Thesis Proposal

Chris Kedzie

Columbia University

kedzie@cs.columbia.edu

December 24, 2018

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.

Contributions: Salience Estimation

- **Feature Based Models of Salience**
  - Incorporate salience regression with biased clustering.
  - Incorporate salience regression with exploration (learning to search).
  - Evaluation in Stream Summarization task.

- **Deep Learning Models of Salience**
  - Several novel deep learning models of sentence and word salience.
  - Extensive ablation studies to determine important lexical/structural features for learning.
  - Model evaluation on Single Document Summarization task
  - Domain adaption experiments to Multi-Document Summarization

Contributions: Faithful Generation

- Text generation is treated as a two player game between, i.e. a speaker and listener.
- The speaker generates a summary from the input.
- The listener uses the summary to reconstruct parts of the input.

Talk Outline

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

Feature Based Models of Sentence Salience

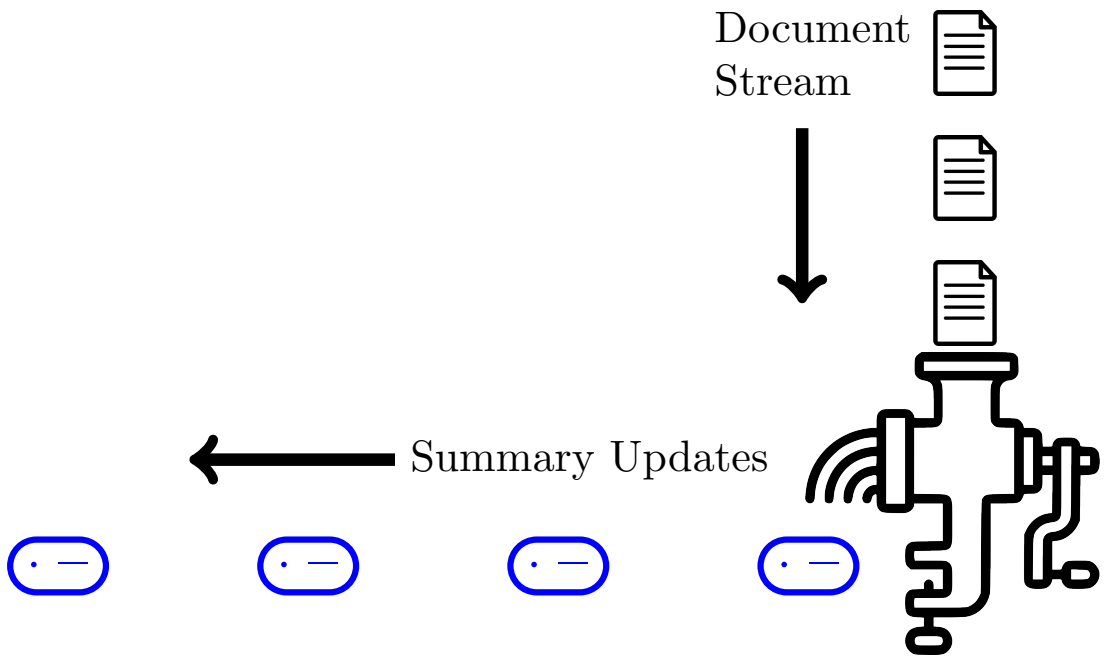
Two models for sentence extractive summarization:

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.

Model evaluation on stream summarization task.

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

Queries and Nuggets

TREC data included event queries, i.e. real life natural and man-made disasters/crises.
E.g. "Hurricane Sandy," "Boston Marathon Bombing"

No reference summaries, but reference *nuggets*:

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

Sentence Salience for Stream Summarization

- Let $S(q)$ be the ordered sequence of sentences from the relevant document stream for query q.
- Let $\mathcal{N}(q)$ be the set of nuggets for query q.
- **Salience** y of a sentence $s \in S(q)$ is computed:

$$y = \max_{n \in \mathcal{N}(q)} \textsc{Similarity}(s, n)$$

where $\textsc{Similarity}(\cdot, \cdot)$ is the semantic similarity method of Guo and Diab, 2012.

Sentence Salience for Stream Summarization

- Let $S(q)$ be the ordered sequence of sentences from the relevant document stream for query q.
- Let $\mathcal{N}(q)$ be the set of nuggets for query q.
- **Salience** y of a sentence $s \in S(q)$ is computed:

$$y = \max_{n \in \mathcal{N}(q)} \textsc{Similarity}(s, n)$$

where $\textsc{Similarity}(\cdot, \cdot)$ is the semantic similarity method of Guo and Diab, 2012.

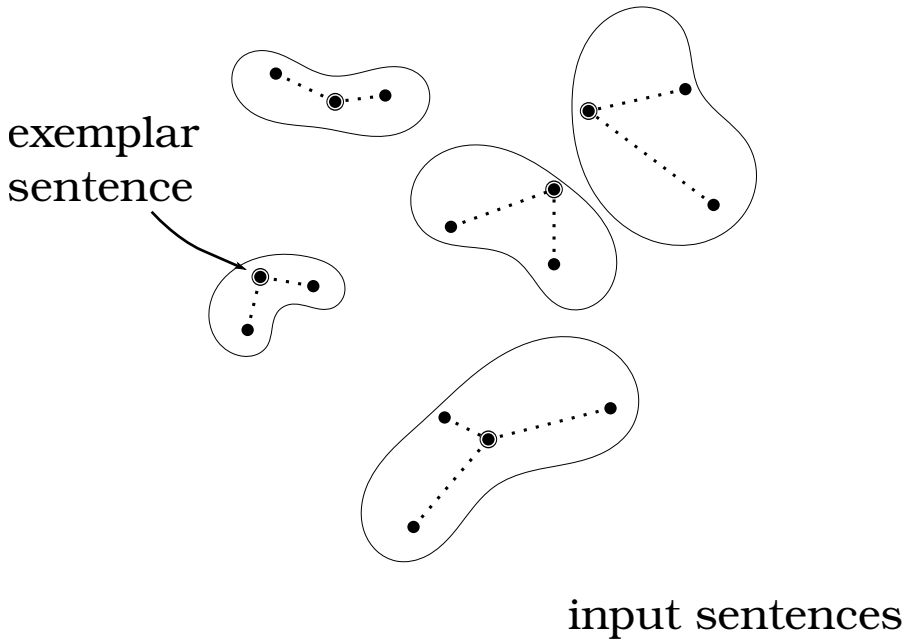
- No nugget knowledge at test time. Learn to predict $p(\hat{y}|s, q)$.

Features

- Surface Features
- Query Features
- Language Model Scores
- Geographic Relevance (SAP)
- Temporal Relevance (SAP)
- Document Frequency (L2S)
- Stream Language Model Scores (L2S)
- Update Similarity (L2S)
- Nugget Probability (L2S)
- Single Document Summarization Rankings (L2S)

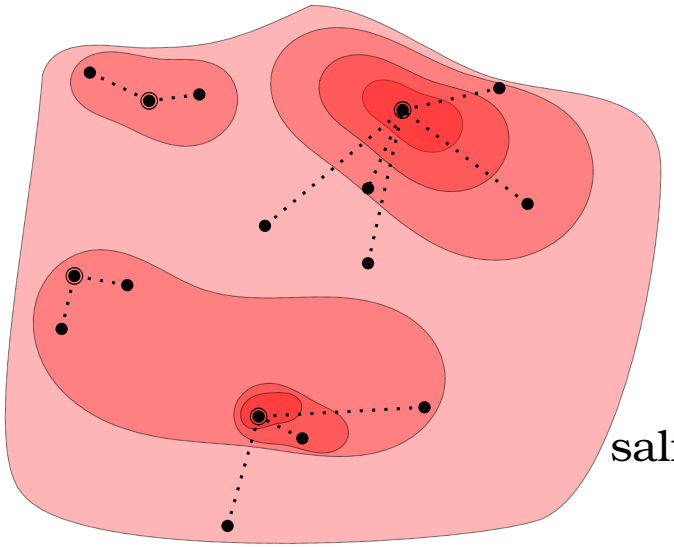
Salience Biased Affinity Propagation

- **Affinity Propagation Clustering**
  - Exemplars (and clusters) determined by similarity.



Salience Biased Affinity Propagation

- **Affinity Propagation Clustering**
  - Exemplars (and clusters) determined by similarity.
- **Salience-biased Affinity Propagation Clustering**
  - Exemplars likely to have higher predicted salience $\hat{y} \sim p(\cdot|s, q)$



Experiments

- Leave-One-Out evaluation
  - 21 TREC 2014 Temporal Summarization events
- 3 events held out for tuning similarity threshold parameters
- trained salience models for each event
  - 1000 sentences sampled for each event
  - At test time, salience prediction is the average of the 20 other models

Evaluation Metrics

- ROUGE
  - "reference" summary generated by concatenating event nugget texts
- TREC Temporal Summarization metrics
  - **Expected Gain** — the average number of novel nuggets in each extracted sentence; $\approx$ nugget precision.
  - **Comprehensiveness** — nugget recall.

Baselines

- SAP – full model: salience-biased affinity propagation
- AP – affinity propagation clustering with no salience
- HAC – hierarchical agglomerative clustering
- RS – rank by salience

Rouge

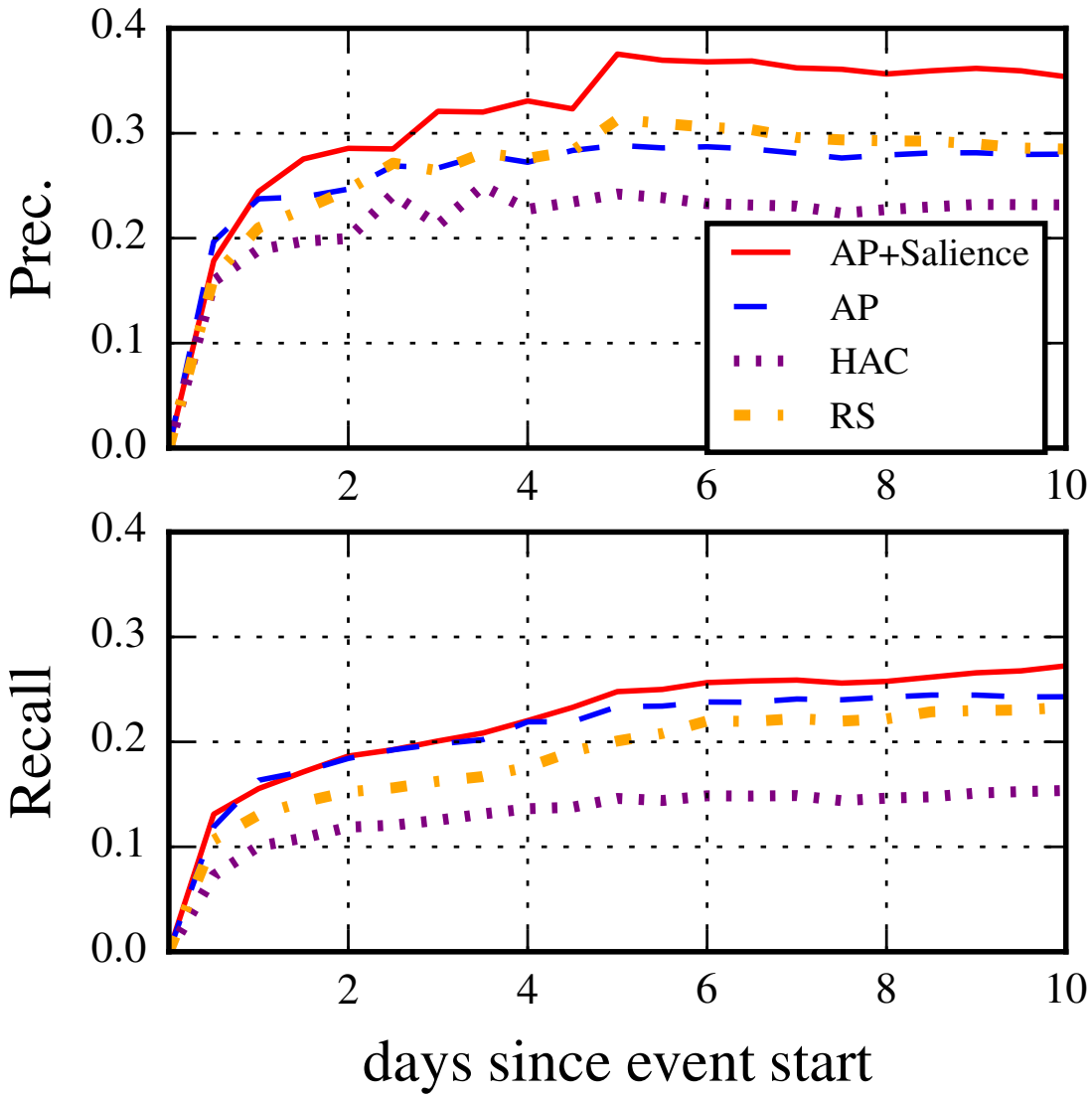
Rouge-1

System	Recall	Prec.	F_1
SAP	**0.282**	**0.344**	**0.306**
AP	0.245	0.285	0.263
RS	0.230	0.271	0.247
HAC	0.169	0.230	0.186

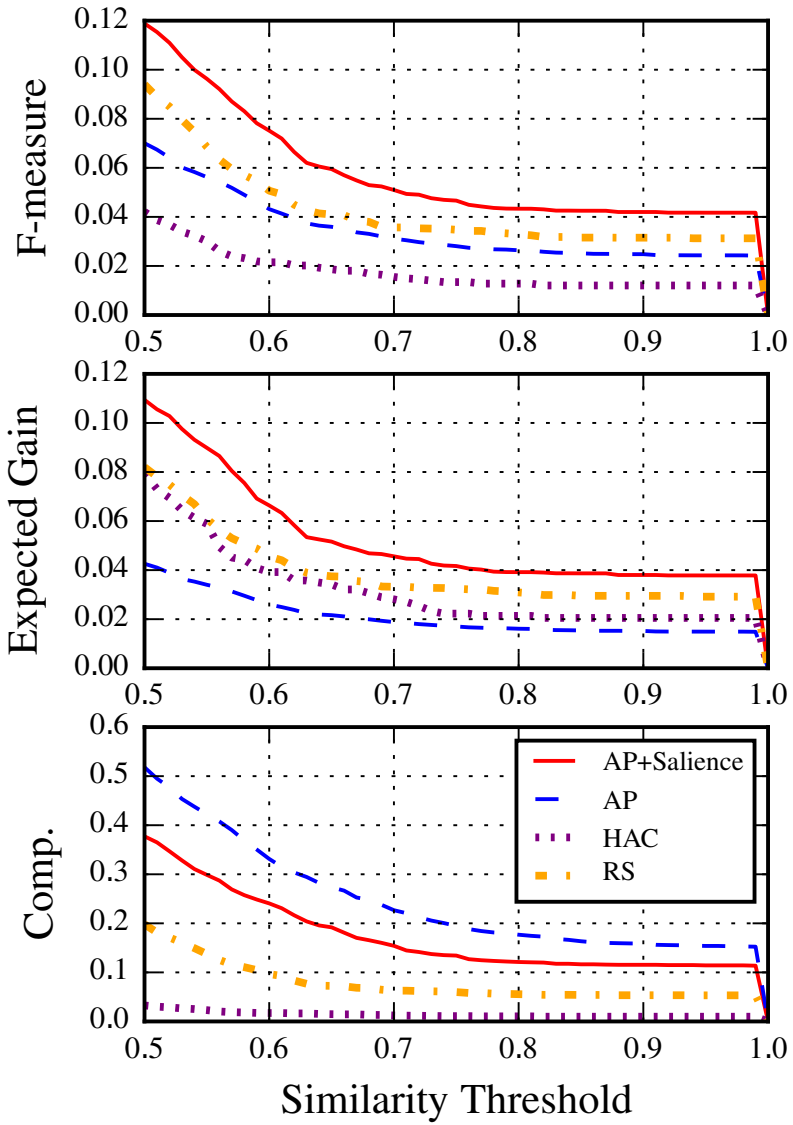
Rouge-2

System	Recall	Prec.	F_1
SAP	**0.045**	**0.056**	**0.049**
AP	0.033	0.038	0.035
RS	0.031	0.037	0.034
HAC	0.017	0.024	0.019

ROUGE-1 over time



$\mathbb{E}$[Gain] and Comprehensiveness

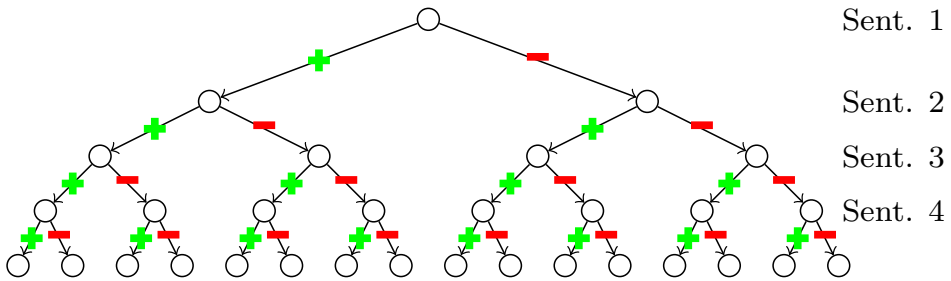


Learning-to-Search Summarization

- SAP 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!

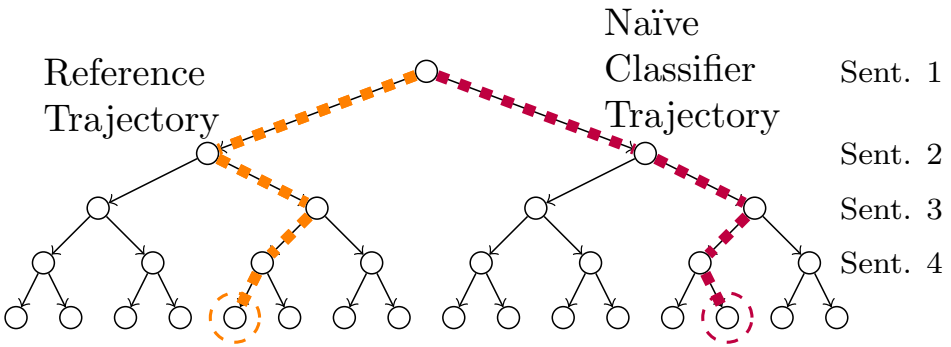
Learning-to-Search Summarization

- SAP 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!



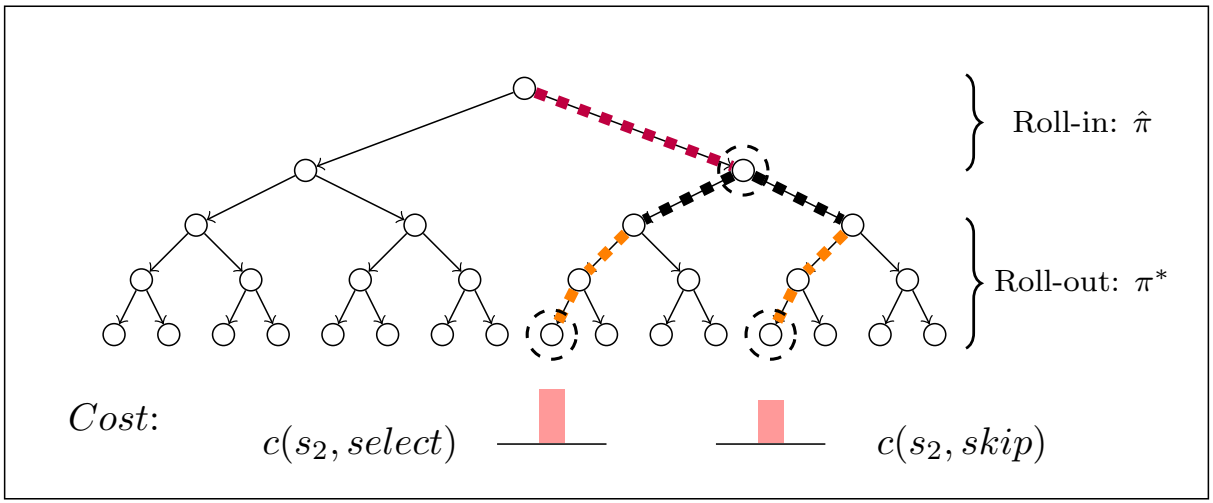
Learning-to-Search Summarization

- SAP 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!



Learning-to-Search Summarization

- SAP 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.

System Comparisons

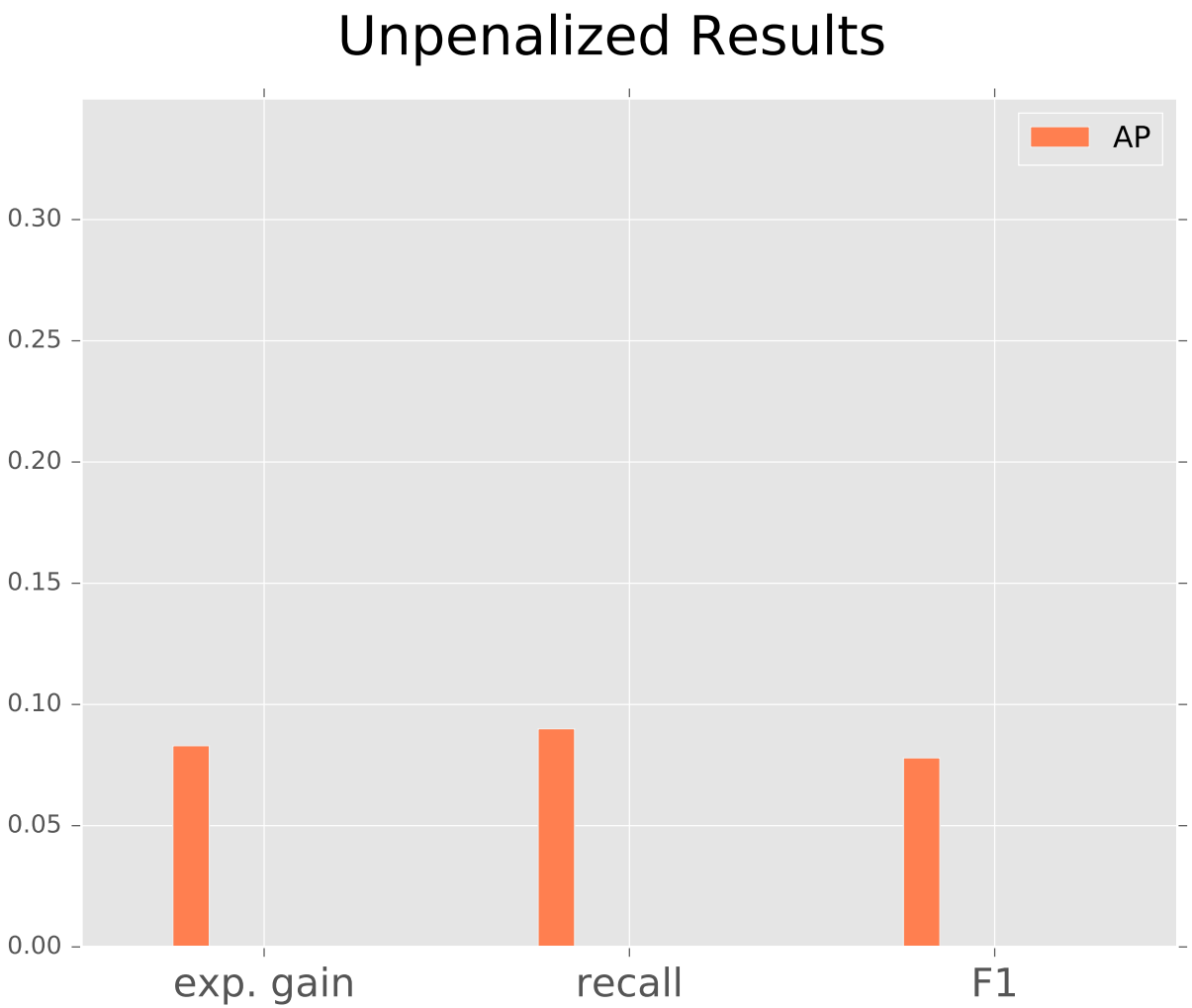
Baselines

- **CosSim** — 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
- **SAP** — Salience Biased Affinity Propagation Clustering

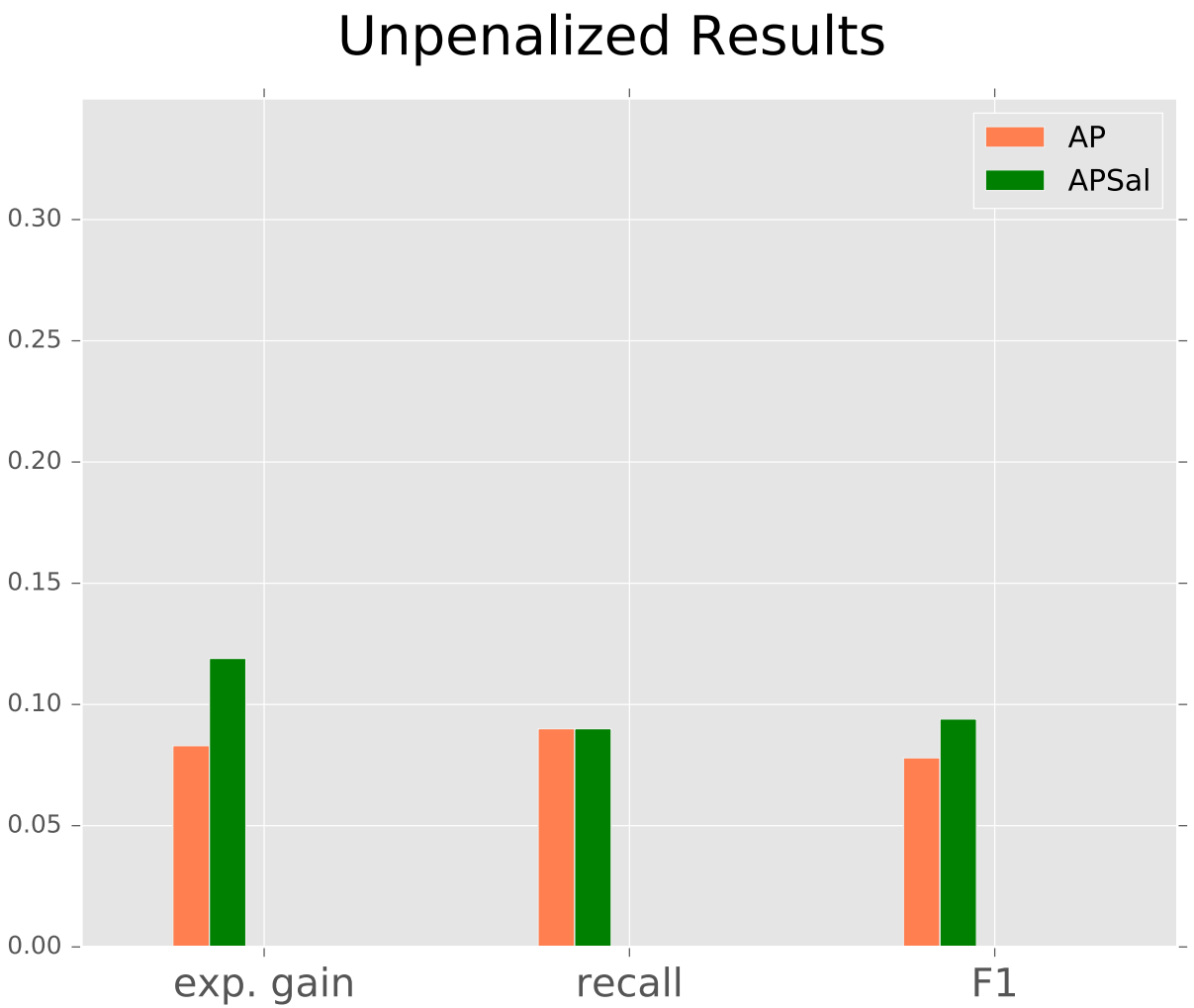
Our Models

- **L2S** — learning to search model
- **L2SCos** — learning to search model with similarity threshold

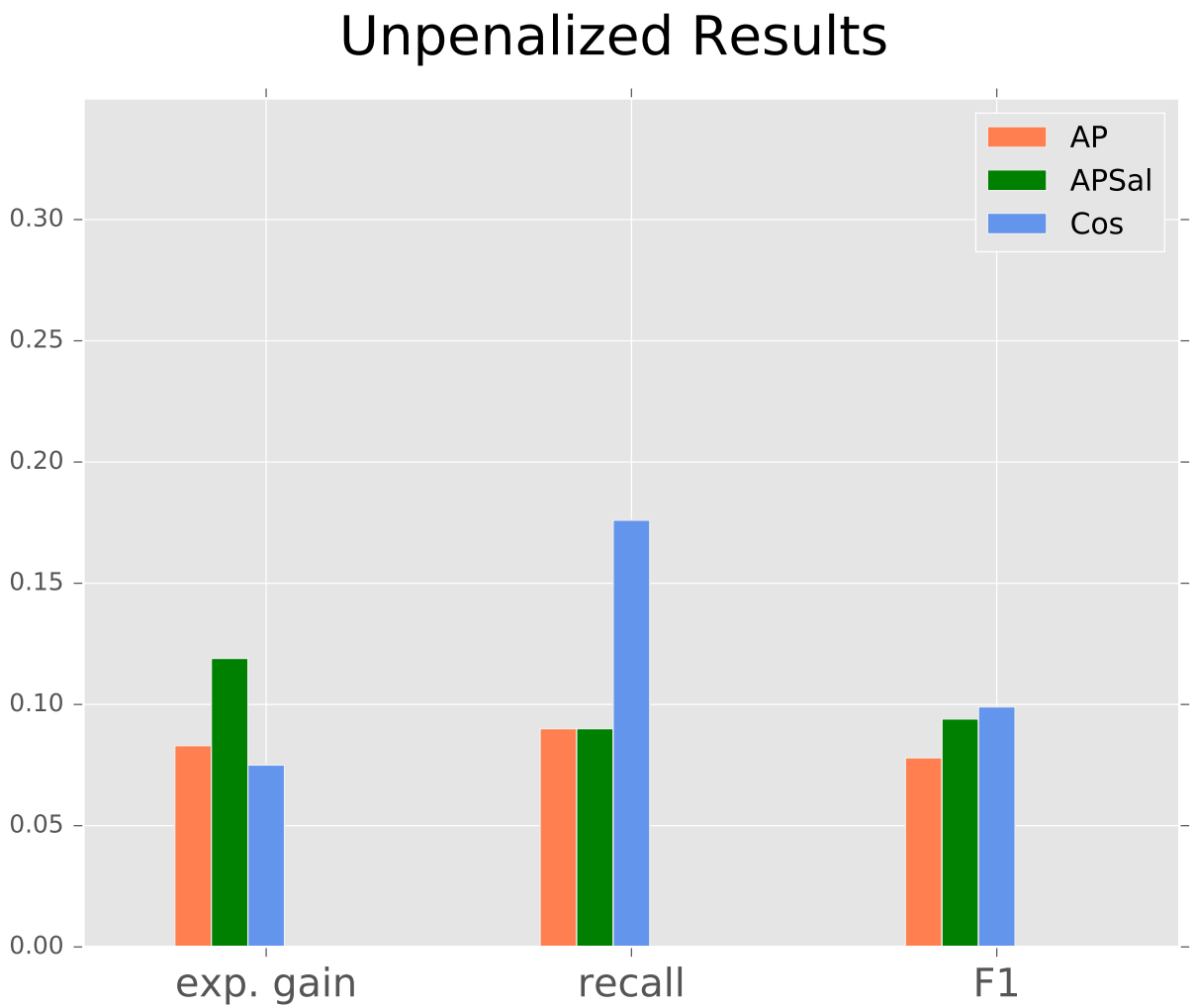
Results


Unpenalized Results
AP
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exp. gain
recall
F1

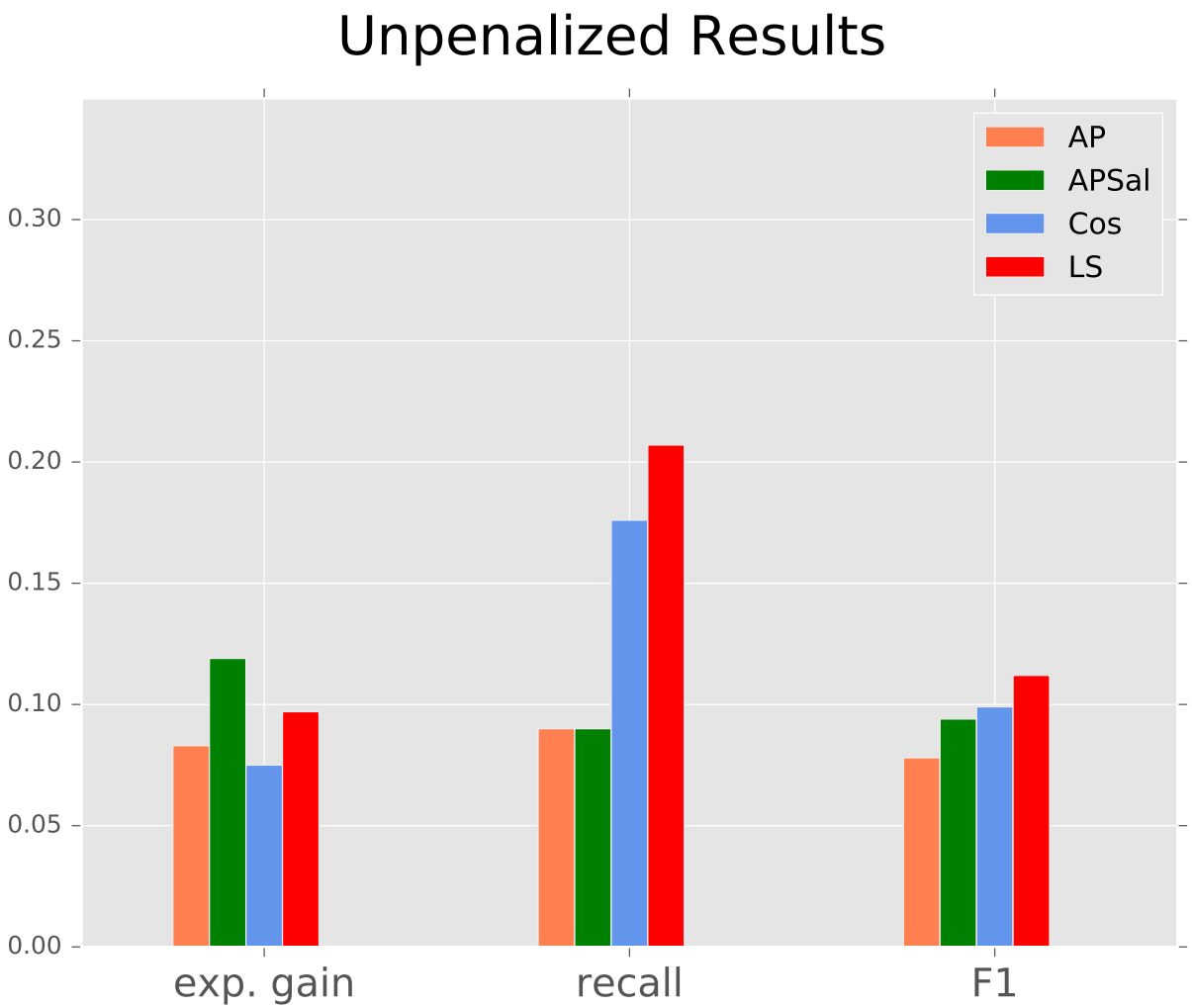
Results


Unpenalized Results
AP
APSal
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exp. gain
recall
F1

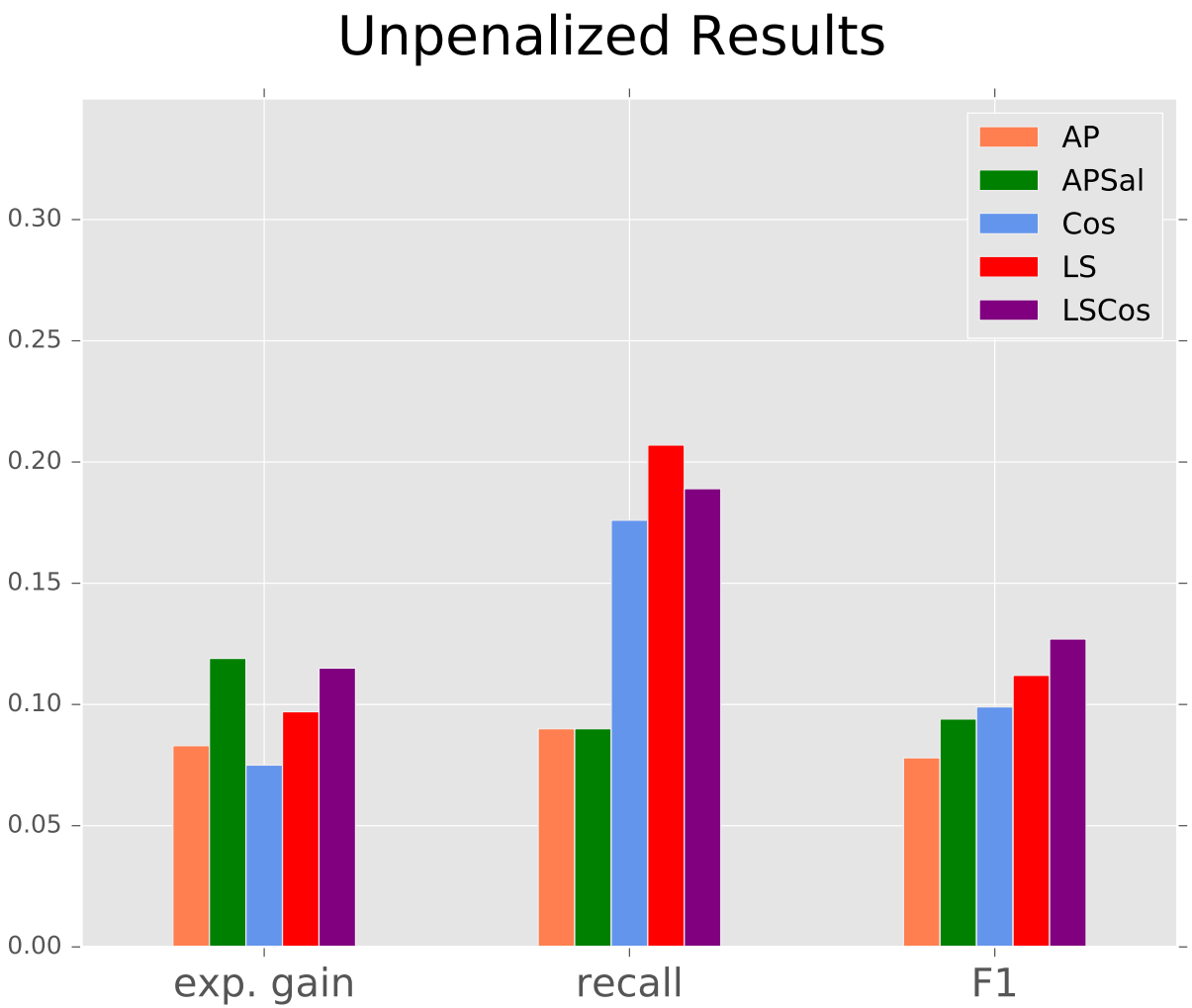
Results


Unpenalized Results
AP
APSal
Cos
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exp. gain
recall
F1

Results


Unpenalized Results
AP
APSal
Cos
LS
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exp. gain
recall
F1

Results


Unpenalized Results
AP
APSal
Cos
LS
LSCos
0.30
0.25
0.20
0.15
0.10
0.05
0.00
exp. gain
recall
F1

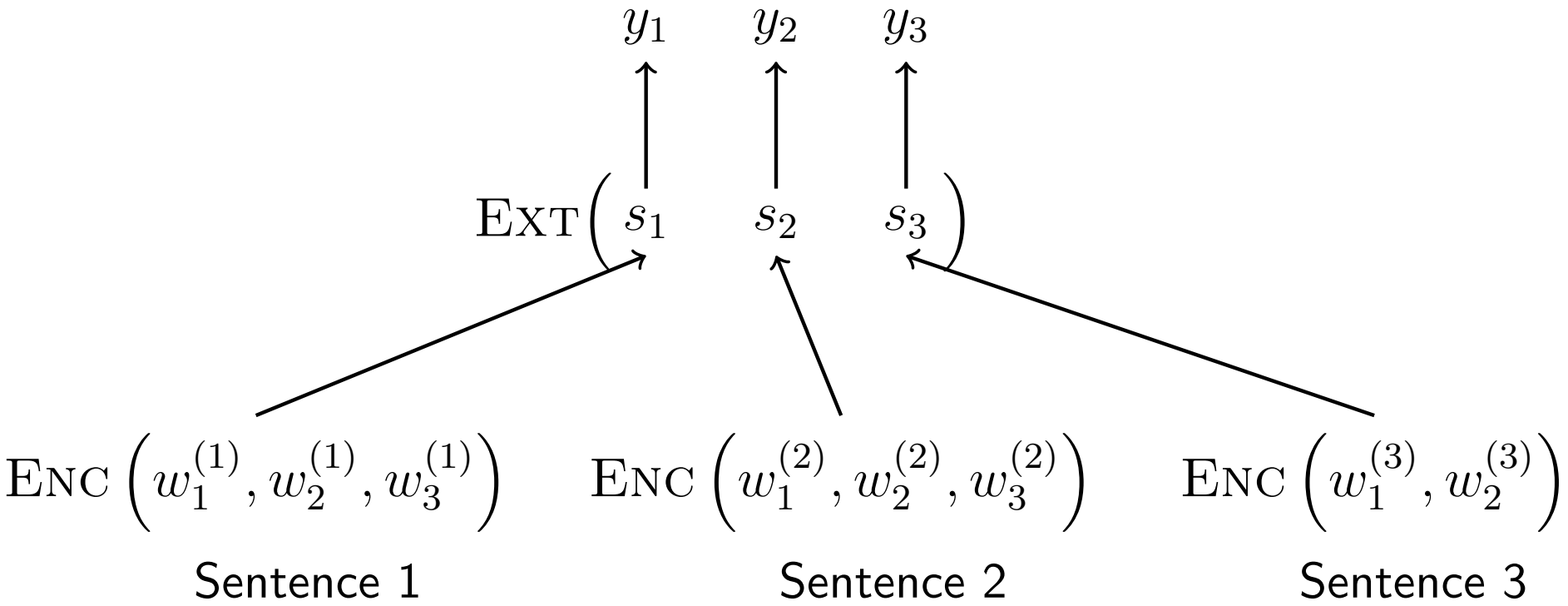
Talk Outline

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

Sentence Salience

- Lots of recent work on deep nets for sentence extractive news summarization. (some citances)
- Salience is modeled as a classification problem, i.e. salience = probability of including a sentence in the extract summary.
- Multiple architecture tweaks in each work, difficult to tell what is actually driving improvements.

Summarizer Architecture

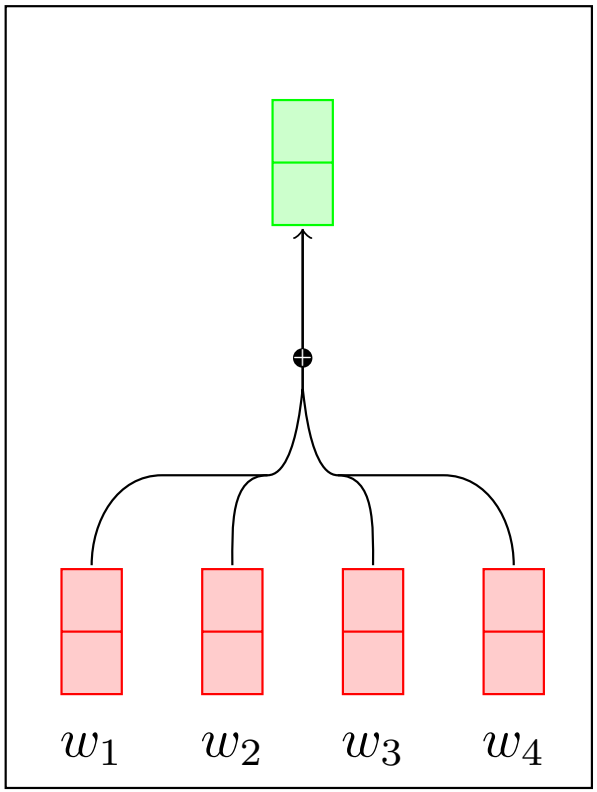


Sentence Salience

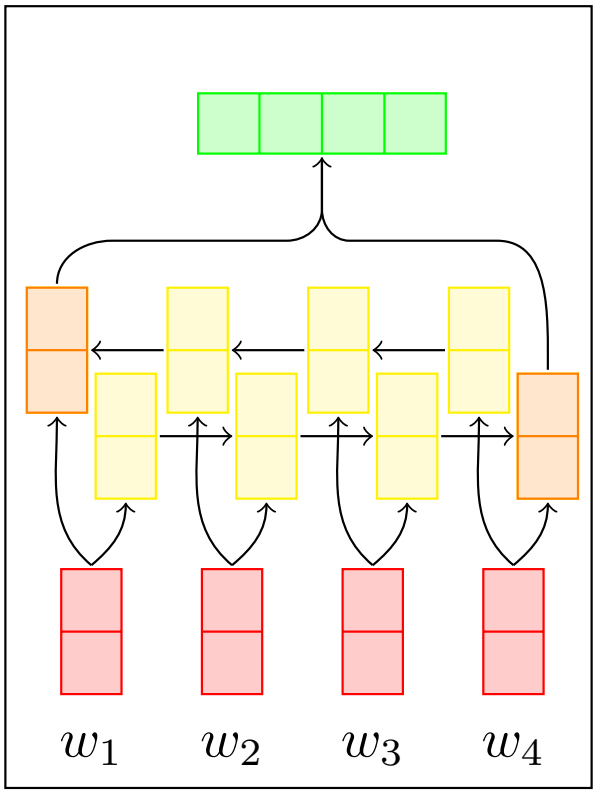
- We experiment with several popular sentence embedding methods.

Sentence Encoders

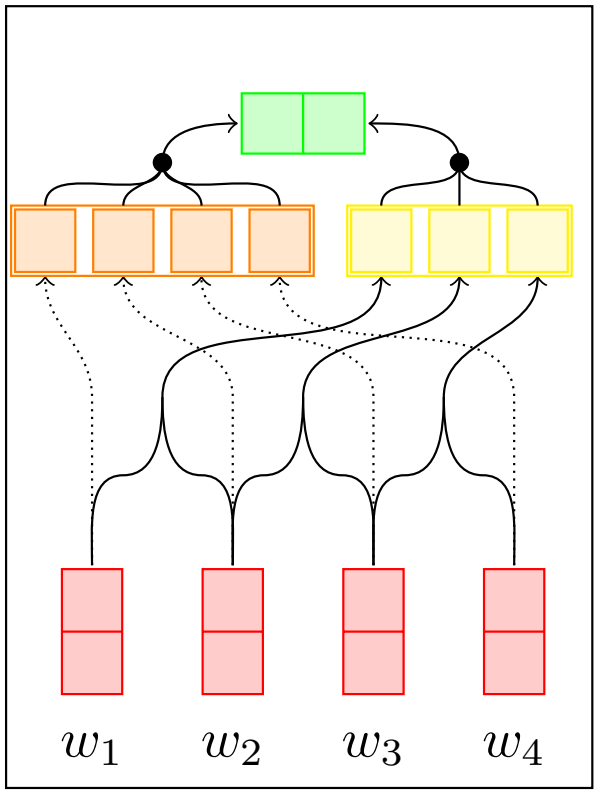
Averaging Encoder



RNN Encoder



CNN Encoder



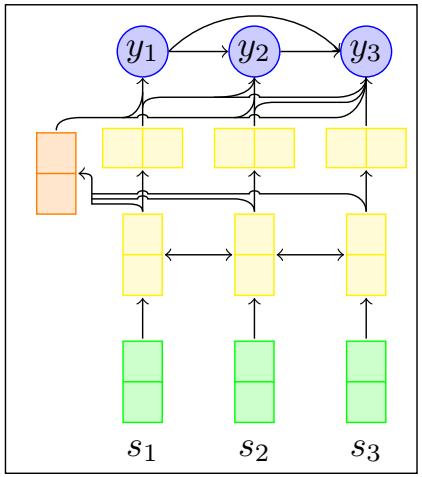
We use pretrained (Wikipedia/Gigaword) Glove word embeddings.

Sentence Salience

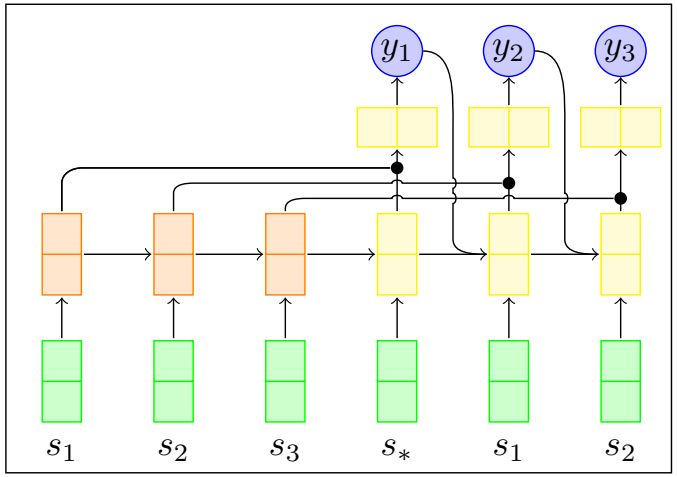
- We experiment with several popular sentence embedding methods.
- We experiment with two state of art sentence classification architectures (Nallapati et al., 2016) and (Cheng & Lapata, 2016), and
- propose simplified versions of each (RNN and Seq2Seq, respectively).

Sentence Classifiers

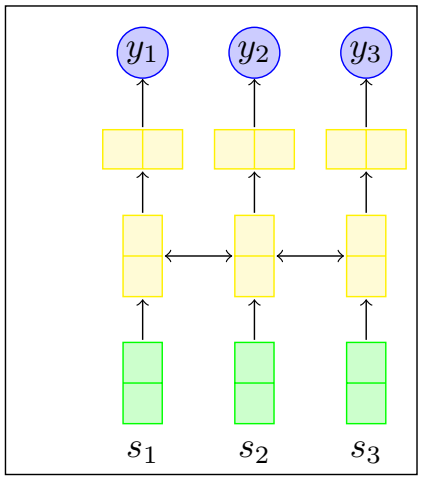
SummaRunner Extractor (Nallapati et al. 2016)



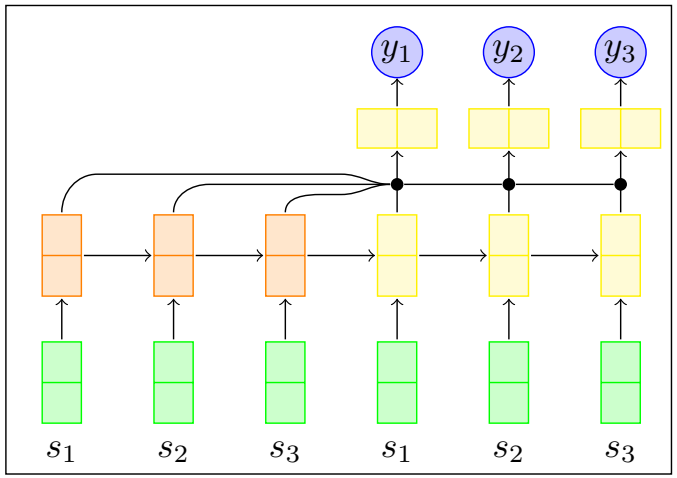
Cheng & Lapata Extractor (Cheng and Lapata, 2016)



RNN Extractor (ours)



Seq2Seq Extractor (ours)



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

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!

Deep Learning Models of Word Salience

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

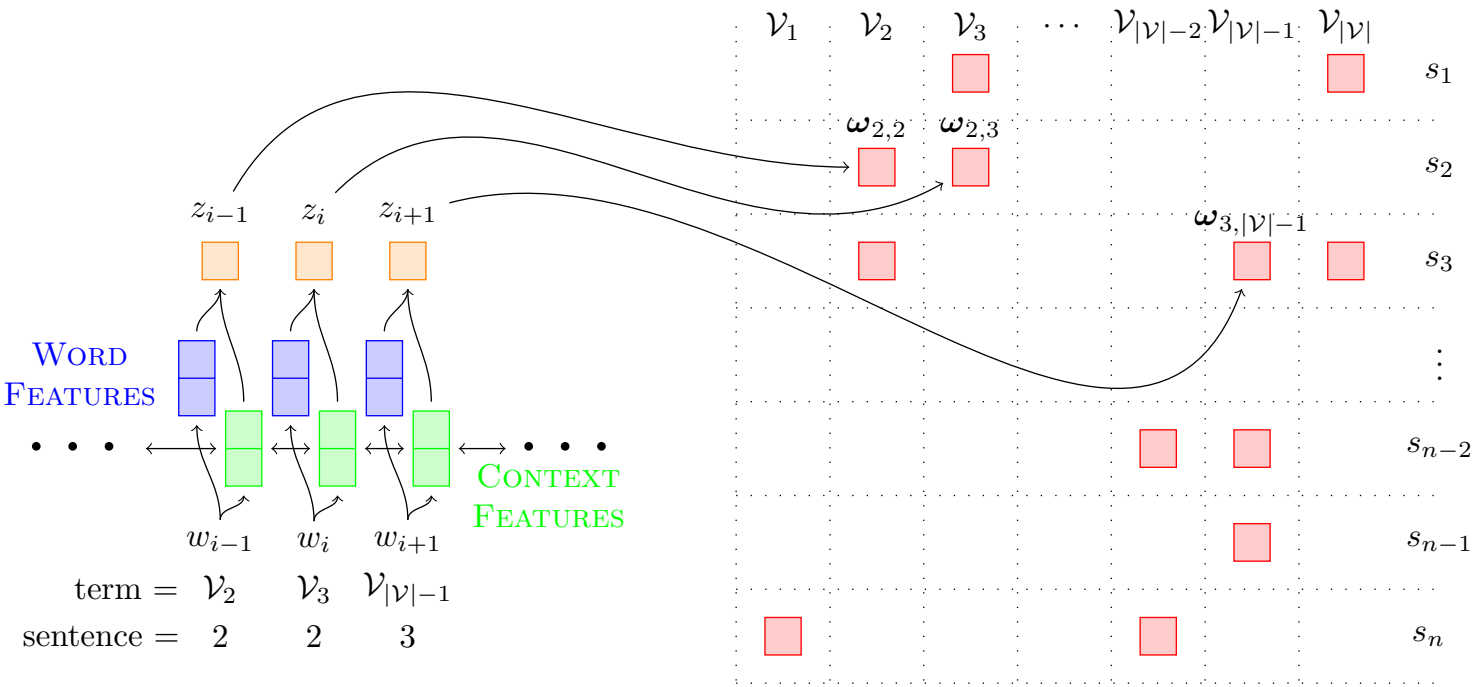
Why?

- Improve explanation.
- Improve generalizability I (to Multi-Doc)
- Improve generalizability II (to Abstractive Summarization)

Word Features

- Feature Embeddings
  - Shallow lexical semantics (Glove Embeddings)
  - Term Frequency
  - Topic Signature
  - POS tag
  - Named-Entity tag
  - Dependency role
  - Dependency depth (distance from root node)
  - LDA Topic/Brown Cluster
- Contextualized representation (Elmo Embeddings)

Proposed Model



$\omega \triangleq$ Sentence $\times$ Term matrix of word salience weights, e.g. $\omega_{i,j}$ is the importance score of term j in sentence i

Proposed Experiments

- Learn word importance scores on large news datasets (CNN/DM, NYT, Newsroom, XSUM)
- Many ways to construct a summary from ω.
- To start, greedy maximization of sum of importance scores, and
- large margin learning frame work.

Large Margin Learning

$\omega \triangleq$ Sentence $\times$ Term matrix of word salience weights, e.g. $\omega_{i,j}$ is the importance score of term j in sentence i

$y \in \mathbb{Z}^n$ is an extractive reference summary of n sentences.

$\eta \triangleq \sum_{j \in \{1,\ldots,|\mathcal{V}|\}} \max_{i \in y} \omega_{i,j}$, the score for the reference summary.

$\hat{y}, \hat{\eta}$ predicted summary indices and score.

$\mathcal{L}(y, \hat{y}) = \max(0, 1 + \hat{\eta} - \eta)$

Other experiments/Plan B's

- We can supervise the individual importance scores using the reference abstracts to get labels, i.e. $z_i \rightarrow 1$ if w_i occurs in the reference summary.
- More sophisticated inference, e.g. knapsack packing.
- Matching performance of sentence extractive models ok if word level scores gives additional explanability.
- We can experiment with selective word class masking for domain adaptation, i.e. train Adj./Adv. only model (with no position) on news and evaluate on Reddit.)

Generalize to multi-document summarization

1. For each document $d \in \{1, \ldots, D\}$ create word importance scores $z_1^{(d)}, \ldots, z_{m_d}^{(d)}$.
2. Compute document set level attention matrix $\Lambda \in \mathbb{R}^{M \times M}$ where $M = \sum_d^D m_d$ and
$$\Lambda_{i,j} = \sigma(h_i^T h_j / \tau + b)$$
and h_i are outputs of the contextual representation of the i-th word (e.g. ELMO embedding).
3. Compute aggregate importance scores $\bar{z}_i = \sum_{j=1}^M z_j \cdot \Lambda_{i,j}$
4. Create ω for all sentences in the document set using the aggregated scores $\bar{z}_i$.
5. Proceed as in the single-document summarization model.

Supervise Attention for Abstractive Summarization

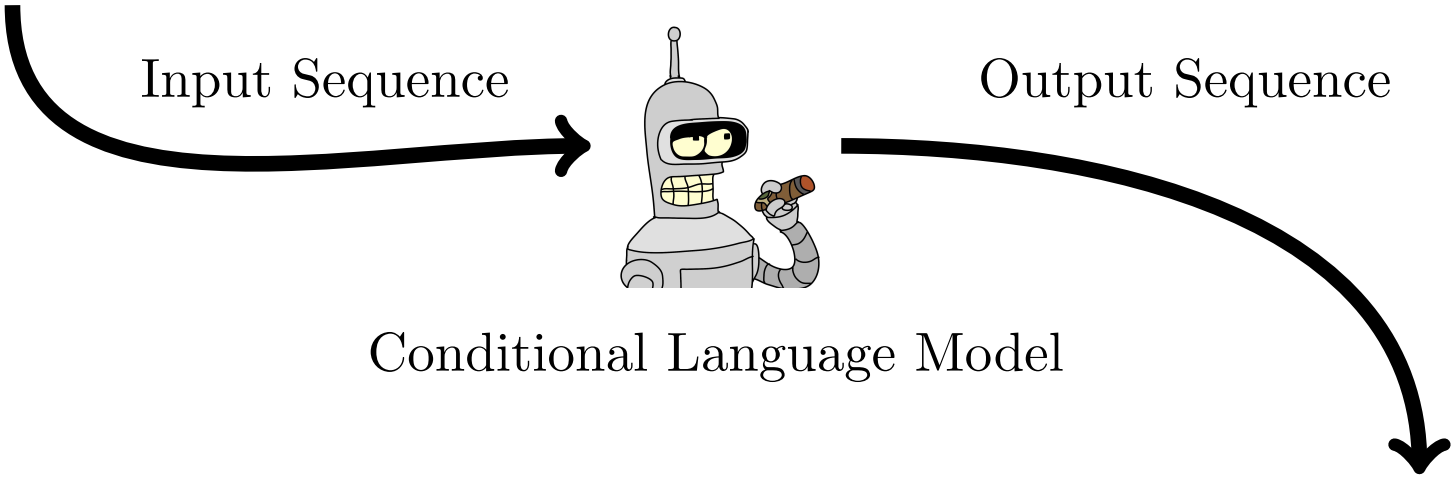
- Normalized importance scores z_i form a distribution over input tokens, like attention.
- This could be used in a couple ways:
  - Attention could be supervised to match the importance scores, i.e. supervising the content selection in the decoder.
  - Directly masking the input to the decoder.

Talk Outline

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

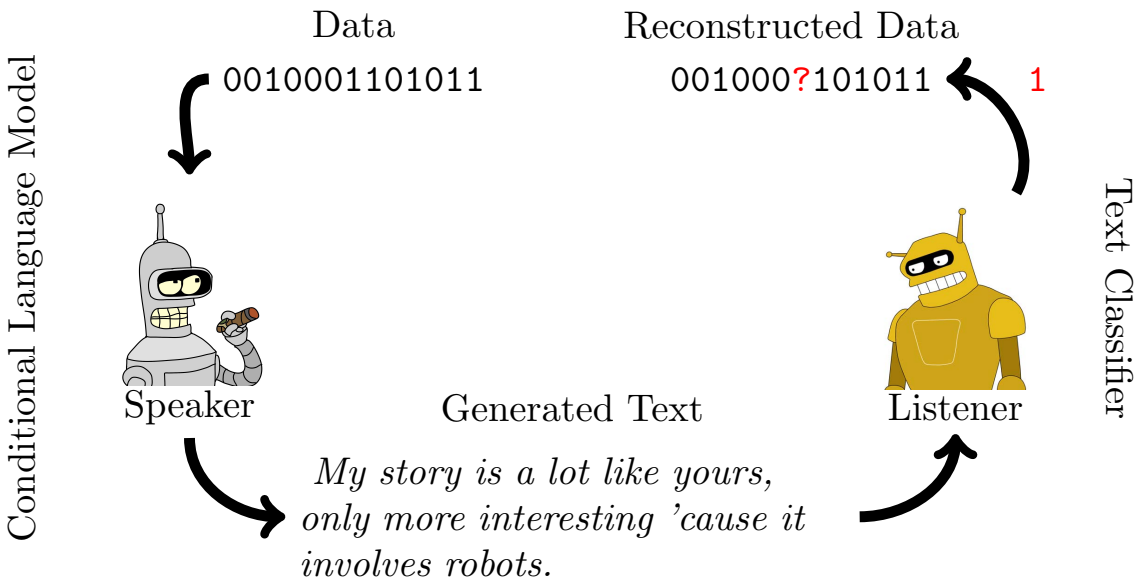
Hallucination in Seq2Seq Models

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...



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

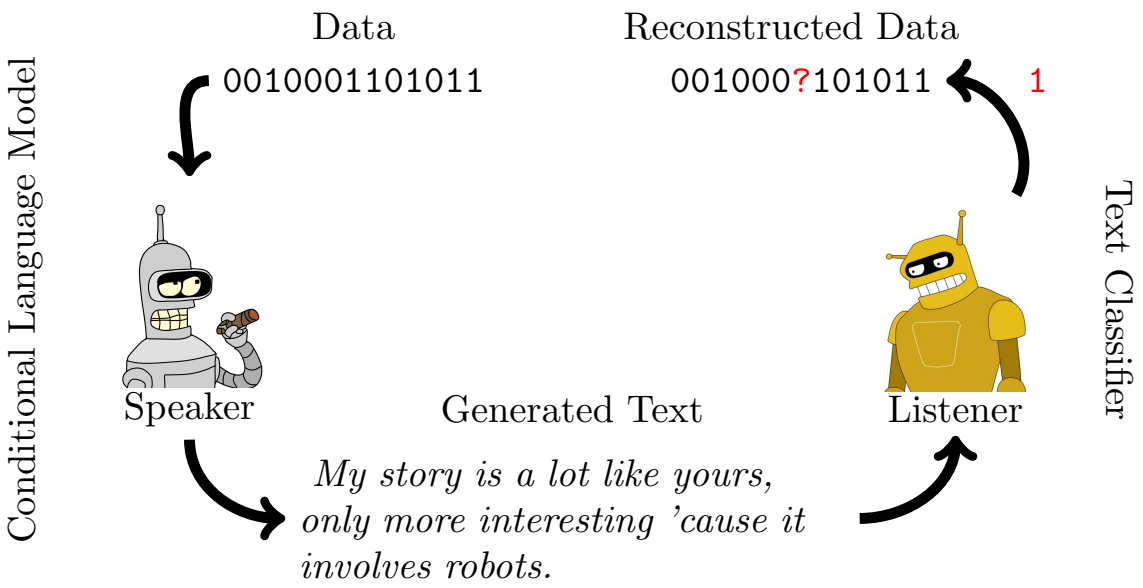
Faithful Generation



Inspired by:

- Rational Speakers and Listeners, (Andreas et al.)
- n-best ranking, (Collins and Koo)
- Round-trip translation

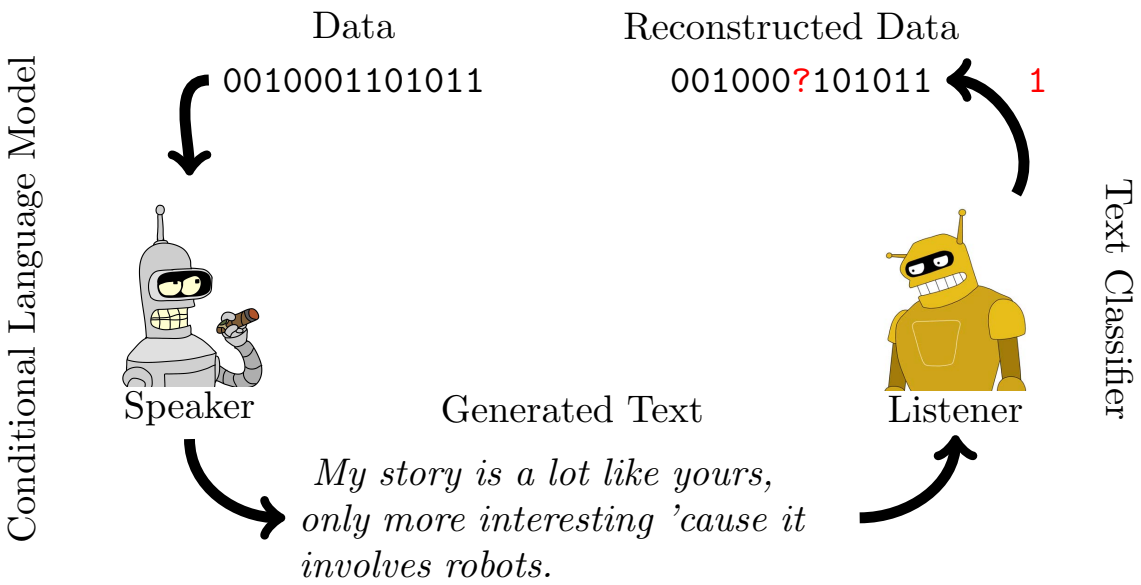
Faithful Generation



Motivation:

- Augment mle training with RL objective to improve accuracy of reconstruction without hurting fluency.
- We can apply this object 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.

Faithful Generation

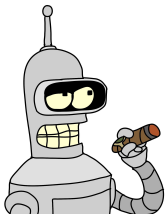


Other possible applications: controllable text generation.

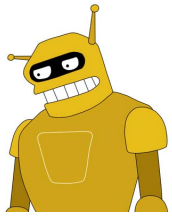
Two Applications

- **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

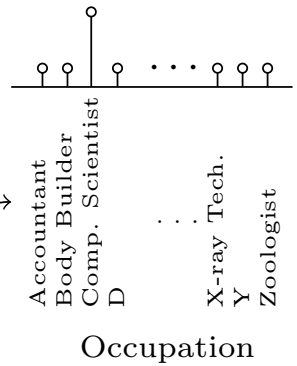


Speaker



Listener

Recovered Table Data



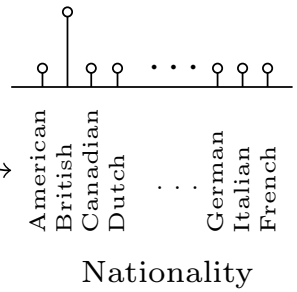


Table Data

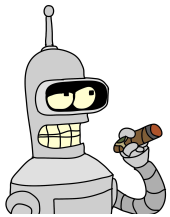
Name: Karen Spärck Jones

Occupation: Computer Scientist

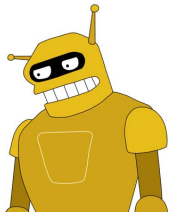
Nationality: British

Karen Spärck Jones was a British computer scientist.

Text-to-Text



Speaker



Listener

Input Document

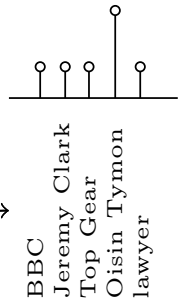
The BBC producer Oisin Tymon, allegedly struck by Jeremy Clarkson, will not press charges against the Top Gear host, his lawyer said...

Generated Summary

Producer Oisin Tymon will not press charges against Jeremy Clarkson, his lawyer says.

Cloze Question

The BBC producer ■ ■ allegedly struck by Jeremy Clarkson will not press charges against the Top Gear host, his lawyer said.



Answer Choices

Recovered Input Text Data

Planned Experiments

- Data-to-Text
  - E2E Dataset – generate restaurant descriptions from metadata.
  - WikiBio Datatest – generate Wikipedia biographical entries from table data.
- Text-to-Text
  - TL;DR Dataset – newly released, Reddit comments with summaries. (non-news dataset!)
  - Lots of news (CNN/DM, NYT, Newsroom, XSUM)

Planned Experiments

- Reinforce (Williams ???) 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.
  - Understand correlation in listener models $\Rightarrow$ enforce independent listener models.
  - Localize error signals with token level explanations from classifier.

Planned Experiments

- 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
   - Features
   - Salience Biased Affinity Propagation
   - Learning-to-Search Summarizer
2. Deep Learning Models of Salience
   - Sentence Salience
   - Word Salience
3. Faithful Generation
4. Research Plan

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

Contribution Status

- **Feature Based Models of Salience**
  - ✓ Incorporate salience regression with biased clustering. (TREC '14, ACL '15)
  - ✓ Incorporate salience regression with learning to search. (TREC '15, IJCAI '16)
  - ✓ Evaluation in Stream Summarization task.
- **Deep Learning Models of Salience**
  - ✓ Sentence Level Salience (EMNLP '18)
    - ✓ Implemented simplified DL models.
    - ✓ Model evaluation on Single Document Summarization task.
    - ✓ Extensive ablation studies to determine important lexical/structural features for learning.
  - ✗ Word level salience.
    - ✗ Develop word level DL model and margin learning framework.
    - ✗ Model evaluation on Single Document Summarization task
    - ✗ Adaption experiments to Multi-Document Summarization task.
    - ✗ Adaption experiments to Abstractive Summarization task.
- **Faithful Generation**
  - ✗ Data-to-Text experiments. (In Progress)
  - ✗ Question generation for summarization.
  - ✗ Text-to-Text experiments.