Text summarization for biomedical domain content

Karen Ouyang | kjouyang@stanford.edu
Biomedical Informatics

MOTIVATION

Biomedical information in the form of scientific articles and electronic medical records is increasing at an alarmingly fast pace. The output of publications in the biomedical domain is estimated to double every 5-10 years, currently with >3000 new articles published per day. As such, there is clear utility in having systems that can automatically handle various natural language processing (NLP) tasks, such as text summarization. Summarization is the task of distilling longer text to a shorter version that retains the key information from the original text. The objective of this project is to apply NLP machine learning models for text summarization that perform well on general language text summarization datasets and further modify/adapt for biomedical domain specific text summarization. I evaluate and compare the performance on general language (CNN-DailyMail) versus biomedical-specific (BioASQ) datasets, and analyze results to leverage general language models for biomedical domain-specific applications.

DATA

- Training: 287,226 pairs
- Validation: 13,368 pairs
- Test: 11,490 pairs

CONTEXT: a cheeky monkey was captured on camera snatching a banana from a female tourist before slapping her gopro when she got too close. filmed in the thai town of kanchanaburi, the monkey approaches the woman, who holds a banana, with its outstretched hands. grabbing it in both hands, the monkey takes a small bite before pulling it from its skin, which he leaves with the lady. (...) getting right up into the camera's lens the monkey appears to sniff it while diverting its eyes, as if hoping that it is food. the cheeky monkey snatches the banana from the woman's hand and begins scoffing it down, realising that it is out of luck, it returns to its original position and continues tucking into the banana. (...) the footage was captured by maja and diano, a pair who describe themselves on their youtube channel as a young married couple with an impulse to explore, film and edit great travel moments. (...)

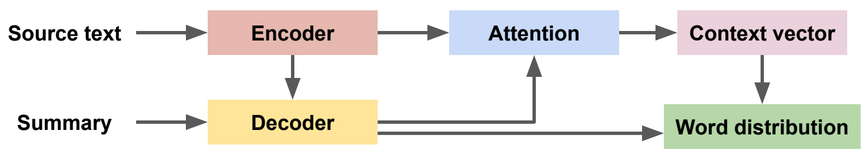
SUMMARY (Ground Truth): the woman holds out a banana, which the monkey quickly snatches. monkey then approaches the camera and sniffs it to see if it is food. woman gets too close to protective monkey and it slaps her gopro. the footage was captured by a couple in thai town of kanchanaburi

- Training: 815 pairs
- Validation: 193 pairs
- Test: 665 pairs

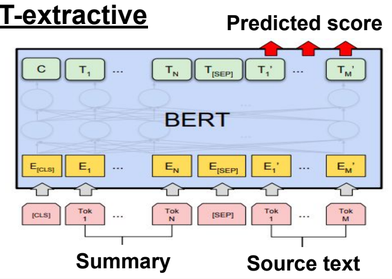
MODEL ARCHITECTURES

- PGEN-abstractive: Sequence-to-sequence model with attention
- BERT-extractive: BERT language model with summarization layers

PGEN-abstractive



BERT-extractive

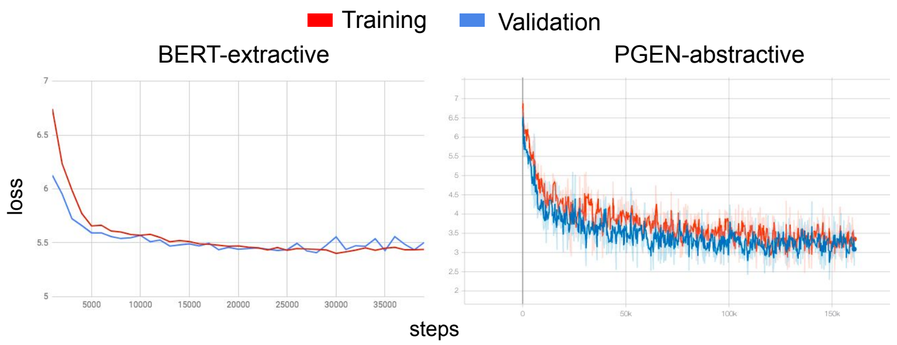


RESULTS & ANALYSIS

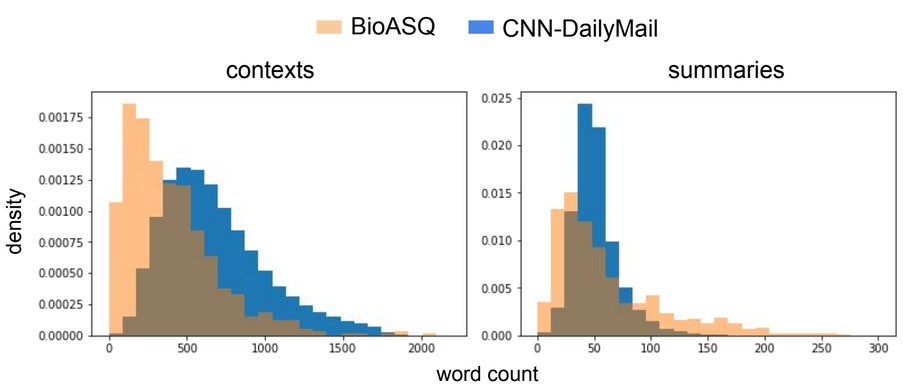
Comparison of CNN-DailyMail and BioASQ with ROUGE

Model	Test dataset	ROUGE-1	ROUGE-2	ROUGE-L
BERT-extractive	CNN-DailyMail	43.16	20.22	39.56
	BioASQ	45.85	32.20	39.93
PGEN-abstractive	CNN-DailyMail	35.39	15.11	32.97
	BioASQ	32.85	17.74	25.54

Model training and validation loss



Word count distributions for contexts and summaries



Examples of predicted candidate summary

CONTEXT:
pyroptosis is an inflammasome-mediated programmed cell death pathway triggered in macrophages by a variety of stimuli, including intracellular bacterial pathogens. C. albicans triggers pyroptosis, a proinflammatory macrophage death. pyroptosis is a caspase-1 dependent pro-inflammatory form of programmed cell death associated with pyroptosis, the pro-inflammatory programmed cell death. our study here identified a novel cell death, pyroptosis in ox-LDL induced human macrophage, which may be implicated in lesion macrophages death and play an important role in lesion instability. caspase-1 induced pyroptosis is an innate immune effector mechanism against intracellular bacteria.

SUMMARY (Ground Truth):
pyroptosis is an inflammasome-mediated programmed cell death pathway.

CANDIDATE (BERT-extractive):
pyroptosis is a caspase-1-dependent pro-inflammatory form of programmed cell death. Caspase-1-induced pyroptosis is an innate immune effector mechanism against intracellular bacteria.

CANDIDATE (PGEN-abstractive):
pyroptosis cell death, pyroptosis in ox-ldl induced human macrophage, may be implicated in lesion macrophages cell death.

Attention visualization

Article

__pyroptotic__ cell death , . __pyroptotic__ cell death . __pyroptosis__ is an __inflammasome-mediated__ programmed cell death pathway triggered in __macrophages__ by a variety of stimuli , including __intracellular__ bacterial pathogens . __pyroptotic__ death . c. __albicans__ triggers __pyroptosis__ , a __proinflammatory__ __macrophage__ death . __pyroptosis__ is a __caspase-1-dependent__ __pro-inflammatory__ form of programmed cell death . associated with __pyroptosis__ , the __pro-inflammatory__ programmed cell death . our study here identified a novel cell death , __pyroptosis__ in __ox-ldl__ induced human __macrophage__ , which may be implicated in lesion __macrophages__ death and play an important role in lesion instability . __pyroptotic__ cell death . __caspase-1-induced__ __pyroptotic__ cell death . __caspase-1-induced__ __pyroptosis__ is an innate immune __effector__ mechanism against __intracellular__ bacteria .

Reference summary

__pyroptosis__ is an __inflammasome-mediated__ programmed cell death pathway .

Generated summary (highlighted = high generation probability)

pyroptosis cell death , pyroptosis in ox-ldl induced human macrophage , may be implicated in lesion macrophages cell death , pyroptosis cell death , pyroptosis in ox-ldl induced human macrophage , which may be implicated in lesion macrophages death .

prob = 0.249

CONCLUSION

- Models overwhelmingly trained on CNN-DailyMail achieve comparable text summarization results on BioASQ dataset.
- Higher ROUGE scores for the BioASQ dataset in the extractive model are likely due to differences in distribution of source text and summary lengths. Slightly lower ROUGE scores in the abstractive model may represent insufficient biomedical text training.