CS388: Natural Language Processing

Lecture 24: Summarization

Greg Durrett

TEXAS
The University of Texas at Austin

Administrivia

- Project 2 back this week
- TACC allocations

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: $\sum_i w_i c_i$

Subject to: $\sum_j l_j s_j \leq L$

$s_j Occ_{ij} \leq c_i, \quad \forall i, j$

$\sum_j s_j Occ_{ij} \geq c_i \quad \forall i$

"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 somewhat 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

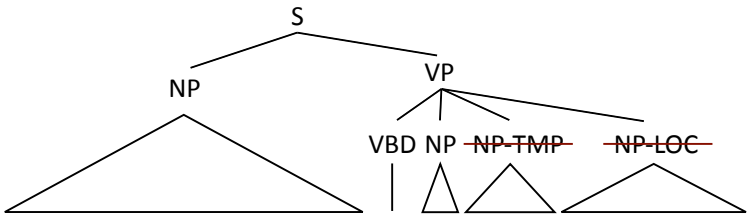
Compressive Summarization

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

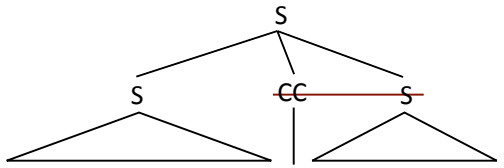


A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles)...

Berg-Kirkpatrick et al. (2011)

Syntactic Cuts

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



At least six people were killed and many others injured

Berg-Kirkpatrick et al. (2011)

Compressive ILP

- Recall the Gillick+Favre ILP: Berg-Kirkpatrick et al. (2011)

Maximize: $\sum_i w_i c_i$

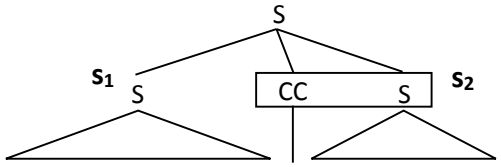
Subject to: $\sum_j l_j s_j \leq L$

$s_j Occ_{ij} \leq c_i, \quad \forall i, j$

$\sum_j s_j Occ_{ij} \geq c_i \quad \forall i$

- 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"



At least six people were killed and many others injured

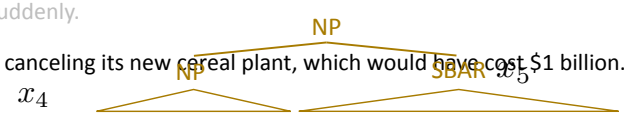
Compressive Summarization

x_1 This hasn't been Kellogg's year.

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

x_3 Its president quit suddenly.

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

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

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

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

- 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*)

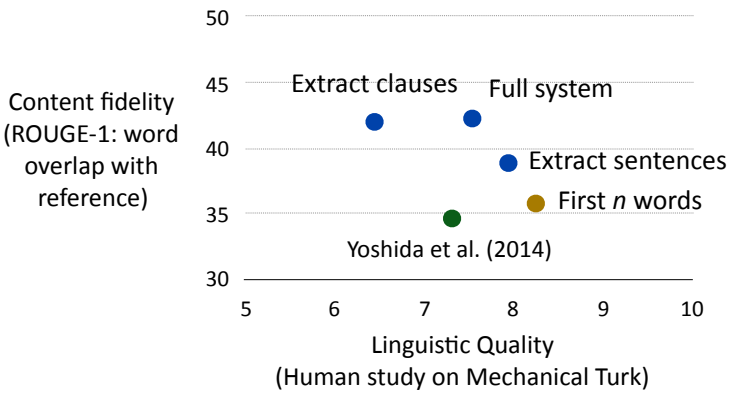
Learning

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

- 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

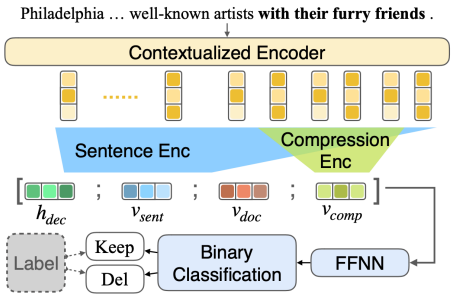
Berg-Kirkpatrick et al. (2011), Durrett et al. (2016)

Results: New York Times Corpus



Neuralizing this Model

- Model is now a neural model that scores sentences and compression options
- Decoding is done by beam search (not ILP), length constraint is not enforced as strongly anymore



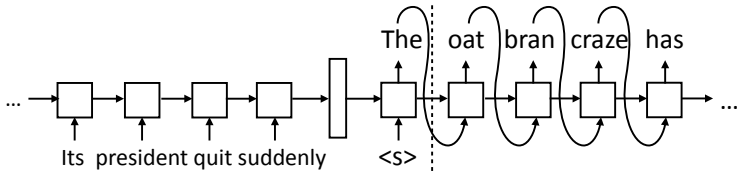
- Stronger results on NYT and on CNN/Daily Mail

Xu and Durrett (2019)

Neural Summarization

Seq2seq Summarization

- Extractive paradigm isn't all that flexible, even with compression
- Can we just use seq2seq models to simplify things?
- Train to produce summary based on document



- Need lots of data: most methods are going to be single-document

Chopra et al. (2016)

Seq2seq Headline Generation

- Headline generation 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.**

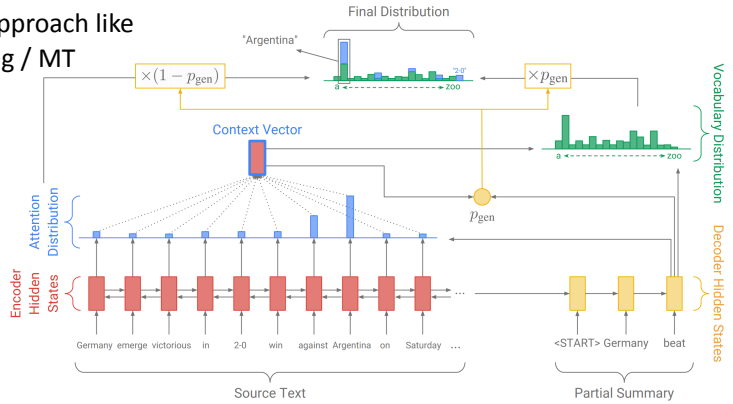
Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to **destabilize nigeria's economy**. UNK says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

- What's wrong with this summary?

See et al. (2017)

Pointer-Generator Model

- Copying approach like in Jia+Liang / MT



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.**

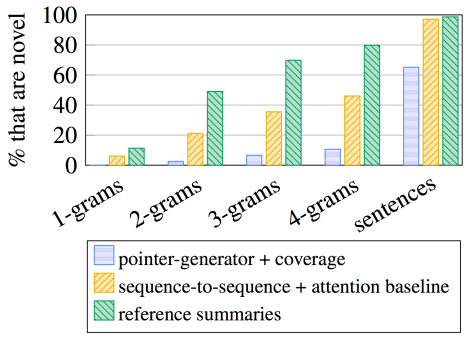
Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

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.

See et al. (2017)

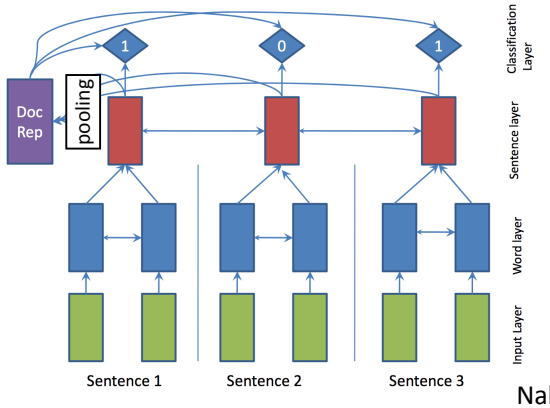
Neural Abstractive Systems

- How abstractive is this, anyway? Mostly doing copying!



See et al. (2017)

Neural Extractive Systems



Nallapati et al. (2017)

Neural Systems: Results

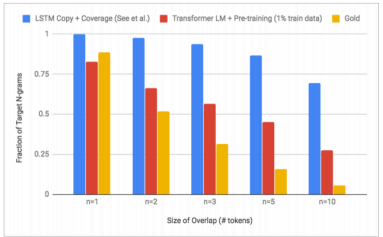
	ROUGE		
	1	2	L
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08
seq-to-seq + attn baseline (50k vocab)	31.33	11.81	28.83
pointer-generator	36.44	15.66	33.42
pointer-generator + coverage	**39.53**	**17.28**	**36.38**
lead-3 baseline (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 barely beat a "lead" baseline on ROUGE
- Best extractive systems now use BERT

See et al. (2017)

Pre-trained Models

- Pre-trained GPT adapted for summarization copies much less than See et al.'s model (but ROUGE is not much better)



	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

- BART: state-of-the-art ROUGE

Khandelwal et al. (2019), Lewis et al. (2019)

BART

Source Document (abbreviated)	BART Summary
The researchers examined three types of coral in reefs off the coast of Fiji ... The researchers found when fish were plentiful, they would eat algae and seaweed off the corals, which appeared to leave them more resistant to the bacterium Vibrio coralliilyticus, a bacterium associated with bleaching. The researchers suggested the algae, like warming temperatures, might render the corals' chemical defenses less effective, and the fish were protecting the coral by removing the algae.	Fisheries off the coast of Fiji are protecting coral reefs from the effects of global warming, according to a study in the journal Science.
Sacoolas, who has immunity as a diplomat's wife, was involved in a traffic collision ... Prime Minister Johnson was questioned about the case while speaking to the press at a hospital in Watford. He said, "I hope that Anne Sacoolas will come back ... if we can't resolve it then of course I will be raising it myself personally with the White House."	Boris Johnson has said he will raise the issue of US diplomat Anne Sacoolas' diplomatic immunity with the White House.

- Are these factual?

Lewis et al. (2019)

Problems in Neural Summarization

Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas even further south.

- Need to evaluate on factuality as well!

Kryściński et al. (2019b)

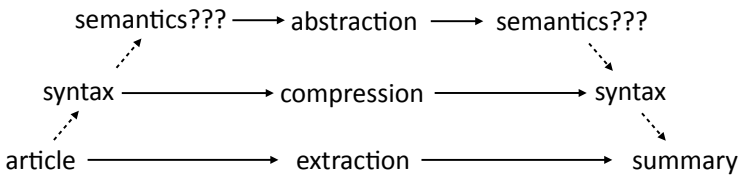
Problems in Neural Summarization

- Better ROUGE scores do not strongly correlate with human judgments of relevance, consistency, fluency, or coherence
- Other results: humans don't agree on what sentences are important, summaries mostly pick first few sentences (but this is a property of newswire)

	Pearson correlation								
	1 Reference			5 References			10 References		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
All Models									
Relevance	0.07	0.03	0.06	0.03	0.02	0.02	0.05	0.03	0.04
Consistency	0.08	0.03	0.07	0.02	0.01	0.01	0.03	0.01	0.02
Fluency	0.08	0.06	0.08	0.05	0.03	0.04	0.05	0.04	0.05
Coherence	0.06	0.05	0.07	0.05	0.04	0.05	0.04	0.03	0.04
Abstractive Models									
Relevance	0.04	0.01	0.05	0.01	0.00	0.00	0.04	0.02	0.03
Consistency	0.07	0.01	0.06	0.00	−0.02	−0.01	0.03	0.01	0.03
Fluency	0.06	0.04	0.07	0.03	0.01	0.02	0.05	0.04	0.04
Coherence	0.04	0.02	0.04	0.02	0.01	0.02	0.03	0.02	0.03
Extractive Models									
Relevance	0.14	0.09	0.13	0.09	0.05	0.07	0.06	0.03	0.04
Consistency	0.10	0.09	0.11	0.07	0.07	0.07	0.00	−0.03	−0.02
Fluency	0.13	0.14	0.13	0.10	0.10	0.08	0.06	0.03	0.04
Coherence	0.15	0.17	0.15	0.13	0.13	0.13	0.08	0.05	0.06

Kryściński et al. (2019)

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 may just copy inputs (or deviate from inputs in bad ways)