Faithful to the Original: Fact Aware Neural Abstractive Summarization

- Ziqiang Cao, Furu Wei, Wenjie Li, Sujian Li

Mitodru Niyogi

Faculty of Mathematics and Computer Science, Heidelberg University, and SAP SE

January 22, 2020

Problem

- Unlike extractive summarization, abstractive summarization has to fuse different parts of the source text, which inclines to create fake facts.
- Author's preliminary study reveals nearly 30% of the abstract summaries outputs from a state-of-the-art (Nallapati et al.) neural summarization system generate fake facts

Source	the repatriation of at least #,### bosnian moslems was postponed friday after the unhcr pulled out of the first joint scheme to return refugees to their homes in northwest bosnia .
Target	**repatriation** of bosnian moslems postponed
s2s	**bosnian moslems** postponed after unhcr pulled out of **bosnia**

Figure: An example of fake summaries generated by the state-of-the-art s2s model

How to handle this problem?

- To achieve this goal, the first step is to extract the facts from the source sentence
- encode existing facts into the summarization system to avoid fake generation
- Use of Open Information Tool (Angeli, Premkumar, and Manning 2015) to encode facts
- OpenIE refers to the extraction of entity relations from the open-domain text
- In OpenIE, a fact is typically interpreted as a relation triple consisting of (subject; predicate; object)
- However, the relation triples are not always extractable, e.g., from the imperative sentences

Sentence	I saw a cat sitting on the desk
Triples	(I; saw; cat)
	(I; saw; cat sitting)
	(I; saw; cat sitting on desk)

Figure: Examples of OpenIE triples in different granularities

Importance of Fact Description:

- words in fact descriptions are 40% more likely to be used in the summary than the words in the original sentence
- It indicates that fact descriptions truly condense the meaning of sentences to a large extent

Source:	Sentence	Fact
AvgLen	31.4	18.2
Count	1	2.7
Copy%	0.12	0.17

Figure: Comparisons between source sentences and relations

Author's key contributions:

- leverage open information extraction and dependency parse technologies to extract actual fact descriptions from the source text
- proposal of dual-attention sequence-to-sequence framework to force the summary generation conditioned on both the source text and the extracted fact descriptions
- Experiments on the Gigaword benchmark dataset demonstrate that the model can greatly reduce fake summaries by 80%.

Fact Description Extraction

- For example, given the source sentence in the above example, the popular OpenIE tool generates two relation triples including (repatriation; was postponed; friday) and (unhcr; pulled out of; first joint scheme)
- fact description representation: triplet(subject + predicate + object)
- **Problems with OpenIE:**
  - OpenIE may extract multiple triples to reflect an identical fact in different granularities
  - This can yield redundant triplets variants for one relation and increases the computation cost of the model
- **Authors' proposal:**
  - remove a relation triple if all its words are covered by another one to balance redundancy and fact completeness
  - use of dependency parser to supplment the absence of fact descriptions triplets with apprpriate tuples

Dependency Parser:

- A dependency parser converts a sentence into the **labeled (governor; dependent) tuples**.
- extract the predicate-related tuples according to the labels: nsubj (nominal subject), nsubjpass(passive nominal subject), csubj (clausal subject), csubjpass (clausal passive subject) and dobj(direct obejct)
- reserve the important modifiers including the adjectival (amod), numeric (nummod) and noun compound (compound) for more complete fact descriptions
- merge the tuples containing the same words, and order words based on the original sentence to form the fact descriptions
- It is found out that on average one key source word is missing in the fact descriptions

- **nsubj:** A nominal subject is a noun phrase which is the syntactic subject of a clause
- **nsubjpass:** A passive nominal subject is a noun phrase which is the syntactic subject of a passive clause
- **dobj:** The direct object of a VP is the noun phrase which is the (accusative) object of the verb
- **csub:** A clausal subject is a clausal syntactic subject of a clause, i.e., the subject is itself a clause
- **csubjpass:** A clausal passive subject is a clausal syntactic subject of a passive clause

Example	Relation
"Clinton defeated Dole"	nsubj(defeated, Clinton)
'Dole was defeated by Clinton"	nsubjpass(defeated, Dole)
"She gave me a raise"	dobj(gave, raise)
"What she said makes sense"	csubj(makes, said)
"That she lied was suspected by everyone"	csubjpass(suspected, lied)

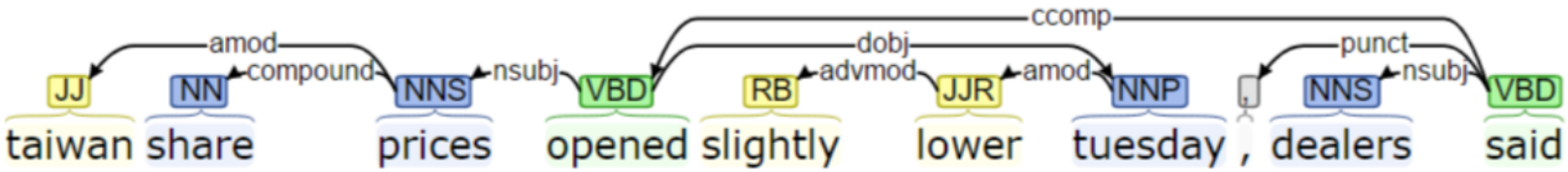

Figure: A dependency tree example. Two fact descriptions being extracted: taiwan share prices opened lower tuesday ||| dealers said

- Based on the dependency parser, predicate-related tuples: (prices; opened) (opened; tuesday) (dealers; said) were filtered and the modify-head tuples: (taiwan; price) (share; price) (lower; tuesday)
- These tuples are then merged to form two fact descriptions: taiwan share prices opened lower tuesday ||| dealers said

Fact Aware Neural Summarization Model Framework

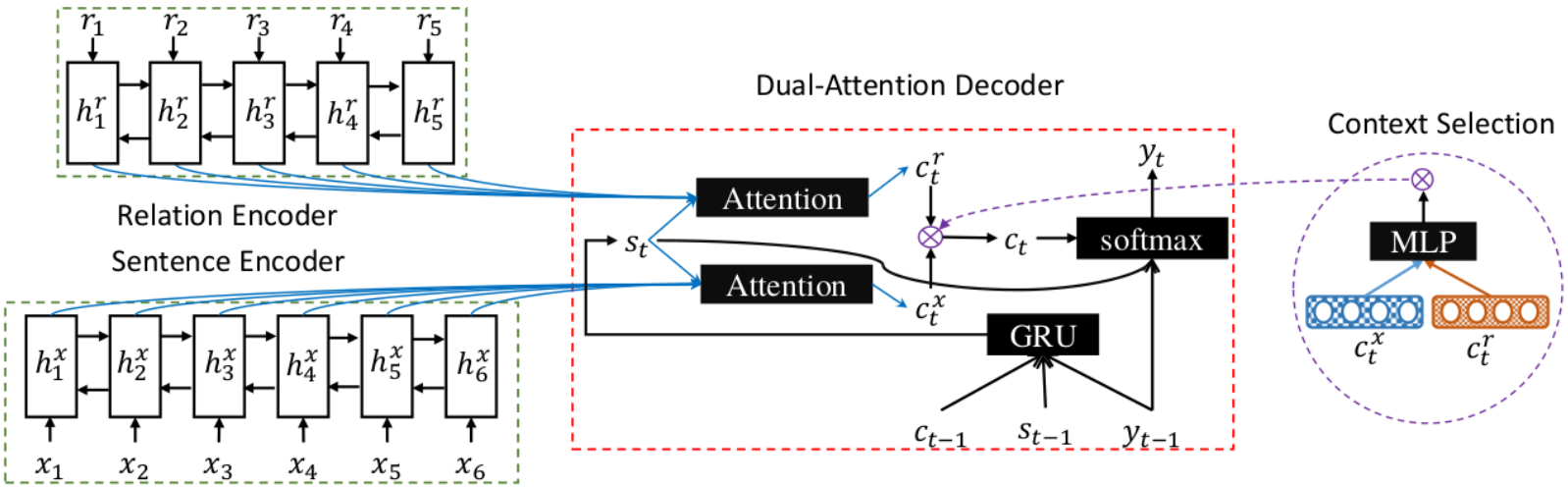

Figure: Modelframework

- model consists of three modules: two encoders and a dual-attention decoder equipped with a context selection gate network.
- The sentence encoder reads the input words x = (x_1,..., x_n) and converts into hidden states representation (h_1^x,..., h_n^x)
- the relation encoder converts the fact descriptions r = (r_1 , $\cdots$ r_k) into hidden states (h_1^r ,..., h_k^r)

- model computes the sentence and relation context vectors (c_t^x and c_t^r) at each decoding time step t following the attention mechanism
- context vectors are merged using the gate network
- The decoder produces summaries y = (y_1 , y_l) word-by-word conditioned on the tailored context vector which embeds the semantics of both source sentence and fact descriptions

Encoders

- input: source sentence x and the fact descriptions r
- For each sequence, the BiGRU encoder is used to construct its semantic representation
- The GRU at the time step i is defined as follows:

$$h_i = GRU(x_i, h_{i-1}) \tag{1}$$

- the composite hidden state represntation of a word: $h_i = [h_i^{\rightarrow}; h_i^{\leftarrow}]$
- For the relation sequence r, introduce boundary indicators γ to separate their hidden states
- γ is defined as follows:

$$\gamma_i = \begin{cases} 0, & r_i \text{ is "|||"} \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

- γ is used to reset the GRU state in Eq. 1:

$$h_i' = \gamma_i h_i \tag{3}$$

Dual-Attention Decoder

- decoder: GRU with attentions (Bahdanau et.)
- At each decoding time step t, new hidden state s_t :

$$s_t = GRU(y_{t-1}, c_{t-1}, s_{t-1}) \tag{4}$$

- the context representation of the sentence at time step t:

$$e^x_{t,i} = MLP(s_t, h^x_i) \tag{5}$$

$$\alpha^x_{t,i} = \frac{\exp(e^x_{t,i})}{\sum_i \exp(e^x_{t,j})} \tag{6}$$

$$c^x_t = \sum_i \alpha^x_{t,i} h^x_i \tag{7}$$

- "$FTSum_c$": concatenates two context vectors:

$$c_t = [c^x_t; c^r_t] \tag{8}$$

- "$FTSum_g$": used MLP to build a gate network

$$g_t = MLP(c^x_t, c^r_t) \tag{9}$$

$$c_t = g_t \odot c^x_t + (1 - g_t) \odot c^r_t \tag{10}$$

- Experiments show that $FTSum_g$ significantly outperforms $FTSum_c$
- **prediction:** softmax layer over previous word y_{t-1} , context vector c_t and current decoder state s_t

$$o_t = W_w[y_{t-1}] + W_c c_t + W_s s_t \tag{11}$$

$$p(y_t \mid y_{<t}) = softmax(W_o o_t) \tag{12}$$

Learning:

- Goal: is to maximize the estimated probability of the actual summary

$$J(\theta) = -\frac{1}{|D|} \sum_{(x,r,y)\epsilon D} \log(p(y \mid x, r)) \tag{13}$$

Dataset

Dataset	Train	Dev.	Test
Count	3.8M	189k	1951
AvgSourceLen	31.4	31.7	29.7
AvgTargetLen	8.3	8.3	8.8

Figure: Data statistics for the English Gigaword. Avg- SourceLen is the average input sentence length and AvgTar- getLen is the average headline length.

- Experiments conducted on the Annotated English Gigaword corpus, as with (Rush, Chopra, and Weston 2015b)
- This parallel corpus is produced by pairing the first sentence in the news article and its headline as the summary with heuristic rules.

Evaluation metrics:

- Rouge 1
- Rouge 2
- ROUGE-L
- Manual inspection whether the generated summaries accord with the facts in the original sentences into three categories: FAITHFUL, FAKE and UNCLEAR

Training:

- Glove embeddings of 200 dimensional size
- All GRU hidden state dimension fixed to 400
- Dropout probability of 0.5
- beam search of size 6 was used to generate the summary, and the maximal length of a summary is 20 words

Baselines:

- **ABS:**(Rush, Chopra, and Weston 2015a) used an attentive CNN encoder and NNLM decoder to summarize the sentence
- **ABS+:** (Chopra et al. 2015) fine tuned ABS model to balnce abstractive and extractive tendency
- **RAS-Elman:** (Chopra et al. 2016) convolutional attention-based encoder and an RNN de coder .
- **Feats2s:** (Nallapati et al. 2016) full s2s RNN model with added POS tag and NER, to enhance the encoder representation
- **Luong-NMT:** (Luong et al. 2015) Two-layer LSTMs MNT model with 500 hidden units in each layer
- **att-s2s:** standard attentional s2s with dl4mt model (conditioned GRU with attention)

Informativeness Evaluation

Model	Perplexity
ABS†	27.1
RAS-Elman†	18.9
s2s-att	24.5
FTSum_c	20.1
FTSum_g	**16.4**

Figure: Final perplexity on the development set

Takeaways:

- Proposed model achieves the lowest perplexity compared against the state-of-the-art systems
- $FTSum_g$ largely outperforms $FTSum_c$ which outlines the importance of context selection
- fact descriptions have significant contribute to the increase of ROUGE scores

Model	RG-1	RG-2	RG-L
ABS†	29.55*	11.32*	26.42*
ABS+†	29.78*	11.89*	26.97*
Feats2s†	32.67*	15.59*	30.64*
RAS-Elman†	33.78*	15.97*	31.15*
Luong-NMT†	33.10*	14.45*	30.71*
s2s+att	34.23*	15.52*	31.57*
FTSum_c	35.73*	16.02*	34.13
FTSum_g	**37.27**	**17.65**	**34.24**

Figure: ROUGE F1 performance. "*" indicates statistical significance of the corresponding model with respect to the baseline model on the 95% confidence interval in the official ROUGE script. RG refers to ROUGE for short.

- Despite ABS+ and Feats2s have utilized a series of hand-crafted features, the proposed model surpasses Feats2s by 13% and ABS+ by 56% on ROUGE-2

Faithful evaluation:

Model	Category	Count
att-s2s	FAITHFUL	68
	FAKE	27
	UNCLEAR	5
FTSum_g	FAITHFUL	87
	FAKE	6
	UNCLEAR	7

Figure: Faithfulness performance on the test set.

- randomly select 100 sentences from the test set. Then, classify the generated summaries as FAITHFUL, FAKE or UNCLEAR
- Around 30% s2s-att outputs gives disinformation
- Proposed model only reports to 6% fake
- 90% of summaries generated by propsoed model is faithful
- It is observed that s2s-att tends to copy the words closer to the predicate and regard them as its subject and object

	Example 1
Source	the repatriation of at least #,### bosnian moslems was postponed friday after the unhcr pulled out of the first joint scheme to return refugees to their homes in northwest bosnia .
Relations	unhcr pulled out of first joint scheme ||| repatriation was postponed friday ||| unhcr return refugees to their homes
Target	repatriation of bosnian moslems postponed
att-s2s	(FAKE) **bosnian moslems** postponed after unhcr pulled out of **bosnia**
FTSum	(FAITHFUL) repatriation of bosnian moslems postponed
	Example 2
Source	davis love said he was thinking of making the world cup of golf a full time occupation after taking a ## stroke lead over japan in the event with us partner fred couples here on saturday .
Relations	making world cup full time occupation ||| taking ## stroke lead
Target	americans lead UNK by ## strokes
att-s2s	(FAKE) davis love says he is thinking of **the world cup**
FTSum	(UNCLEAR) love **in** the world cup of golf
	Example 3
Source	the us space shuttle atlantis separated from the orbiting russian mir space station early saturday , after three days of test runs for life in a future space facility , nasa announced .
Relations	us space shuttle atlantis separated from orbiting russian mir space station ||| us space shuttle atlantis runs after three days of test for line in future space facility
Target	atlantis mir part ways after three-day space collaboration by emmanuel UNK
att-s2s	(UNCLEAR) space shuttle atlantis separated after # days of test runs for life
FTSum	(FAITHFUL) space shuttle atlantis separated from mir

Figure: Examples of defective outputs. Bold font to indicate the problematic parts

- **Example 1:** att-s2s treats "bosnian moslems" as the subject of "postponed" and "bosnia" as the object of "pulled out of"
- **Example 2:** att-s2s mismatches the object while the proposed model fails to produce a complete sentence due to absence of high-quality fact descriptions (main clause is hard to summarize)
- **Example 3:** inability of the model to utilize multiple long fact descriptions for generation. Model utilizes one fact instead of 2 long fact descriptions
- As a result, despite the high faithfulness, the informativeness is somewhat damaged

Gate Analysis: What does the gate network (Eq. 9) actually learns?

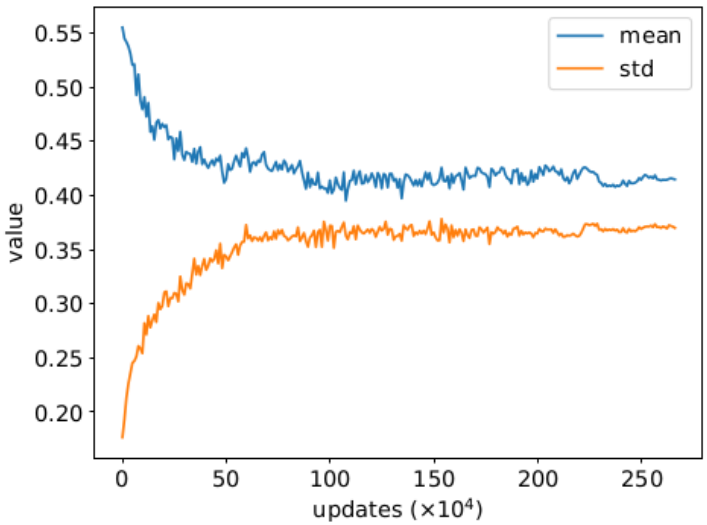

Figure: Gates change during training

- gate values apparently reflect the relative reliability of sentence and fact descriptions
- At the beginning, the average gate value exceeds 0.5, which means the generation is biased to the source sentence

- As training proceeds, drop in gate value results that the fact descriptions are more reliable
- the average gate value is gradually stabilized to 0.415
- the ratio of sentence and relation gate values i.e., (1 - 0.415)/0.415 $\approx$ 1.41, is extremely close to the ratio of copying proportions i.e., 0.17/0.12 $\approx$ 1.42
- It seems that the model predicts the copy proportion and normalizes it as the gate value

Conclusion

- investigates the faithfulness problem in abstractive summarization
- popular OpenIE and dependency parse tools were used to extract fact descriptions in the source sentence
- introduces the dual-attention s2s framework to force the generation conditioned on both source sentence and the fact descriptions
- fact descriptions indeed increases ROUGE scores
- Experiments on the Gigaword benchmark demonstrate that the proposed model greatly reduce fake summaries by 80%.

Pros:

- significance tests were performed on results
- first to test the authenticity of the generated abstract summaries
- strong baselines
- massive gain to reduce the fake summaries

Cons:

- Failure of the study as to how the model fails to utilize multiple fact descriptions to improve informativeness
- The author should have explained more about the concept of dependency parser and relation terms in general
- unavailability of code
- didn't present as how the loss converges during the training
- They didn't try out with other LSTM based encoders

Reference


Ziqiang Cao, Furu Wei, Wenjie Li, Sujian Li. Faithful to the Original: Fact Aware Neural Abstractive Summarization. *arXiv:1711.04434, 2017*


de Marneffe, Marie-Catherine and Manning, Christopher D. 2008. Stanford typed dependencies manual. Technical report, Stanford University.

Thank you very much for your attention.

Do you have any question?