Neural Argument Generation Augmented with Externally Retrieved Evidence

Presented by Derek Joyce, Matthew Monjarrez, and Noah Lee

Xinyu Hua and Lu Wang

Association for Computational Linguistics, 2018

Research Question & Motivation

How can we automatically generate arguments of a different stance for a given statement?

- Argumentation is crucial in communication.
  - We want to avoid biased perception and uninformed decisions.
- Persuasion is complicated.
  - Being informative is already non-trivial, not to mention being persuasive.

Prior & Related Work

- Argument Component Detection
  - Evidence detection [Rinott et al, 2015]
  - Classification of types of supports [Hua and Wang, 2017]
- Argument and Evidence Retrieval
  - Argument search engine [Wachsmuth et al, 2017; Stab et al, 2018]
- Argument Component Generation
  - Retrieval based argument generation [Sato et al, 2015]
  - Argument strategy based generation [Zukerman et al, 2000]
- Argument Generation with Retrieval, Planning, and Realization [Hua, Hu, Wang, 2019]

Goal

- Design a counterargument using external evidence (Wikipedia)

- Challenges:

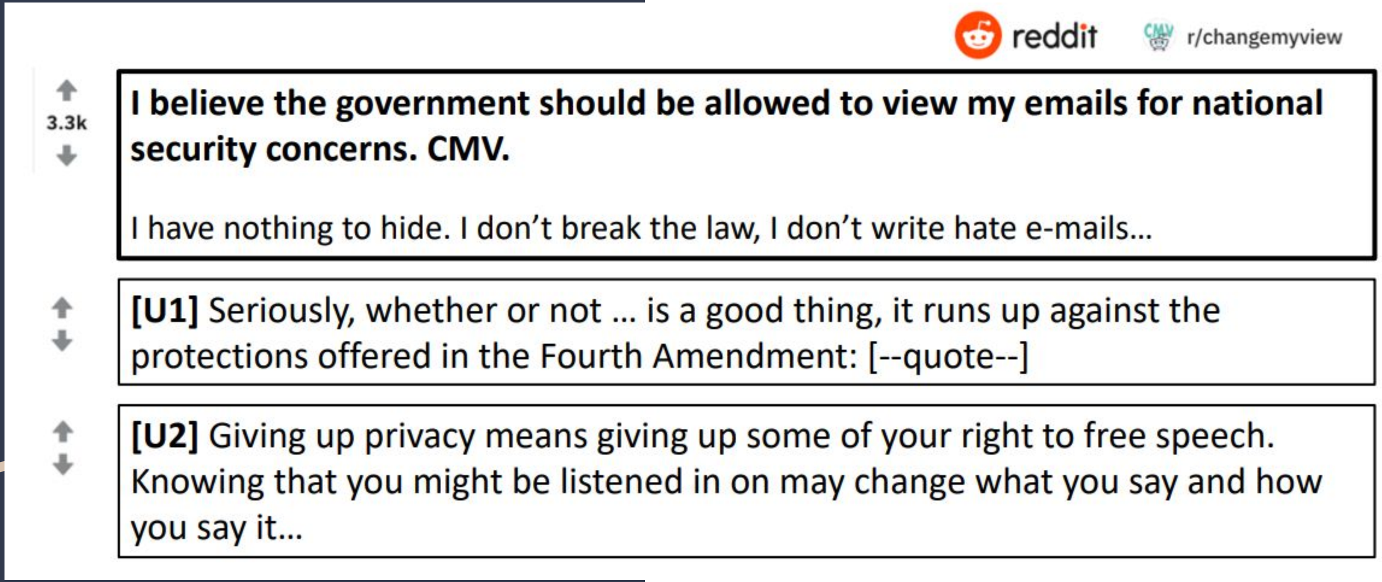
  1. Understanding the topic and stance

  2. Application of common sense knowledge

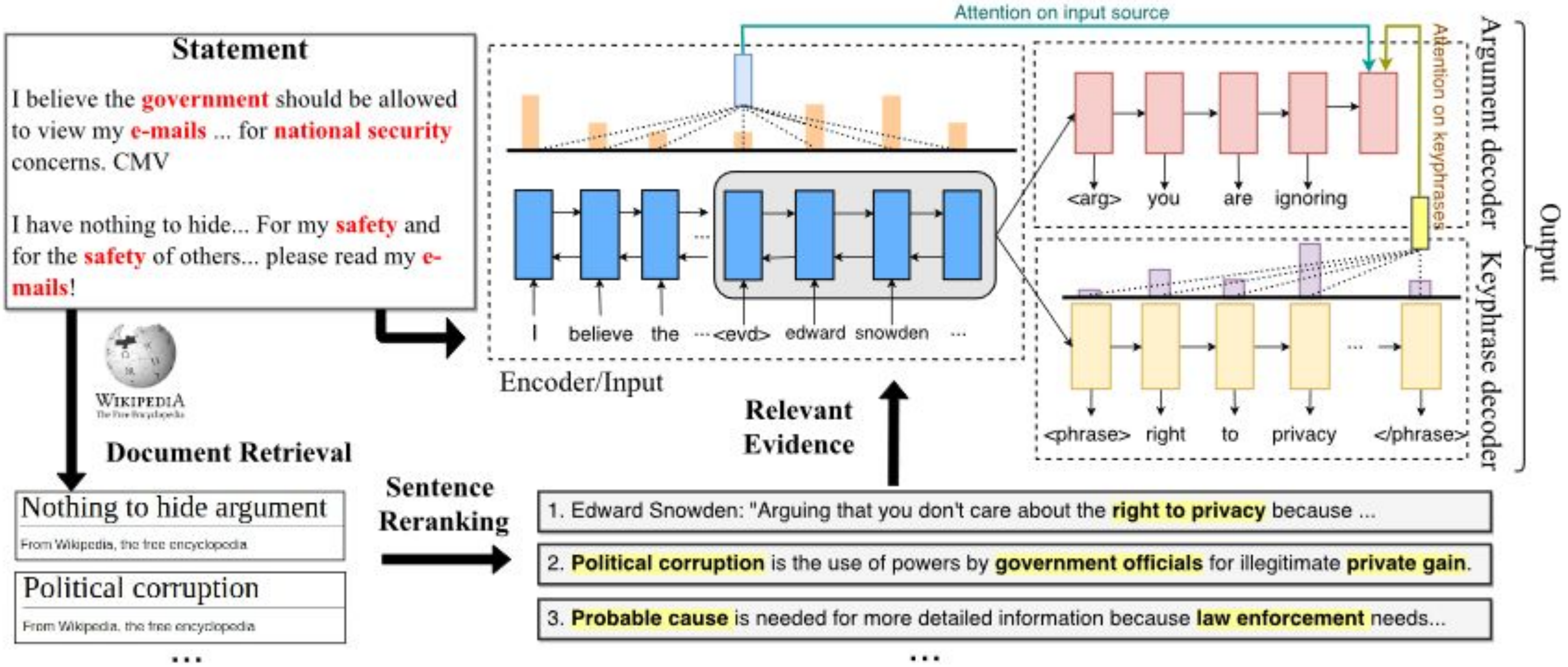
  3. Generating arguments in natural language texts

Data

- Use of r/ChangeMyView
- Posts from Jan 2013 - Jun 2017
- Political topics

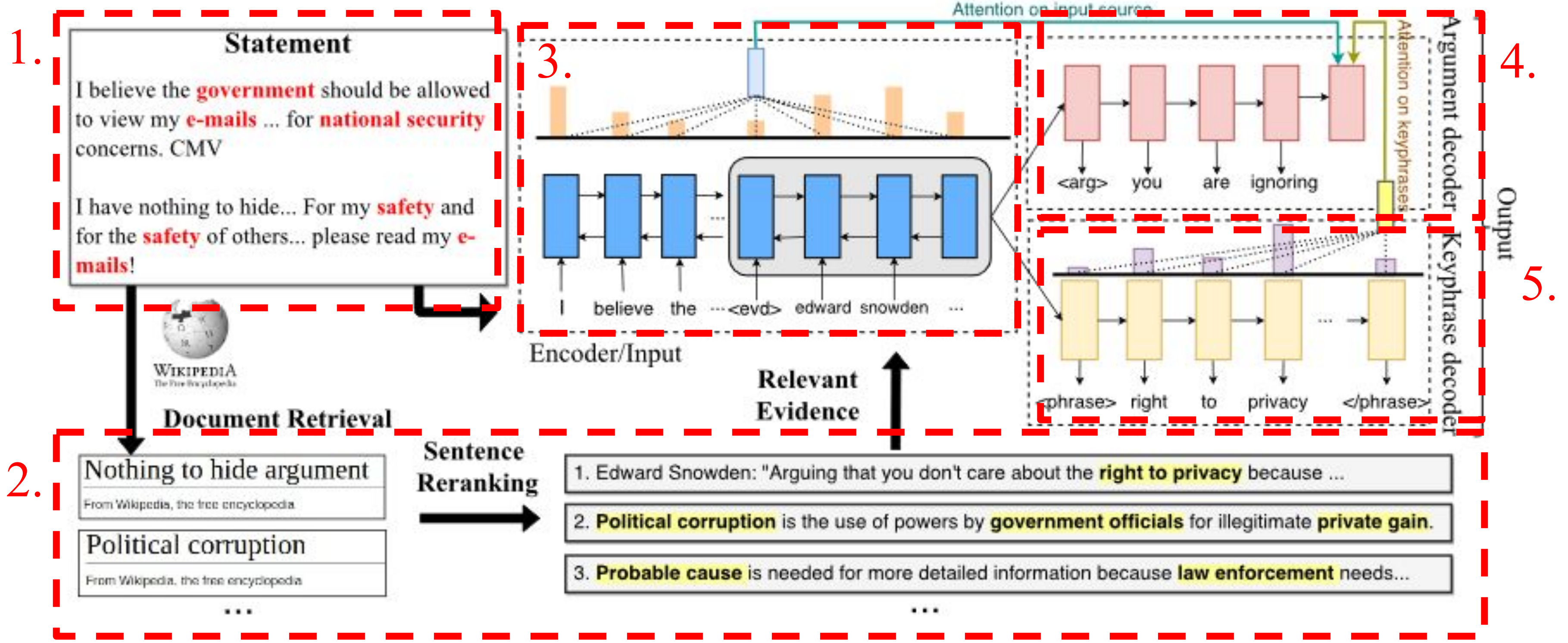


The Model



Statement
I believe the government should be allowed to view my e-mails ... for national security concerns. CMV
I have nothing to hide... For my safety and for the safety of others... please read my e-mails!
Attention on input source
Attention on keyphrases
I believe the ... <evd> edward snowden ...
Encoder/Input
<arg> you are ignoring
Argument decoder
<phrase> right to privacy </phrase>
Keyphrase decoder
Output
WIKIPEDIA
Document Retrieval
Nothing to hide argument
From Wikipedia, the free encyclopedia
Political corruption
From Wikipedia, the free encyclopedia
...
Sentence Reranking
Relevant Evidence
1. Edward Snowden: "Arguing that you don't care about the right to privacy because ...
2. Political corruption is the use of powers by government officials for illegitimate private gain.
3. Probable cause is needed for more detailed information because law enforcement needs...
...

The Model



1.
Statement
I believe the government should be allowed to view my e-mails ... for national security concerns. CMV
I have nothing to hide... For my safety and for the safety of others... please read my e-mails!
Wikipedia
Document Retrieval
2.
Nothing to hide argument
From Wikipedia, the free encyclopedia
Political corruption
From Wikipedia, the free encyclopedia
...
Sentence Reranking
1. Edward Snowden: "Arguing that you don't care about the right to privacy because ...
2. Political corruption is the use of powers by government officials for illegitimate private gain.
3. Probable cause is needed for more detailed information because law enforcement needs...
...
Relevant Evidence
3.
I believe the ... <evd> edward snowden ...
Encoder/Input
Attention on input source
Attention on keyphrases
4.
<arg> you are ignoring
Argument decoder
5.
<phrase> right to privacy ... </phrase>
Keyphrase decoder
Output

Query Extraction

Construct one query per sentence using topic signature words and search for relevant Wikipedia articles.

- Given
  - t: a word that appears in the input
  - T: cluster of articles on a given topic (input)
  - NT: articles not on topic T (background corpus)
- Decide if t is a topic word or not
- Words that have (almost) the same probability in T and NT are not topic words

H1:	$P(t \mid T) = P(t \mid NT) = p$ (t is not a descriptive term for the topic)
H2:	$P(t \mid T) = p_1$ and $P(t \mid NT) = p_2$ and $p_1 > p_2$ (t is a descriptive term)

Input statement

3.3k

I believe the government should be allowed to view my emails for national security concerns. CMV.

I have nothing to hide. I don't break the law...

Evidence Retrieval

Sort the retrieved articles for the top 10 sentences by the TF-IDF metric.

$$TFIDF\ score\ for\ term\ i\ in\ document\ j = TF(i,j) * IDF(i)$$

$where$

$IDF = Inverse\ Document\ Frequency$

$TF = Term\ Frequency$

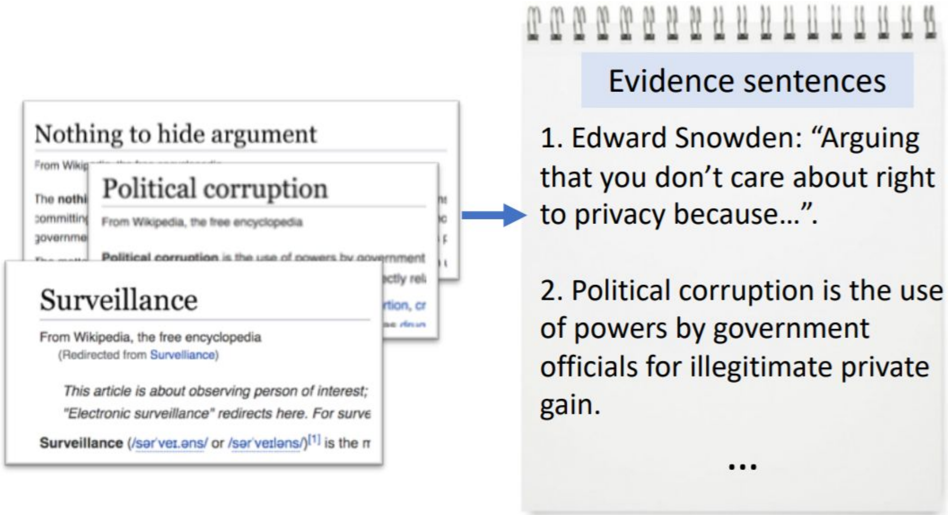
$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

$$IDF(i) = \log_2\left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$$

and

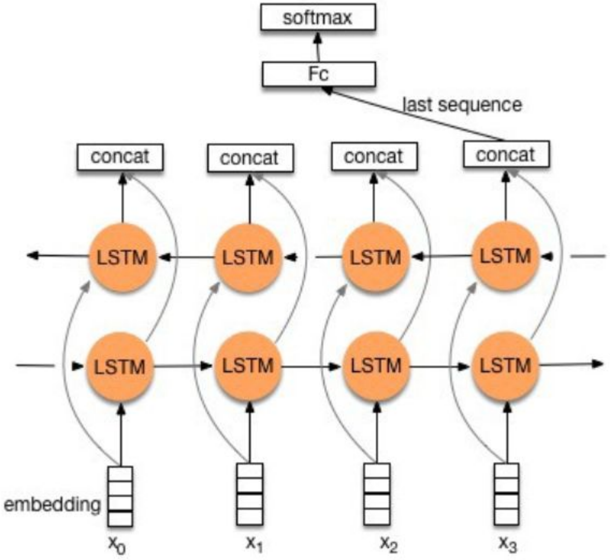
$t = Term$

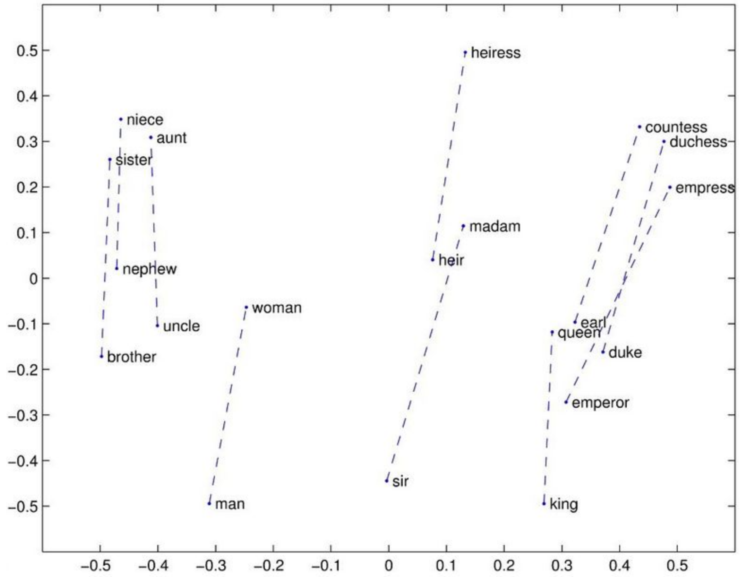
$j = Document$



Input Encoding

- Uses a bidirectional LSTM to encode the input.
- Pre-trained with 200 dimensional GloVe embeddings.





Keyphrase Decoding

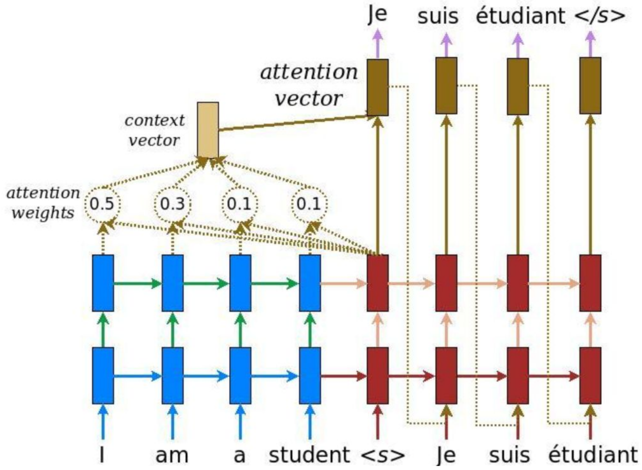
- Uses a unidirectional LSTM
- Extracts noun phrases and verb phrases
- Length of keyphrases between 2 and 10 words
- Contains not many “stop” words

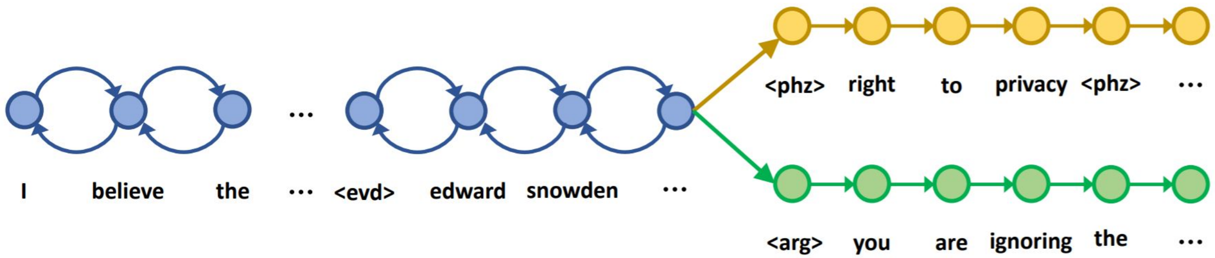
Numerous civil rights groups and privacy groups oppose surveillance as a violation of people's **right to privacy**.

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

Argument Decoding

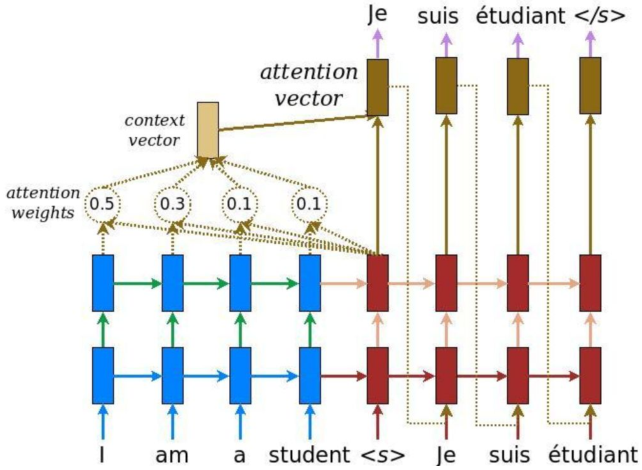
- Decoder is initialized with the last hidden state from the encoder or the keyphrase decoder
- An attention mechanism over the input and generated key phrases is used

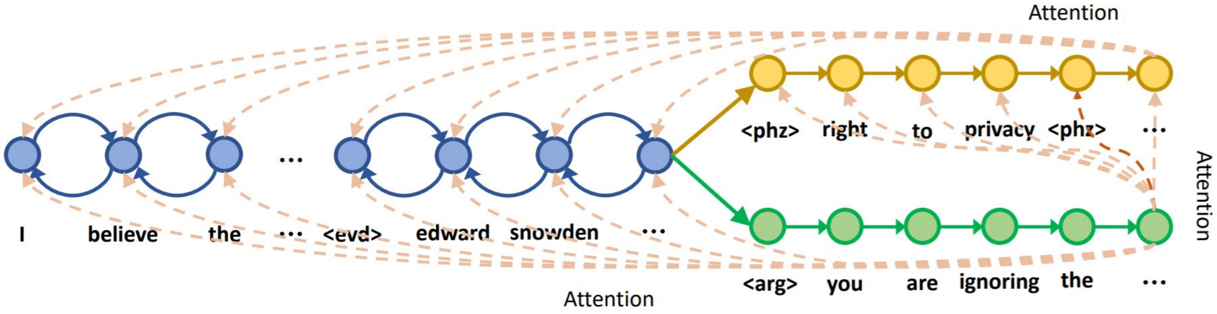




Argument Decoding

- Decoder is initialized with the last hidden state from the encoder or the keyphrase decoder
- An attention mechanism over the input and generated key phrases is used





Dataset

- Training Set: 224,553 examples (9,737 Original Posts)
- Validation Set: 13,911 examples (640 Original Posts)
- Testing Set: 30,417 examples (1,892 Original Posts)

Component	Stage 1	Stage 2	Stage 3
Encoder			
OP	50	150	400
Evidence	0	80	120
Decoder			
Keyphrases	0	80	120
Target Argument	30	80	120

Baseline & Comparisons

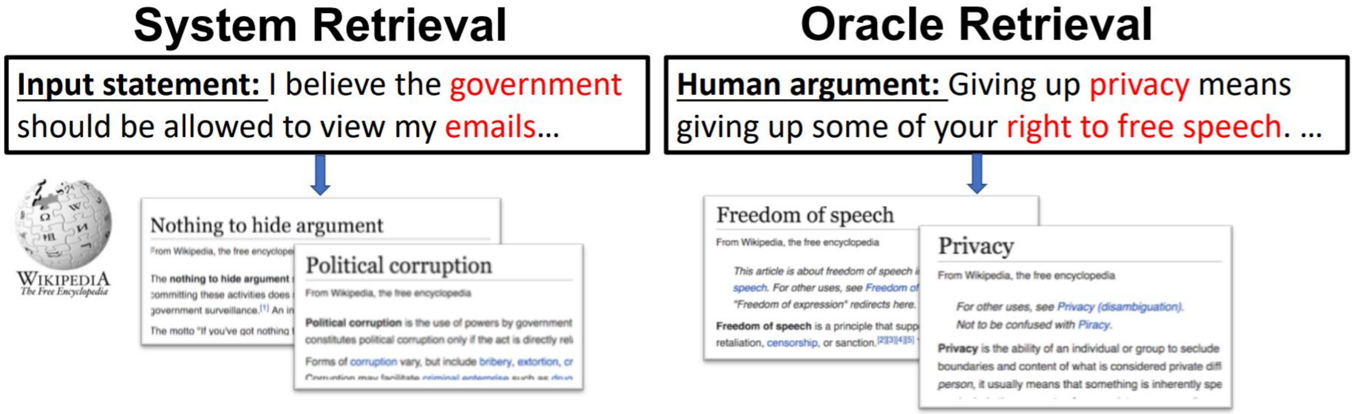
- Retrieval
- Seq2Seq
- Seq2Seq + encode evd
- Seq2Seq + encode keyphrases
- Decoder - Separate
- Decoder - Shared
- Decoder - Separate + attend KP
- Decoder - Shared + attend KP

Baseline & Comparisons

- Retrieval
- Seq2Seq
- Seq2Seq + encode evd
- Seq2Seq + encode keyphrases
- Decoder - Separate
- Decoder - Shared
- Decoder - Separate + attend KP
- Decoder - Shared + attend KP

System vs. Oracle Retrieval

- In the System setup, evidence can only be retrieved based on the input statement
- In the Oracle setup, human arguments are also used to retrieve evidence



Results

- Use of both automatic and manual evaluation methods
- Automatic evaluation performed with existing metrics and a novel technique
- Manual evaluation performed using human subjects

BLEU

- Goal: the closer the machine translation is to something that a human would output the better
- Scored against gold-standard translations (in this case human generated arguments).

METEOR

- Designed to address issues with the BLEU metric.
- Has a focus on precision and recall
    - Compare with BLEU’s focus on accuracy
    - The more in common a machine translation with a provided human one, the higher the score.

Automatic Evaluation

- The new models all achieve higher BLEU scores than any other method.
- Retrieval has the highest METEOR scores. Why?
  - The retrieval method yields extremely long results, which “match” easier with gold-standard translations.
- System retrieval had the highest BLEU scores because it produces more BLEU-favoring generic arguments

	w/ System Retrieval			*w/ Oracle Retrieval*		
	BLEU	**MTR**	**Len**	**BLEU**	**MTR**	**Len**
Baseline						
RETRIEVAL	15.32	**12.19**	151.2	10.24	**16.22**	132.7
Comparisons						
SEQ2SEQ	10.21	5.74	34.9	7.44	5.25	31.1
+ *encode evd*	18.03	7.32	67.0	13.79	10.06	68.1
+ *encode KP*	21.94	8.63	74.4	12.96	10.50	78.2
Our Models						
DEC-SHARED	21.22	8.91	69.1	15.78	11.52	68.2
+ *attend KP*	**24.71**	10.05	74.8	11.48	10.08	40.5
DEC-SEPARATE	24.24	10.63	88.6	17.48	13.15	86.9
+ *attend KP*	24.52	11.27	88.3	**17.80**	13.67	86.8

Novel Evaluation Method

- As we have seen, existing evaluation methods tend to favor generic arguments.
- However we contend that more specific arguments are more interesting.
- Solution: Train a model that scores topic relevance
  - Pair of OP and argument fed into the model. Trained on CMV data as gold standard.

Novel Evaluation Method

- When scored on topic relevance, our models score higher than other techniques.

- 29 common generic responses were chosen (such as “I don’t think so”).
    - Over 75% of seq2seq outputs contained a generic response compared to 16% of the newer models output.

	Standard Decoder		Our Decoder	
	MRR	P@1	MRR	P@1
Baseline				
RETRIEVAL	81.08	65.45	-	-
Comparisons				
SEQ2SEQ	75.29	58.85	74.46	57.06
+ *encode evd*	83.73	71.59	88.24	78.76
Our Models				
DEC-SHARED	79.80	65.57	**95.18**	**90.91**
+ *attend KP*	**94.33**	**89.76**	93.48	87.91
DEC-SEPARATE	86.85	76.74	91.70	84.72
+ *attend KP*	88.53	79.05	92.77	86.46

Human Evaluation

- Hire three trained human judges.
- They will score results on grammaticality, informativeness, and relevance on a scale of 1 to 5.

System	Gram	Info	Rel
RETRIEVAL	**4.5** $\pm$ 0.6	**3.7** $\pm$ 0.9	**3.3** $\pm$ 1.1
SEQ2SEQ	3.3 $\pm$ 1.1	1.2 $\pm$ 0.5	1.4 $\pm$ 0.7
OUR MODEL	2.5 $\pm$ 0.8	1.6 $\pm$ 0.8	1.8 $\pm$ 0.8

Conclusion

- Both automatic and human evaluation methods score our novel argument generation higher than popular existing seq2seq methods.
- Thank you for listening!

References

- Xinyu Hua, Lu Wang. 2018. Neural Argument Generation with Externally Retrieved Evidence. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL).
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pages 1532–1543.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 376–380. http://www.aclweb.org/anthology/W14-3348.