Supplementary: Document Modeling with External Attention for Sentence Extraction

Shashi Narayan[*]
University of Edinburgh
shashi.narayan@ed.ac.uk

Ronald Cardenas[*]
Charles University in Prague
ronald.cardenas@matfyz.cz

Nikos Papasarantopoulos[*]
University of Edinburgh
nikos.papasa@ed.ac.uk

Shay B. Cohen **Mirella Lapata**
University of Edinburgh
{scohen,mlap}@inf.ed.ac.uk

Jiangsheng Yu **Yi Chang**
Huawei Technologies
{jiangsheng.yu,yi.chang}@huawei.com

1 Extractive document Summarization

1.1 An Example CNN Article with External Information

It is a challenging task to rely only on the main body of the document for extraction cues, as it requires document understanding. Documents in practice often have additional information, such as the title, image captions, videos, images and twitter handles, along with the main body of the document. These types of information are often available for newswire articles. Figure 1 shows an example of a newswire article taken from CNN (CNN.com). It shows the additional information such as the title (first block) and the images with their captions (third block) along with the main body of the document (second block). The last block shows a manually written summary of the document in terms of "highlights" to allow readers to quickly gather information on stories. As one can see in this example, gold highlights focus on sentences from the fourth paragraph, i.e., on key events such as the "PM's resignation", "bribery scandal and its investigation", "suicide" and "leaving an important note". Interestingly, the essence of the article is explicitly or implicitly mentioned in the title and the image captions of the document.

1.2 An Example of Summaries for Human Evaluation

Figure 2 shows output summaries from various systems for the article shown in Figure 1. As can be seen, both XNET and POINTERNET were able to select the most relevant sentence for the summary from anywhere in the article, but XNET is better at producing summaries which are close to human authored summaries.

1.3 Implementation Details

We used the CNN training data to train word embeddings using the *Word2vec* skip-gram model (Mikolov et al., 2013) with context window size 6, negative sampling size 10 and hierarchical softmax 1. For known words, word embedding variables were initialized with pre-trained word embeddings of size 200. For unknown words, embeddings were initialized to zero, but optimized during training. All sentences, including titles and image captions, were padded with zeros to a sentence length of 100. For the convolutional sentence encoder, we followed Kim et al. (2016), and used a list of kernels of widths 1 to 7, each with output channel size of 50. This leads the sentence embedding size in our model to be 350. For the recurrent neural network component in document encoder and sentence extractor, we used a single-layered LSTM network with size 600. All input documents were padded with zeros to a maximum document length of 126. For each document, we consider a maximum of 10 image captions. We experimented with various numbers (1, 3, 5, 10 and 20) of image captions on the validation set and found that our model performed best with 10 image captions. We performed mini-batch cross-entropy training with a batch size of 20 documents for 10 training epochs. After each epoch, we evaluated our model on the validation set and chose the best performing model for the test set. We trained our models with the optimizer Adam (Kingma and Ba, 2015) with initial learning rate 0.001. Our system is fully implemented in TensorFlow (Abadi

[*] The first three authors made equal contributions to the paper. The work was done when the second author was visiting Edinburgh. Our TensorFlow code and datasets are publicly available at https://github.com/shashiongithub/Document-Models-with-Ext-Information.

South Korean Prime Minister Lee Wan-koo offers to resign

Seoul (CNN) South Korea's Prime Minister Lee Wan-koo offered to resign on Monday amid a growing political scandal.

Lee will stay in his official role until South Korean President Park Geun-hye accepts his resignation. He has transferred his role of chairing Cabinet meetings to the deputy prime minister for the time being, according to his office.

Park heard about the resignation and called it "regrettable," according to the South Korean presidential office.

Calls for Lee to resign began after South Korean tycoon Sung Woan-jong was found hanging from a tree in Seoul in an apparent suicide on April 9. Sung, who was under investigation for fraud and bribery, left a note listing names and amounts of cash given to top officials, including those who work for the President.

Lee and seven other politicians with links to the South Korean President are under investigation. cont...

South Korean PM offers resignation over bribery scandal

Suicide note leads to government bribery investigation

- Calls for Lee Wan-koo to resign began after South Korean tycoon Sung Woan-jong was found hanging from a tree in Seoul
- Sung, who was under investigation for fraud and bribery, left a note listing names and amounts of cash given to top officials

Figure 1: A CNN news article with story highlights and additional information. The second block is the main body of the article. It comes with additional information such as the title (first block) and the images with their captions (third block). The last block is the story highlights that assist in gathering information on the article quickly. These highlights are often used as the gold summary of the article in summarization literature.

LEAD
- Seoul (CNN) South korea's Prime Minister Lee Wan-koo offered to resign on monday amid a growing political scandal
- Lee will stay in his official role until South Korean President Park Geun-hye accepts his resignation
- He has transferred his role of chairing cabinet meetings to the deputy Prime Minister for the time being , according to his office

POINTERNET
- South Korea's Prime Minister Lee Wan-koo offered to resign on Monday amid a growing political scandal
- Lee will stay in his official role until South Korean President Park Geun-hye accepts his resignation
- Lee and seven other politicians with links to the South Korean President are under investigation

XNET
- South Korea's Prime Minister Lee Wan-Koo offered to resign on Monday amid a growing political scandal
- Lee will stay in his official role until South Korean President Park Geun-hye accepts his resignation
- Calls for Lee to resign began after South Korean tycoon Sung Woan-jong was found hanging from a tree in Seoul in an apparent suicide on April 9

HUMAN
- Calls for Lee Wan-koo to resign began after South Korean tycoon Sung Woan-jong was found hanging from a tree in Seoul
- Sung, who was under investigation for fraud and bribery, left a note listing names and amounts of cash given to top officials

Figure 2: Summaries produced by various systems for the article shown in Figure 1.

et al., 2015).[1]

2 Answer Selection for Machine Reading Comprehension

2.1 An Example: LRXNET vs ISF

In Table 1 we contrast LRXNET with ISF on a question from the NewsQA dataset. We see the main pitfall behind using ISF in an isolated man-

[1] Our TensorFlow code is publicly available at https://anonymized.com.

	pos	score	sentence
LRXNet	3	0.49	Raymond UNK Chuck UNK Foster , 44 , was indicted on **second-degree murder charges** Wednesday in the November death of Cynthia Lynch , 43 , of Tulsa , Oklahoma.
	7	0.16	Two other men , including Foster 's son , Shane Foster , were indicted on a count of obstruction of justice , and a woman , Danielle Jones , was indicted on one count of being an accessory after the fact.
	1	0.12	Raymond UNK Chuck UNK Foster is reputed to be the leader of the Klan Group.
ISF	29	3.57	Officials tracked down those two members and arrested them , then arrested others at the campsite and Foster.
	7	3.18	Two other men , including Foster 's son , Shane Foster , were indicted on a count of obstruction of justice , and a woman , Danielle Jones , was indicted on one count of being an accessory after the fact .
	8	3.12	Wood said Thursday Foster 's case was assigned to one judge and the others ' cases were assigned to another.

Table 1: Example of a case in which LRXNET returns the correct sentence for the question *What charges does UNK Chick UNK Foster 's son and two others face?*. The ISF score, however, does not identify the correct sentence because of large overlap between the question and incorrect sentences. "pos" denotes the sentence index in the document, "score" denotes LRXNET's score or the normalized ISF score. Answer is in bold.

ner: it chooses sentences that have large lexical overlap with the question, but that do not necessarily contain the direct answer. LRXNET, on the other hand, with its machine reading capability, is able to fine tune the ISF score and identify the correct answer.

2.2 Optimization and Implementation Details

We use the One Billion Words Benchmark (Chelba et al., 2013) to pre-train word embeddings of size 200 using the *word2Vec* skip-gram model (Mikolov et al., 2013) with context window size 6 and negative sampling size 10. Known words are initialized with the pre-trained embeddings and unknown words are learned during training. All architectures were optimized using Adam (Kingma and Ba, 2015).

PAIRCNN uses the same filter lengths as XNet+ for each dataset, with feed forward layer of size 100.

Given the size of the considered datasets -except WikiQA-, we subsample each training set in order to conduct hyper-parameter search and ablation studies. We leverage the fact that NewsQA and SQuAD present several questions per document. We randomly pick one instance per document using uniform distribution without replacement. For MSMarco, we uniformly pick 12% of the training set without replacement. The resulting subsampled training sets consists of 11,468 instances for NewsQA; 7,709 for SQuAD; and 8,608 for MSMarco. We perform hyper-parameter optimization for each dataset using the Bayesian optimization framework SigOpt.

The number of sentences per document and the number of tokens per sentence are determined by the 95% cut-off value in the corresponding cumulative frequency plot based on the whole training set. For NewsQA, these values are 64 and 50, respectively. For SQuAD, 80 and 16. For WikiQA, 100 and 30. For MSMarco, 150 and 10.

XNET and XNET+ were trained for 20 epochs without dropout for all datasets. For XNET, the gradients were clipped to 10, whereas no gradient clipping was applied for XNET+. For both architectures and all datasets, filter lengths from 5 to 8 performed best in the convolutional layer.

For XNET, the batch size is set to 22, 56, 64, 62; learning rate to $2 \times 10^{-4}, 5.3 \times 10^{-5}, 6.3 \times 10^{-5}, 2 \times 10^{-5}$; LSTM state size is set to 447, 1446, 1021, 1521; sentence embedding size to 80, 368, 556, 852; for WikiQA, SQuAD, NewsQA, and MSMarco respectively.

For XNET+, the batch size is set to 44, 20, 42, 50; learning rate to $1.7 \times 10^{-4}, 1.7 \times 10^{-3}, 1.4 \times 10^{-5}, 1.9 \times 10^{-5}$; LSTM state size is set to 100, 100, 100, 121; sentence embedding size to 164, 796, 1024, 528; for WikiQA, SQuAD, NewsQA, and MsMarco respectively.

LRXNET is trained with a L_2 regularization parameter $C = 100, 10, 10, 0.01$ for WikiQA, SQuAD, NewsQA, and MSMarco respectively.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh

Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005* .

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*. Phoenix, Arizona USA, pages 2741–2749.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, California, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. pages 3111–3119.