# CS388: Natural Language Processing
# Lecture 20: Summarization

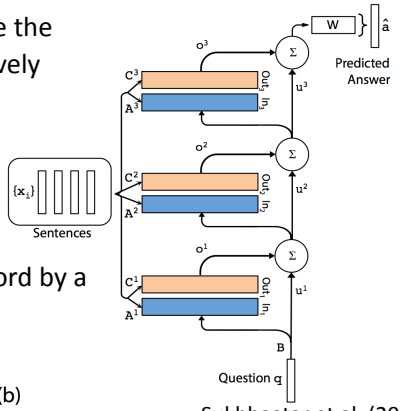
Greg Durrett

## Administrivia

- Proposals due Thursday

## Recall: Memory Networks

- Three layers of memory network where the query representation is updated additively based on the memories at each step
- How to encode the sentences?
  - Bag of words (average embeddings)
  - Positional encoding: multiply each word by a vector capturing position in sentence




(b)

Sukhbaatar et al. (2015)

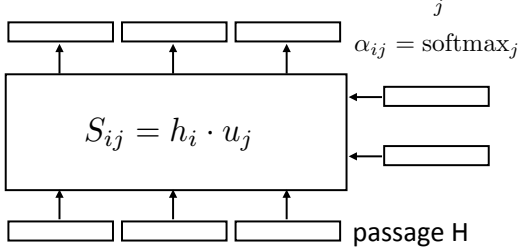
## Recall: Bidirectional Attention Flow

- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

$$\tilde{u}_i = \sum_j \alpha_{ij} u_j$$

$$\alpha_{ij} = \mathrm{softmax}_j(S_{ij})$$

- query "specialized" to the *i*th word
- dist over query




Seo et al. (2016)

# Final Projects

## EMNLP 2018: New Approaches

- Language modeling as pretraining is really effective
  - Minimizing/distilling BERT/ELMo will be important, but this will require a lot of compute (almost certainly infeasible as a final project)
- Transformers seem to be on the rise: linguistically-informed self attention, BERT, etc.
  - Understand transformer heads? Better inductive biases?
- Unsupervised MT: lots of problems here including lexicon induction, how to use syntax, etc.
- Variational autoencoders: still don't work great but there's optimism about them

## EMNLP 2018: New Datasets

- New QA settings:
  - "Conversational" machine reading (need to ask clarification questions to the user): https://arxiv.org/pdf/1809.01494.pdf
  - Questions as dialogue (user asks clarification questions to the system): https://arxiv.org/pdf/1808.07036.pdf
- emrQA: https://arxiv.org/pdf/1809.00732.pdf
- Commonsense: "Can a suit of armor conduct electricity?": https://arxiv.org/pdf/1809.02789.pdf
- Lots of new QA datasets, many will not prove that useful…sometimes hard to know in advance

## This Lecture

- Extractive systems for multi-document summarization
- Extractive + compressive systems for single-document summarization
- Single-document summarization with neural networks

# Summarization


- What makes a good summary?

# Summarization

BAGHDAD/ERBIL, Iraq (Reuters) - A strong earthquake hit large parts of northern Iraq and the capital Baghdad on Sunday, and also caused damage in villages across the border in Iran where state TV said at least six people had been killed.

There were no immediate reports of casualties in Iraq after the quake, whose epicenter was in Penjwin, in Sulaimaniyah province which is in the semi-autonomous Kurdistan region very close to the Iranian border, according to an Iraqi meteorology official.

But eight villages were damaged in Iran and at least six people were killed and many others injured in the border town of Qasr-e Shirin in Iran, Iranian state TV said.

The US Geological Survey said the quake measured a magnitude of 7.3, while an Iraqi meteorology official put its magnitude at 6.5 according to preliminary information.

Many residents in the Iraqi capital Baghdad rushed out of houses and tall buildings in panic.
…

# Summarization

Indian Express — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

The quake has been felt in several Iranian cities and eight villages have been damaged. Electricity has also been disrupted at many places, suggest few TV reports.

Summary

A massive earthquake of magnitude 7.3 struck Iraq on Sunday. The epicenter was close to the Iranian border. Eight villages were damaged and six people were killed in Iran.

# What makes a good summary?

Summary

A strong earthquake of magnitude 7.3 struck Iraq and Iran on Sunday. The epicenter was close to the Iranian border. Eight villages were damaged and six people were killed in Iran.

- Content selection: pick the right content
  - Right content was repeated within and across documents
  - Domain-specific (magnitude + epicenter of earthquakes are important)
- Generation: write the summary
  - Extraction: pick whole sentences from the summary
  - Compression: compress those sentences but basically just do deletion
  - Abstraction: rewrite + reexpress content freely

# Extractive Summarization

## Extractive Summarization: MMR

- Given some articles and a length budget of *k* words, pick some sentences of total length <= *k* and make a summary
- Pick important yet diverse content: maximum marginal relevance (MMR)

While summary is < *k* words

Calculate $MMR \stackrel{\text{def}}{=} Arg \max_{D_i \in R \setminus S} \left[ \lambda(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)) \right]$

"max over all sentences not yet in the summary"

"make this sentence similar to a query"

"make this sentence maximally different from all others added so far"

Add highest MMR sentence that doesn't overflow length

Carbonell and Goldstein (1998)

## Extractive Summarization: Centroid

- Represent the documents and each sentences as bag-of-words with TF-IDF weighting

While summary is < *k* words

Calculate score(sentence) = cosine(sent-vec, doc-vec)

Discard all sentences whose similarity with some sentence already in the summary is too high

Add the best remaining sentence that won't overflow the summary

Radev et al. (2004)

## Extractive Summarization: Bigram Recall

- Count number of *documents* each bigram occurs in to measure importance

score(*massive earthquake*) = 3 score(*magnitude 7.3*) = 2

score(*six killed*) = 2 score(*Iraqi capital*) = 1

- Find summary that maximizes the score of bigrams it covers
- ILP formulation: *c* and *s* are indicator variables indexed over bigrams ("concepts") and sentences, respectively

Maximize: $\sum_i w_i c_i$ $s_j Occ_{ij} \leq c_i, \quad \forall i, j$

Subject to: $\sum_j l_j s_j \leq L$ $\sum_j s_j Occ_{ij} \geq c_i \quad \forall i$

"set $c_i$ to 1 iff some sentence that contains it is included"

sum of included sentences' lengths can't exceed L

Gillick and Favre (2009)

# Evaluation: ROUGE

- Rouge-n: n-gram precision/recall/F1 of summary w.r.t. gold standard
- Rouge-2 correlates well with human judgments for multi-document summarization tasks

~~A~~ massive earthquake ~~of~~ magnitude 7.3 struck Iraq ~~on~~ Sunday — prediction

~~An~~ earthquake ~~was~~ detected ~~in~~ Iraq ~~on~~ Sunday — reference

ROUGE 2 recall = 1 correct bigram (Iraq, Sunday) / 4 reference bigrams

ROUGE 2 precision = 1 correct bigram (Iraq, Sunday) / 6 predicted bigrams

- Many hyperparameters: stemming, remove stopwords, etc.
- Historically: ROUGE recall @ k {words, characters}. Now: ROUGE F1

Lin (2004)

# Results

| Model | R-1 | R-2 | R-4 |
|---|---|---|---|
| Centroid | 36.03 | 7.89 | 1.20 |
| LexRank | 35.49 | 7.42 | 0.81 |
| KLSum | 37.63 | 8.50 | 1.26 |
| CLASSY04 | 37.23 | 8.89 | 1.46 |
| ICSI | 38.02 | **9.72** | **1.72** |
| Submodular | 38.62 | 9.19 | 1.34 |
| DPP | **39.41** | 9.57 | 1.56 |
| RegSum | 38.23 | 9.71 | 1.59 |
| Better centroid: | 38.58 | 9.73 | 1.53 |

Gillick and Favre / bigram recall

- Caveat: these techniques all work better for multi-document than single-document!

Ghalandri (2017)

# Multi-Document vs. Single Document

- "a massive earthquake hit Iraq" "a massive earthquake struck Iraq" — lots of redundancy to help select content in multi-document case
- When you have a lot of documents, there are more possible sentences to extract:

But eight villages were damaged in Iran and at least six people were killed and many others injured in the border town of Qasr-e Shirin in Iran, Iranian state TV said.

The quake has been felt in several Iranian cities and eight villages have been damaged.

- Multi-document summarization is easier?

# Compressive Summarization

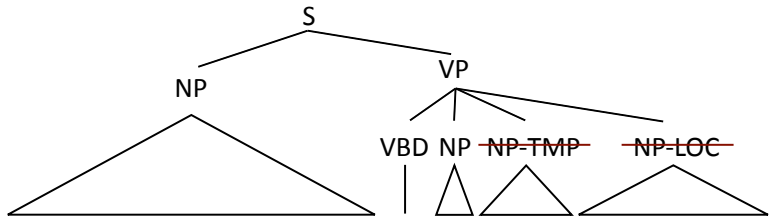
# Compressive Summarization

Indian Express — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

- Sentence extraction isn't aggressive enough at removing irrelevant content
- Want to extract sentences and also delete content from them

# Syntactic Cuts

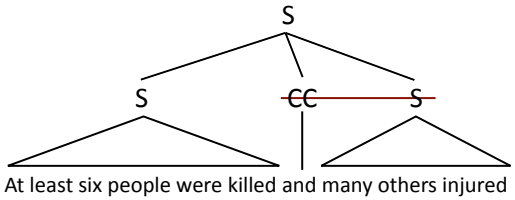
- Use syntactic rules to make certain deletions
- Delete adjuncts




Berg-Kirkpatrick et al. (2011)

# Syntactic Cuts

- Use syntactic rules to make certain deletions
- Delete second parts of coordination structures




Berg-Kirkpatrick et al. (2011)

# Compressive ILP

- Recall the Gillick+Favre ILP:

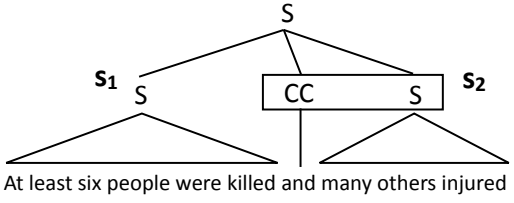
Berg-Kirkpatrick et al. (2011)

$$\text{Maximize: } \sum_i w_i c_i \qquad s_j Occ_{ij} \leq c_i, \quad \forall i, j$$

$$\text{Subject to: } \sum_j l_j s_j \leq L \qquad \sum_j s_j Occ_{ij} \geq c_i \quad \forall i$$

- Now $s_j$ variables are nodes or sets of nodes in the parse tree
- New constraint: $s_2 \leq s_1$

  "$s_1$ is a prerequisite for $s_2$"



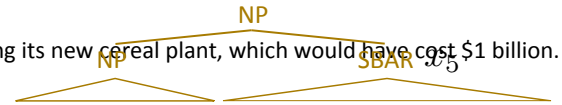
# Compressive Summarization

$x_1$ This hasn't been Kellogg's year.

$x_2$ The oat-bran craze has cost Kellogg market share.

$x_3$ Its president quit suddenly.

$x_4$ And now Kellogg is canceling its new cereal plant, which would have cost $1 billion.




ILP: $\max_{\mathbf{x}} \left(w^\top f(\mathbf{x})\right)$ $s.t.$ summary($\mathbf{x}$) obeys length limit
summary($\mathbf{x}$) is grammatical
summary($\mathbf{x}$) is coherent

# Constraints

$\max_{\mathbf{x}} \left(w^\top f(\mathbf{x})\right)$ $s.t.$ summary($\mathbf{x}$) obeys length limit
summary($\mathbf{x}$) is grammatical
summary($\mathbf{x}$) is coherent

Grammaticality constraints: allow cuts within sentences

Coreference constraints: do not allow pronouns that would refer to nothing

- If we're confident about coreference, rewrite the pronoun (it→Kellogg)
- Otherwise, force its antecedent to be included in the summary

Durrett et al. (2016)

# Features

$\max_{\mathbf{x}} \left(w^\top f(\mathbf{x})\right)$ $s.t.$ summary($\mathbf{x}$) obeys length limit
summary($\mathbf{x}$) is grammatical
summary($\mathbf{x}$) is coherent

- Now uses a feature-based model, where features identify good content

$f$(And now Kellogg is canceling its new cereal plant) =
- Centrality: $\mathbb{I}$(NumContentWords=4)
- Document position: $\mathbb{I}$(SentenceIndex=4)
- Lexical features: $\mathbb{I}$(FirstWord=*And*)

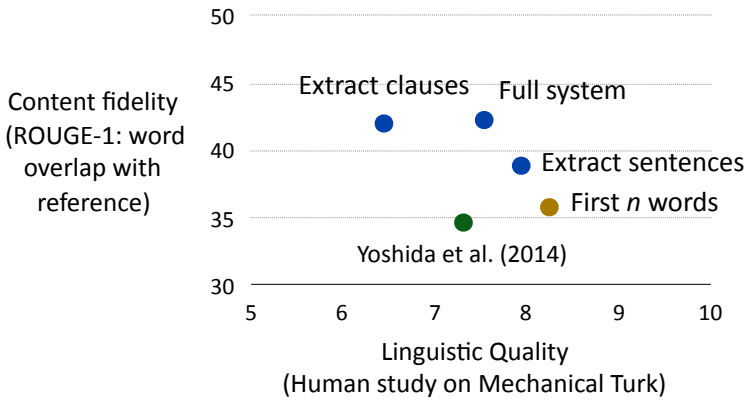
# Learning

$\max_{\mathbf{x}} \left(w^\top f(\mathbf{x})\right)$ $s.t.$ summary($\mathbf{x}$) obeys length limit
summary($\mathbf{x}$) is grammatical
summary($\mathbf{x}$) is coherent

- Train on a large corpus of New York Times documents with summaries (100,000 documents)
- Structured SVM with ROUGE as loss function
  - Augment the ILP to keep track of which bigrams are included or not, use these for loss-augmented decode

Berg-Kirkpatrick et al. (2011), Durrett et al. (2016)

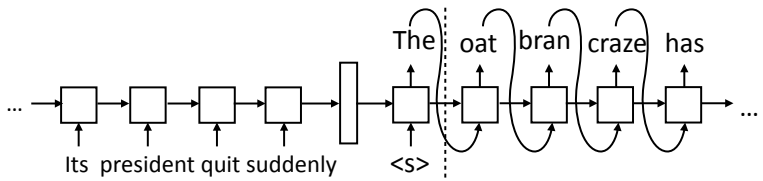
# Results: New York Times Corpus




# Neural Summarization

# Seq2seq Summarization

- Extractive paradigm isn't all that flexible, even with compression
- Training is hard! ILPs are hard! Maybe just use seq2seq?
- Train to produce summary based on document




Chopra et al. (2016)

# Seq2seq Summarization

- Task: generate headline from first sentence of article (can get lots of data!)

**I(1):** brazilian defender pepe is out for the rest of the season with a knee injury , his porto coach jesualdo ferreira said saturday . — sentence
**G:** football : pepe out for season — reference
**A+:** ferreira out for rest of season with knee injury — no attention
**R:** brazilian defender pepe out for rest of season with knee injury — with attention

- Works pretty well, though these models can generate incorrect summaries (who has the knee injury?)
- What happens if we try this on a longer article?

Chopra et al. (2016)

# Seq2seq Summarization

**Original Text (truncated):** 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. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, **he said his administration is confident it will be able to thwart criminals** and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. **the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.**

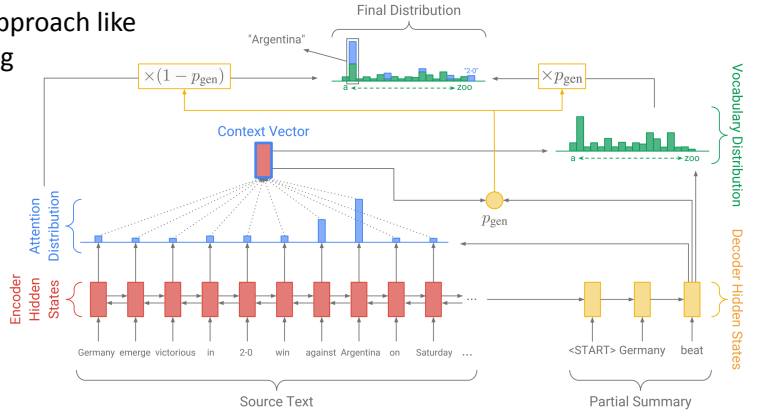
**Baseline Seq2Seq + Attention:** **UNK UNK** says his administration is confident it will be able to **destabilize nigeria's economy**. **UNK** says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

- What's wrong with this summary?

See et al. (2017)

# Pointer-Generator Model

- Copying approach like in Jia+Liang




See et al. (2017)

# Seq2seq Summarization

- Solutions: copy mechanism, coverage, just like in MT…

**Baseline Seq2Seq + Attention:** **UNK UNK** says his administration is confident it will be able to **destabilize nigeria's economy**. **UNK** says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

**Pointer-Gen:** *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

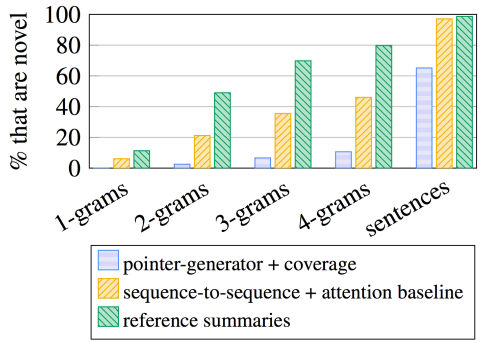
**Pointer-Gen + Coverage:** *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

- Things might still go wrong, no way of preventing this…

See et al. (2017)

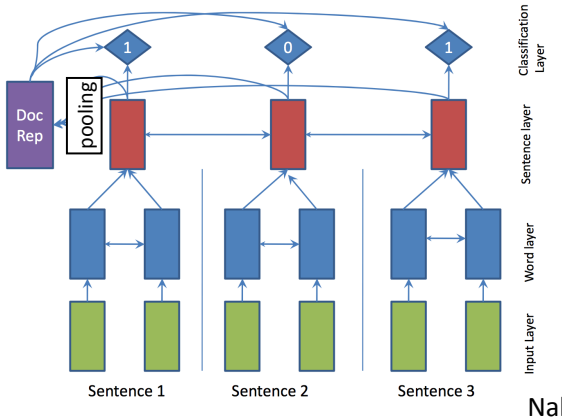
# Neural Abstractive Systems

- How abstractive is this, anyway?




See et al. (2017)

# Neural Extractive Systems




Nallapati et al. (2017)

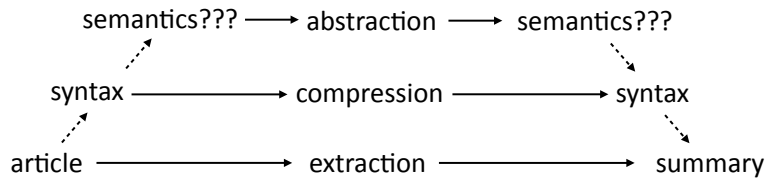
# Neural Systems: Results

| | ROUGE | | |
|---|---|---|---|
| | 1 | 2 | L |
| abstractive model (Nallapati et al., 2016)* | 35.46 | 13.30 | 32.65 |
| seq-to-seq + attn baseline (150k vocab) | 30.49 | 11.17 | 28.08 |
| seq-to-seq + attn baseline (50k vocab) | 31.33 | 11.81 | 28.83 |
| pointer-generator | 36.44 | 15.66 | 33.42 |
| pointer-generator + coverage | **39.53** | **17.28** | **36.38** |
| lead-3 baseline (ours) | 40.34 | 17.70 | 36.57 |
| lead-3 baseline (Nallapati et al., 2017)* | 39.2 | 15.7 | 35.5 |
| extractive model (Nallapati et al., 2017)* | 39.6 | 16.2 | 35.3 |

- Copy mechanism and coverage help substantially
- Abstractive systems don't/barely beat a "lead" baseline on ROUGE

See et al. (2017)

# Challenges of Summarization




- True abstraction?
  - Not really necessary for articles
  - Generating from structured information can usually be done with templates…

# Takeaways

- Extractive systems built on heuristics / ILPs work pretty well
- Compression can make things better, especially in the single-document setting
- Neural systems (like MT models) can do abstractive summarization, but they often just copy inputs (or deviate from inputs in bad ways)