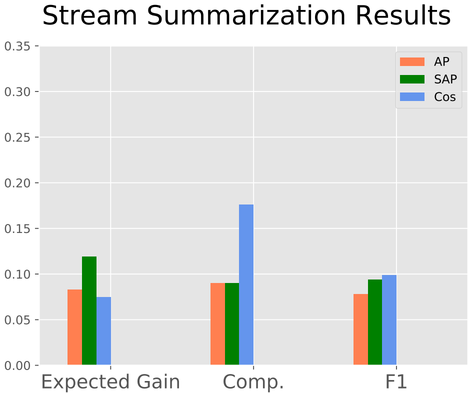
Evaluation Metrics

- TREC Temporal Summarization metrics
  - **Expected Gain** — the average number of novel nuggets in each extracted sentence; $\approx$ nugget precision.
  - **Comprehensiveness** — nugget recall.

System Comparisons

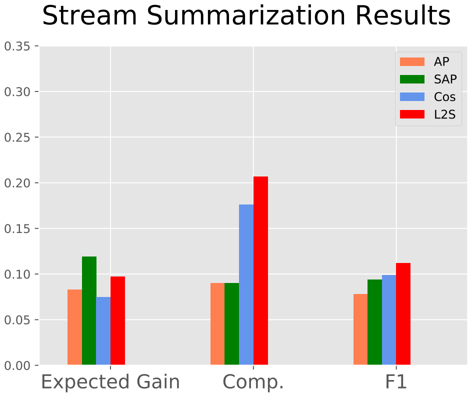
- **Baseline Models**
  - **Cos** — Cosine Similarity Threshold
    - Selects sentence if it's max similarity to any previous update is below a threshold.
    - Only examines first sentences of article.
  - **AP** — Affinity Propagation Clustering
- **Our Models**
  - **SAP** — Salience-Biased Affinity Propagation Clustering
  - **L2S** — Learning to Search
  - **L2SCos** — Learning to Search with cosine similarity threshold

Results

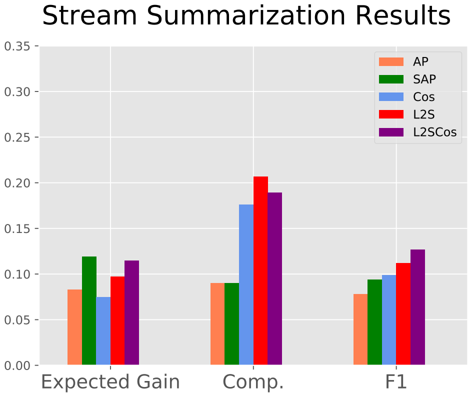


Stream Summarization Results
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
AP
SAP
Cos
Expected Gain
Comp.
F1

Results


Stream Summarization Results
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Expected Gain
Comp.
F1
AP
SAP
Cos
L2S

Results


Stream Summarization Results
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
AP
SAP
Cos
L2S
L2SCos
Expected Gain
Comp.
F1

Stream Summarization Takeaways

- Learning-to-Search approach can beat tough lead-based baseline.
- Dynamic summary features important for achieving performance.

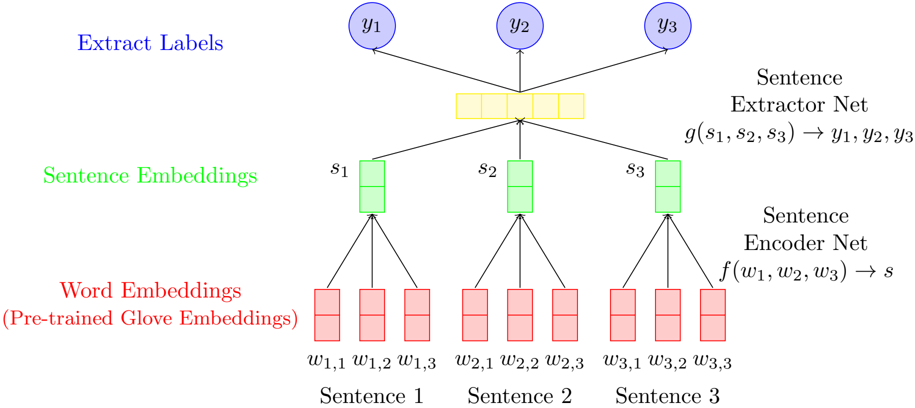
Talk Outline

1. Feature Based Models of Sentence Salience
   - Stream Summarization
   - Learning-to-Search Summarizer
2. Deep Learning Models of Salience
   - Sentence Salience
   - Word Salience
3. Faithful Generation
4. Research Plan and Contributions

Sentence Salience

- Lots of recent work on deep nets for sentence extractive news summarization.
  - Cheng and Lapata, (2016); Nallapati et al., (2016); Narayan et al., (2018), and more...
- Salience is modeled as a sentence classification problem,
  - i.e. salience = probability of including a sentence in an extract summary.
- Multiple architecture tweaks in each work
  - Difficult to tell what is actually driving improvements.

Summarizer Architecture



Sentence Salience

- We experiment with several popular sentence encoder methods.
  - Word Embedding Averaging
  - Recurrent Neural Nets (RNNs)
  - Convolutional Neural Nets (CNNs)

Sentence Salience

- We experiment with several popular sentence encoder methods.
  - Word Embedding Averaging
  - Recurrent Neural Nets (RNNs)
  - Convolutional Neural Nets (CNNs)

- We experiment with two state of art sentence extractor architectures (Nallapati et al., 2016) and (Cheng & Lapata, 2016), and

Sentence Salience

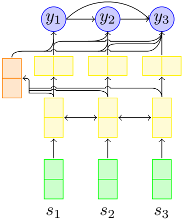
- We experiment with several popular sentence encoder methods.
  - Word Embedding Averaging
  - Recurrent Neural Nets (RNNs)
  - Convolutional Neural Nets (CNNs)

- We experiment with two state of art sentence extractor architectures (Nallapati et al., 2016) and (Cheng & Lapata, 2016), and

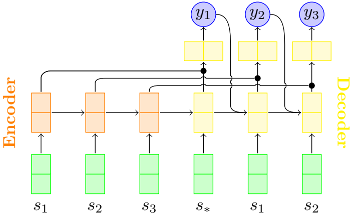
- propose simplified versions of each (RNN and Seq2Seq, respectively).

Sentence Extractors

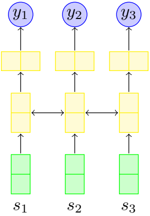
SummaRunner Extractor (Nallapati et al. 2016)



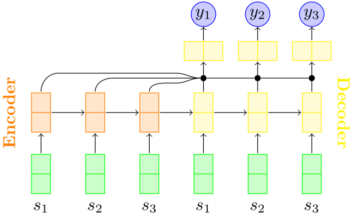
Cheng & Lapata Extractor (Cheng and Lapata, 2016)



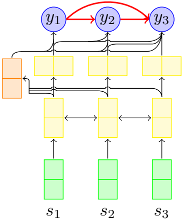
RNN Extractor (ours)



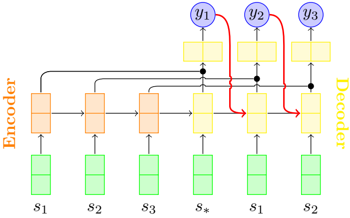
Seq2Seq Extractor (ours)



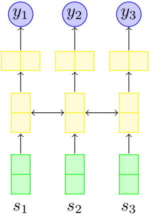
SummaRunner Extractor (Nallapati et al. 2016)



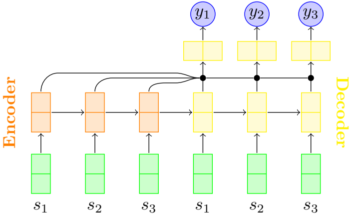
Cheng & Lapata Extractor (Cheng and Lapata, 2016)



RNN Extractor (ours)



Seq2Seq Extractor (ours)



Datasets

Dataset	Genre	Train	Valid	Test	Refs
CNN/DailyMail	News	287,113	13,368	11,490	1
NYT	News	44,382	5,523	6,495	1.93
Reddit (Ouyang et al., 2017)	Personal Narratives	404	24	48	2
PubMed	Medical Journal Articles	21,250	1,250	2,500	1

Sizes of the training, validation, test splits for each dataset and the average number of test set human reference summaries per document.

Sentence Extractor Evaluation

Simpler extractors are just as good if not better!

	Rouge-2 Recall	
Sentence Extractor	**CNN/DM**	**NYT**
RNN	25.4	34.7
Seq2Seq	**25.6**	**35.7**
Cheng & Lapata	25.3	35.6
SummaRunner	25.4	35.4

Sentence Extractor Evaluation

Simpler extractors are just as good if not better!

Similar story on non-news datasets.

	Rouge-2 Recall	
Sentence Extractor	Reddit	PubMed
RNN	11.4	17.0
Seq2Seq	**13.6**	**17.7**
Cheng & Lapata	13.6	17.7
SummaRunner	13.4	17.2

Sentence Extractor Evaluation

Simpler extractors are just as good if not better!

Similar story on non-news datasets.

Sentence Extractor	**Rouge-2 Recall** **Reddit**	**PubMed**
RNN	11.4	17.0
Seq2Seq	**13.6**	**17.7**
Cheng & Lapata	13.6	17.7
SummaRunner	13.4	17.2

Similar results for choice of sentence encoder: averaging is as good as RNN and CNN encoders.

Additional experiments

What are these models learning?

How are important sentences identified? Lexical information?

Additional experiments

What are these models learning?

How are important sentences identified? Lexical information?

1. Word Embedding Fine Tuning:

Additional experiments

What are these models learning?

How are important sentences identified? Lexical information?

1. Word Embedding Fine Tuning:
   - No significant improvement!
   - In fact, worse performance on average (.3-.7 pts worse on news)

Additional experiments

What are these models learning?

How are important sentences identified? Lexical information?

1. Word Embedding Fine Tuning:
   - No significant improvement!
   - In fact, worse performance on average (.3-.7 pts worse on news)

2. POS class ablations: Remove nouns, verbs, adj/adv, and function words.

Additional experiments

What are these models learning?

How are important sentences identified? Lexical information?

1. Word Embedding Fine Tuning:
   - No significant improvement!
   - In fact, worse performance on average (.3-.7 pts worse on news)

2. POS class ablations: Remove nouns, verbs, adj/adv, and function words.
   - News datasets mostly unaffected (-0.1pt)
   - Reddit sees modest drop (-2pts) when removing adj./adv.

Additional experiments

What are these models learning?

How are important sentences identified? Lexical information?

1. Word Embedding Fine Tuning:
   - No significant improvement!
   - In fact, worse performance on average (.3-.7 pts worse on news)

2. POS class ablations: Remove nouns, verbs, adj/adv, and function words.
   - News datasets mostly unaffected (-0.1pt)
   - Reddit sees modest drop (-2pts) when removing adj./adv.

3. Sentence order shuffling:

Additional experiments

What are these models learning?

How are important sentences identified? Lexical information?

1. Word Embedding Fine Tuning:
   - No significant improvement!
   - In fact, worse performance on average (.3-.7 pts worse on news)

2. POS class ablations: Remove nouns, verbs, adj/adv, and function words.
   - News datasets mostly unaffected (-0.1pt)
   - Reddit sees modest drop (-2pts) when removing adj./adv.

3. Sentence order shuffling:
   - Large drops in performance on news and PubMed.

Shuffled vs In-Order

Ext.	Order	CNN/DM	NYT	Reddit	PubMed
Seq2Seq	In-Order	**25.6**	**35.7**	**13.6**	**17.7**
	Shuffled	21.7	25.6	**13.5**	14.9

Shuffled model is trained on shuffled sentence order documents.

Both models evaluated on in-order data.

Large **performance drops** on news and PubMed!

Shuffled vs In-Order

Ext.	Order	CNN/DM	NYT	Reddit	PubMed
Seq2Seq	In-Order	**25.6**	**35.7**	**13.6**	**17.7**
	Shuffled	21.7	25.6	**13.5**	14.9

Shuffled model is trained on shuffled sentence order documents.

Both models evaluated on in-order data.

Large **performance drops** on news and PubMed!

- Simple network architectures good enough.
- Without care, document structure dominates learning.

Deep Learning Models of Word Salience

Goal: Make lexical information more useful for DL models of sentence extractive summarization.

Why?

- Improve performance.
- Improve explanation.
- Improve generalizability.

Word Features

- Feature Embeddings
  - Shallow lexical semantics (Glove Embeddings)

Word Features

- Feature Embeddings
  - Shallow lexical semantics (Glove Embeddings)
  - Syntax
    - POS tag
    - Dependency role
    - Dependency depth (distance from root node)

Word Features

- Feature Embeddings
  - Shallow lexical semantics (Glove Embeddings)
  - Syntax
    - POS tag
    - Dependency role
    - Dependency depth (distance from root node)

  - Word Type/Topic
    - Named-Entity tag
    - LDA Topic/Brown Cluster

Word Features

- Feature Embeddings
  - Shallow lexical semantics (Glove Embeddings)
  - Syntax
    - POS tag
    - Dependency role
    - Dependency depth (distance from root node)

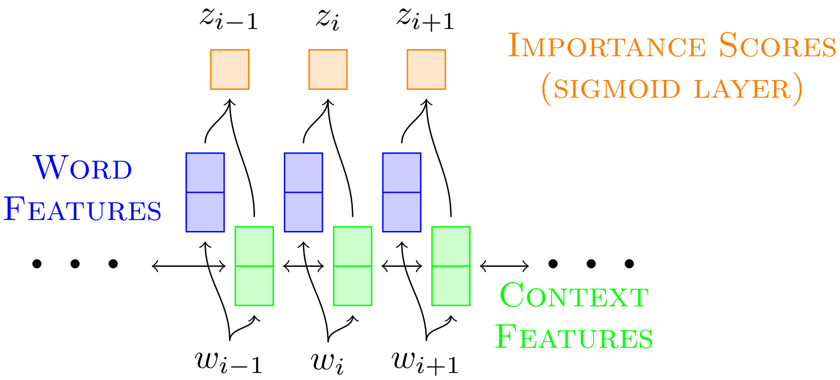
  - Word Type/Topic
    - Named-Entity tag
    - LDA Topic/Brown Cluster

  - Term Probability
  - Topic Signature
    - Log Likelihood Ratio (LLR): Log ratio of a term's document probability to background corpus probability.
    - Separate embeddings for words with highest LLR, and/or
    - separate embeddings for different thresholds of LLR.

Word Features

- Feature Embeddings
  - Shallow lexical semantics (Glove Embeddings)
  - Syntax
    - POS tag
    - Dependency role
    - Dependency depth (distance from root node)
  - Word Type/Topic
    - Named-Entity tag
    - LDA Topic/Brown Cluster
  - Term Probability
  - Topic Signature
    - Log Likelihood Ratio (LLR): Log ratio of a term's document probability to background corpus probability.
    - Separate embeddings for words with highest LLR, and/or
    - separate embeddings for different thresholds of LLR.
- Contextualized representation (Elmo Embeddings)

Proposed Model



Proposed Experiments

- Several ways to learn model:

Proposed Experiments

- Several ways to learn model:
    - Learn as part of a sentence extractve summarization task.

Proposed Experiments

- Several ways to learn model:
    - Learn as part of a sentence extractve summarization task.
        - Margin based learning objective with greedy inference.

Proposed Experiments

- Several ways to learn model:
    - Learn as part of a sentence extractve summarization task.
        - Margin based learning objective with greedy inference.
        - Explore other combinatorial optimization algorithms, e.g. dynamic programming or ILP.

Proposed Experiments

- Several ways to learn model:
  - Learn as part of a sentence extractve summarization task.
    - Margin based learning objective with greedy inference.
    - Explore other combinatorial optimization algorithms, e.g. dynamic programming or ILP.
  - Directly supervision using reference summaries to classify words as in or out of the summary.

Proposed Experiments

- Learn word importance scores on large news datasets:

Proposed Experiments

- Learn word importance scores on large news datasets:
  - Newsroom (Grusky et al. 2018)
    - 1.3 million single document/summary pairs from news domain.
    - Filterable subsections: extractive, abstractive, mixed.

Proposed Experiments

- Learn word importance scores on large news datasets:
  - Newsroom (Grusky et al. 2018)
    - 1.3 million single document/summary pairs from news domain.
    - Filterable subsections: extractive, abstractive, mixed.
  - XSUM (Narayan et al. 2018)
    - 200k single document/summary pairs from news domain.
    - More abstractive than previous datasets;
    - requires combining information from multiple parts of the document.

Other experiments/Plan B's

- Evaluate explainability: compare human judgements of word importance to learned predictions
- Genre Adaptation
  - E.g. train adj/adv only news model and evaluate on Reddit.
- Domain Adaptation to Multi-document Summarization
  - Encode documents individually
  - Aggregate importance scores using cross document attention

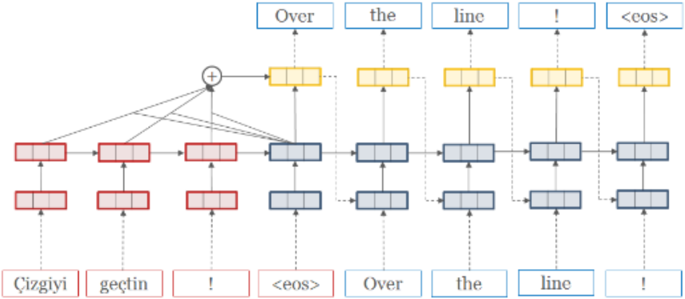
Takeaways

- Simple network architectures good enough.
- Without care, document structure dominates learning.
- To learn more from content, we propose to design better word representations.

Talk Outline

1. Feature Based Models of Sentence Salience
   - Stream Summarization
   - Learning-to-Search Summarizer
2. Deep Learning Models of Salience
   - Sentence Salience
   - Word Salience
3. Faithful Generation
4. Research Plan and Contributions

Sequence-to-Sequence Models for Abstractive Summarization (See et al. 2017)



- Encoder + Decoder Attention must **identify important content**.
- Decoder must learn to generate fluent and **correct** output.

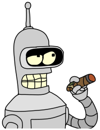
Better Content Selection for Seq2Seq Models

- Help decoder attention with word importance model:
  - Auxilliary Objective to encourage attention distribution to match word importance score distribution.
  - Redact or gate words from decoder with low importance.
  - Training word importance model end-to-end with seq2seq model.

Hallucination in Seq2Seq Models

Lagos, Nigeria (CNN) — A day after winning Nigeria's presidency, Muhammadu Buhari told CNN's Christiane Amanpour that he plans to aggressively fight corruption that has long plagued Nigeria and go after the root of the nation's unrest…

Input Sequence

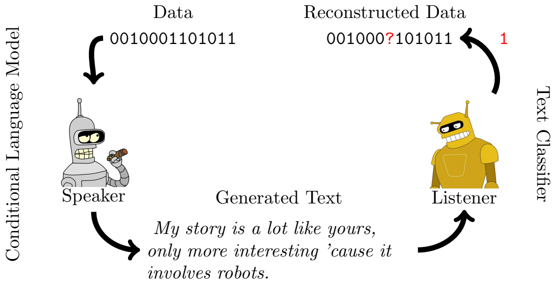


Output Sequence

Conditional Language Model

Muhammadu Buhari says his administration is confident it will be able to destabilize Nigeria's economy.

Faithful Generation



Motivation:

- The speaker generates an utterance from input data.
- The listener tries to reconstruct the input data.
- The speaker learns directly from the listeners feedback, ensuring that it prefers output likely to be true (w.r.t. the input data).

Faithful Generation Training

Given pairs of input data x and output text y:

- Pre-train the speaker to generate text from input data: $\max p(y|x)$.

Faithful Generation Training

Given pairs of input data x and output text y:

- Pre-train the speaker to generate text from input data: $\max p(y|x)$.

- Train listener to predict input from from text y: $\max q(x|y)$.

Faithful Generation Training

Given pairs of input data x and output text y:

- Pre-train the speaker to generate text from input data: $\max p(y|x)$.
- Train listener to predict input from from text y: $\max q(x|y)$.
- Augment maximum likelihood objective with auxilliary objective to maximize the correctness of the input reconstruction under the listener: $\max \mathbb{E}_{y \sim p(\cdot|x)}[q(x|y)]$.

Faithful Generation Training

Given pairs of input data x and output text y:

- Pre-train the speaker to generate text from input data: $\max p(y|x)$.
- Train listener to predict input from from text y: $\max q(x|y)$.
- Augment maximum likelihood objective with auxilliary objective to maximize the correctness of the input reconstruction under the listener: $\max \mathbb{E}_{y \sim p(\cdot|x)}[q(x|y)]$.
- We can apply this objective to entire beam search to encourage diverse but accurate generation outputs.

Faithful Generation Training

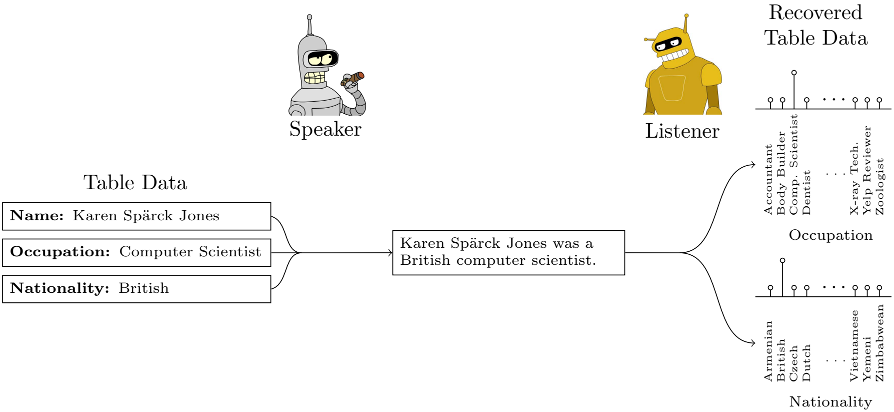
Given pairs of input data x and output text y:

- Pre-train the speaker to generate text from input data: $\max p(y|x)$.
- Train listener to predict input from from text y: $\max q(x|y)$.
- Augment maximum likelihood objective with auxilliary objective to maximize the correctness of the input reconstruction under the listener: $\max \mathbb{E}_{y \sim p(\cdot|x)}[q(x|y)]$.
- We can apply this objective to entire beam search to encourage diverse but accurate generation outputs.
- We can use the listener to give our confidence in the correctness of outputs.

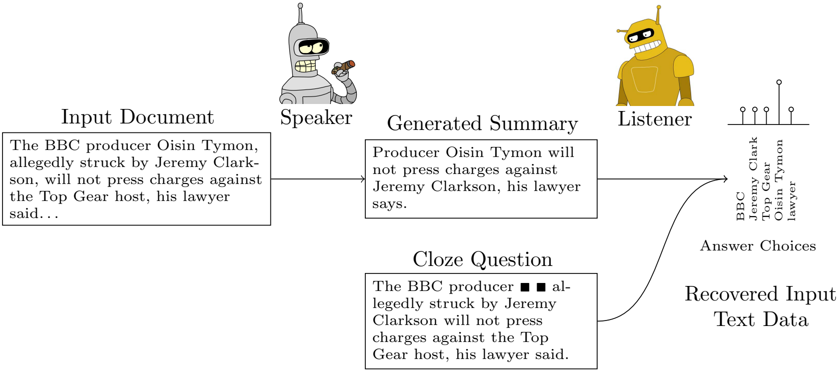
Two Applications

- **Data-to-Text**
  - Table data $\rightarrow$ text description $\rightarrow$ reconstruct table
- **Text-to-Text**
  - Document text $\rightarrow$ text summary $\rightarrow$ answer cloze style questions from document

Data-to-Text


Speaker
Listener
Recovered
Table Data
Table Data
Name: Karen Spärck Jones
Occupation: Computer Scientist
Nationality: British
Karen Spärck Jones was a British computer scientist.
Accountant
Body Builder
Comp. Scientist
Dentist
. . .
X-ray Tech.
Yelp Reviewer
Zoologist
Occupation
Armenian
British
Czech
Dutch
. . .
Vietnamese
Yemeni
Zimbabwean
Nationality

Text-to-Text


Input Document
The BBC producer Oisin Tymon, allegedly struck by Jeremy Clarkson, will not press charges against the Top Gear host, his lawyer said. . .
Speaker
Generated Summary
Producer Oisin Tymon will not press charges against Jeremy Clarkson, his lawyer says.
Listener
BBC
Jeremy Clark
Top Gear
Oisin Tymon
lawyer
Answer Choices
Cloze Question
The BBC producer ■ ■ allegedly struck by Jeremy Clarkson will not press charges against the Top Gear host, his lawyer said.
Recovered Input Text Data

Planned Experiments

- Data-to-Text
  - E2E Dataset (Novikova et al. 2017)
    - Artificially constructed dataset of restaurant data and descriptions
    - 50k Meaning Representations/text description pairs.

Planned Experiments

- Data-to-Text
  - E2E Dataset (Novikova et al. 2017)
    - Artificially constructed dataset of restaurant data and descriptions
    - 50k Meaning Representations/text description pairs.
  - WikiBio Datatest (Lebret et al. 2017)
    - Biographical data paired with text descriptions, taken from Wikipedia.
    - 700k table/text pairs.

Planned Experiments

- Data-to-Text
  - E2E Dataset (Novikova et al. 2017)
    - Artificially constructed dataset of restaurant data and descriptions
    - 50k Meaning Representations/text description pairs.
  - WikiBio Datatest (Lebret et al. 2017)
    - Biographical data paired with text descriptions, taken from Wikipedia.
    - 700k table/text pairs.
- Text-to-Text

Planned Experiments

- Data-to-Text
  - E2E Dataset (Novikova et al. 2017)
    - Artificially constructed dataset of restaurant data and descriptions
    - 50k Meaning Representations/text description pairs.
  - WikiBio Datatest (Lebret et al. 2017)
    - Biographical data paired with text descriptions, taken from Wikipedia.
    - 700k table/text pairs.
- Text-to-Text
  - TL;DR Dataset (Völske et al., 2017)
    - ≈ 4 million Reddit comments with summaries.
    - Non-news dataset!
    - Comments shorter than news article: possibly easier to generate cloze questions.

Planned Experiments

- Data-to-Text
  - E2E Dataset (Novikova et al. 2017)
    - Artificially constructed dataset of restaurant data and descriptions
    - 50k Meaning Representations/text description pairs.
  - WikiBio Datatest (Lebret et al. 2017)
    - Biographical data paired with text descriptions, taken from Wikipedia.
    - 700k table/text pairs.
- Text-to-Text
  - TL;DR Dataset (Völske et al., 2017)
    - ≈ 4 million Reddit comments with summaries.
    - Non-news dataset!
    - Comments shorter than news article: possibly easier to generate cloze questions.
  - Lots of news (CNN/DM, NYT, Newsroom, XSUM)

Planned Experiments

- REINFORCE (Williams, 1992) style learning objective to maximize correct classification by the listener.

Planned Experiments

- REINFORCE (Williams, 1992) style learning objective to maximize correct classification by the listener.
- While incorrect statements in best beam candidate might be rare, errors more likely in remainder of beam.

Planned Experiments

- REINFORCE (Williams, 1992) style learning objective to maximize correct classification by the listener.
- While incorrect statements in best beam candidate might be rare, errors more likely in remainder of beam.

$\Rightarrow$ Optimizing over whole beam should be easier to demonstrate improvements.

Planned Experiments

- Interesting angles to take even if performance improvements are not staggering:
    - Apply listener as beam re-ranking criterion during generation.

Planned Experiments

- Interesting angles to take even if performance improvements are not staggering:
  - Apply listener as beam re-ranking criterion during generation.
  - Localize error signals with token level explanations from classifier.

Planned Experiments

- Interesting angles to take even if performance improvements are not staggering:
  - Apply listener as beam re-ranking criterion during generation.
  - Localize error signals with token level explanations from classifier.
  - Understand correlation in listener models $\Rightarrow$ enforce independent listener models.

Planned Experiments

- Interesting angles to take even if performance improvements are not staggering:
  - Apply listener as beam re-ranking criterion during generation.
  - Localize error signals with token level explanations from classifier.
  - Understand correlation in listener models $\Rightarrow$ enforce independent listener models.

- We can also focus on the cloze question generation aspect.
  - Many heuristics for creating cloze style questions.
  - Incorporate word importance model.
  - Guided question generation to improve training.

Talk Outline

1. Feature Based Models of Sentence Salience
   - Stream Summarization
   - Learning-to-Search Summarizer
2. Deep Learning Models of Salience
   - Sentence Salience
   - Word Salience
3. Faithful Generation
4. Research Plan and Contributions

Research Plan

Task	Date
Faithful Gen. Impl.	December-February 2019
Auto and Human evaluation	February 2019 - March 2019
Word Importance (SDS)	April 2019 - May 2019
Word Importance (MDS/Genre)	July 2019
Word Importance (Abstractive)	June 2019
Write Thesis	August 2019 - February 2020
¡Defend!	March 2020

Contributions

- **Salience Estimation**
  - ✓ Two state-of-the-art sentence extractive stream summarization models. (TREC ‘14, ACL ‘15, TREC ‘15, IJCAI ‘16)
  - ✓ State-of-the-art deep learning based sentence extractive summarizarization models. (EMNLP ‘18)
  - ✓ Extensive ablation studies to determine important lexical/structural features for learning. (EMNLP ‘18)
  - A deep learning model of word importance estimation for single-document news summarization.
  - Adaptation of the word importance model to non-news genre and multi-document summarization.

Contributions

- **Faithful Generation**
  - Supervised attention with word importance estimation.
  - Round-trip Speaker/Listener learning model for:
    - data-to-text generation
    - text-to-text generation/abstractive summarization.