Abstractive Text Summarization for Email Subject Line Generation

Ying Jin
yj1461@nyu.edu
Center for Data Science
New York University

Danfeng Li
dl3983@nyu.edu
Center for Data Science
New York University

Shuyu Wang
sw2860@nyu.edu
Center for Data Science
New York University

Yuntian Ye
yy1947@nyu.edu
NYU Shanghai
New York University

Abstract

Informal and noisy text summarization is a meaningful task with practical use potentials. In this project, we adopted the data-driven attention-based neural model on the task to generate email titles. The experiment evaluations show that our model achieves 21.22 Rouge-L and 17.10 BLEU. We showed that the attention-based model outperforms the extractive methods, though informal text summarization remains to be a challenge.

1 Introduction

The recent developments in text summarization of natural language understanding has been focusing on clean and formal text such as news headline generation task. However, a large collection of informal text, such as emails and group chats, is not well studied by modern summarization techniques. In practice, text summarization would be useful to individuals when applied to informal texts from everyday conversations. Informal languages differ from formal languages in the tone, vocabulary choice and oftentimes grammatical structure, and thus summarization of informal text is technically challenging.

In this work[1], we focused on informal text summarization, specifically conversational paragraphs in the form of emails (Zajic et al., 2008). The result of this research would have the potential to be generalized into other forms of conversational text.

Inspired by the recent developments in abstractive sentence summarization, we proposed using the new encoder-decoder neural network with an attention-based encoder as seen in Rush et al. (2015) on the task of summarizing email subject line. This model adopted the attention-based system in neural machine translation task (Bahdanau et al., 2014), as described in detail in Section 3. The goal is to outperform extractive models on email text - a broader genre in comparison with formal newswire text.

2 Background

Similar to neural machine translation, text summarization is equivalent to taking an input words $x_1, \cdots, x_m$ of length M and finding an output y that maximizes the probability of $p(y|x)$. When translation models seek to output y in a different language, summarization models seek to generate a shorter y with length $L < M$. Words in both x and y are coming from the same set of indexed vocabulary. Sentences are presented as series of tokens' indexes.

The parametrized model was fit by the training data to maximize the conditional probability of body-title pairs. Then the model uses greedy decoder to output sentence that maximize the conditional probability given any new input text.

The abstractive nature of the encoder-decoder model comes from the search in optimal sequence from the set of all possible sentences $\mathcal{Y}$,

$$\arg\max_{y \in \mathcal{Y}} s(x, y) \tag{1}$$

where $s : \mathcal{X} \times \mathcal{Y} \mapsto R$ is a scoring function. Whereas extractive summarization models simply transfer words from input to output by keeping only those that maximize a score function,

$$\arg\max_{m \in \{1,\cdots,M\}^N} s(x, x_{[m_1,\cdots,m_N]}) \tag{2}$$

(Rush et al., 2015), upon which Rush et al. gives a much more comprehensive probabilistic definition.

[1]The code for data processing, model training and evaluation is made public in our GitHub repository: NLU-email-title.

3 Model

We adopted Facebook's abstractive attention based system (Rush et al., 2015, ABS) which is inspired by the work done in neural machine translation (Bahdanau et al., 2014). The model utilized an attention based encoder-decoder network to generate an output probability. Given the output matrix, the result is computed using greedy decoding. We also experimented with both random and pretrained word embedding. The result of the network is compared with a simple baseline, TextRank, which extracts the most important sentences with clue words (Carenini et al., 2007). The network is trained using PyTorch.

3.1 Baseline

In this work, we used TextRank Model (Mihalcea and Tarau, 2004; Barrios et al., 2016) as baseline to test whether the state-of-art abstractive model outperforms. TextRank is a extractive graph-based algorithm that generates importance of vertices within the graph by recursively computing the global information of the graph. A vertex has higher importance if more links are on the vertex (Mihalcea and Tarau, 2004). In our work, texts are modeled as the weighted directed graph. This model was widely used before neural models surged.

Other variations of TextRank such as BM25 / Okapi-BM25 were also introduced as the state of art information retrieval which is used in the implementation for Summa (Barrios et al., 2016).

3.2 Attention-Based Model

Our Attention-based Encoder-Decoder Model follows the same strategy as in Rush et al. (2015) by modifying the implementation of the work in neural machine translation as in Bahdanau et al. (2014). The model architecture consists of a Bi-GRU encoder and an attention-based decoder.

Encoder-Decoder Similar to attention-based neural machine translation model, we used a learned soft alignment between the input and the summary instead of a single representation of the input sentence. We utilized bi-directional RNN, specifically a bi-GRU (Gated Recurrent Unit). After computing the decoder state, we got the probability of the target word y_i by applying a softmax function. Cross-entropy loss is used to maximize the probability of selecting the correct word at each step.

Attention To alleviate the problem of a decline in performance when dealing with long sentences, we used an attention mechanism to allow the model only to pay attention to the most relevant input sentences. Specifically, we used the MLP with tanh-activation to both encoder and decoder hidden states to get the attention energy.

4 Data Source

The Enron Corpus (Klimt and Yang, 2004) were used in this work, as it is a public, uncensored, naturally generated email corpus from a corporate environment. Early works mainly focused text extraction on the dataset, providing critical insights for our study. One of the previous works (Carenini et al., 2008) focused on the sentence extraction method for using graph-based summarization approach. Another more recent study (Yousefi-Azar and Hamey, 2017) focuses on query-oriented extractive approach using unsupervised deep learning.

Considering the purpose of our study, we decided to make use of only first initiated emails as they are the most relevant to subject line content. The original corpus[2] contains 517,401 emails and was reduced to 111,044 after we removed any forwarded, replied, and duplicate emails. This dataset is then split into train, development and test sets (7:2:1), and the original subject line were compared with machine generated text in evaluation.

5 Training

Two variances of TextRank are used, one is the original type which retrieves top 10 keywords (TextRank4Keywords[3]) and the other is a improved version which generates a summary with length of 15 for each given text (summa[4] text rank). Both are unsupervised models and thus no training is required.

To speed up training for the abstractive model, only emails with less than 500 tokens and words occurring more than 5 times were included. This also helped us eliminate newsletter-typed emails.

[2]Enron corpus downloadable here: enron-email-dataset
[3]TextRank for Keywords: TextRank4Keywords
[4]Python implementation for TextRank: summa

	TextRank4Keywords			Summa			Attention-based System		
Rouge-1	P: 13.06	R: 32.30	F1: 17.06	P: 5.96	R: 23.86	F1: 8.86	P: 17.00	R: 25.85	F1: **19.43**
Rouge-2	P: 1.83	R: 5.10	F1: 2.41	P: 2.49	R: 10.44	F1: 3.68	P: 9.33	R: 13.22	F1: 10.54
Rouge-3	P: 0.39	R: 1.18	F1: 0.52	P: 1.31	R: 5.09	F1: 1.90	P: 6.45	R: 8.85	F1: 7.28
Rouge-4	P: 0.12	R: 0.40	F1: 0.17	P: 0.83	R: 3.00	F1: 1.18	P: 4.98	R: 6.77	F1: 5.64
Rouge-L	P: 14.47	R: 31.38	F1: 18.55	P: 7.37	R: 23.91	F1: 10.63	P: 18.78	R: 26.83	F1: **21.22**
Rouge-W	P: 10.50	R: 21.32	F1: 12.39	P: 5.32	R: 16.94	F1: 7.23	P: 16.60	R: 20.52	F1: 16.86
BLEU	0.21			1.71			**17.10**		

Table 1: Validation set results on ROUGE and BLEU.

All words are turned into lower case, then tokenized with SpaCy's tokenizer. The vocabulary of size 32,513 are built from the training body of emails with randomly initialized word embeddings. We've also tested vocabularies and pretrained word embeddings from GloVe[5] and Senna[6], but the result did not improve compared to random initialization. Many important words in our context, such as '1st' or words with typo, are not included in the vocabulary, resulting in the model predicting <unk> (unknown) token repetitively.

The encoder-decoder model implementation is adopted from the Pytorch implementation of Bastings (2018) with minor changes and adjustments to fit our work better.

6 Results

ROUGE and BLEU Evaluation One of the most popular evaluation metrics for abstractive summarization is ROUGE[7], which evaluates the overlaps between generated text and the original subject line. In addition to ROUGE, we also reported BLEU[8] score, a measure of how likely a given language model will predict the test data correctly.

Our experiments shows that the model performs the best on ROUGE-L when the embedding size is around 300 and the hidden dimension is around 256. With Adam optimizer, the best learning rate is found to be near 0.0003.

We reported the validation results from the two baselines and the attention-based system in Table 1. The attention based system outperforms the two extractive systems significantly, especially on its ability in predicting relevant n-grams.

[5]GloVe Vectors

[6]Senna Embeddings

[7]Python package for ROUGE evaluation: py-rouge

[8]Python package for BLEU evaluation: sacrebleu

Human Evaluation We also performed a DUC-style human-evaluations of 50 random samples from our TextRank and Attention models' predictions. This framework is suggested by Liu et al. (2018). Each of the samples was rated by six individual volunteers. Five metrics related to our tasks are assessed: focus (F), grammaticality (G), non-redundancy (N-d), referential clarity (RC), and structure&coherence (S&C). They are evaluated on the scale of 1 to 5. The aggregated average rating is shown in table 2.

	F	G	N-d	RC	S&C
TextRank4Keywords	1.06	0.96	0.98	1.01	0.85
Summa	1.59	2.38	1.32	1.49	1.89
Attention-based	3.13	3.23	3.29	2.68	3.08

Table 2: Linguistic quality human evaluation scores on scale 1-5 (higher is better).

Despite unavoidable subjectivity in human evaluation, we observed that the Attention Based System outperforms the baselines in every aspect, especially in non-redundancy and coherence since the attention can better capture the key idea of the original texts. Summa performs well in grammar since it extracts part of the text which preserves the integrality of original sentences and regular. TextRank is not desired in this case since center words might not reflect the main ideas correctly.

Examples of Prediction Table 3 compares the results of the attention based system on the informal Enron corpus with the two formal datasets, DUC-2004 and Gigaword[9]. Note that although the Rouge-1 and Rouge-L evaluation on the informal corpus is much lower than that of formal text, it performs comparably well on Rouge-2 evaluation.

Issues such as typo and arbitrary acronyms are

[9]Results are reported by Rush et al. (2015).

found frequently in the email text, which add to the challenge for text summarization.

	Formal		Informal
	DUC-2004	Gigaword	Enron
Rouge-1	28.18	31.00	19.43
Rouge-2	8.49	12.65	10.54
Rouge-L	23.81	28.34	21.22

Table 3: Result comparison: Formal vs. Informal.

Figure 1 shows 4 representative examples in different context. In general, attention-based model performs much better than its extractive counterpart, as seen in the first three examples. As TextRank, an unsupervised extractive method, picks the sentence directly from source based on similarity measure. The resulting summary tends to be lengthy, and does not correspond to a subject line title that fits the common best practice. The attention model learns the composition structure of email title from the training set, thus generates results that are much better following the structures of a title. In the first email, the machine generated title can be considered as even better than its original title. It not only follows the correct grammatical structure, but also the common best practice of using a Noun or Noun Phrase in the title.

We also identified some issues from the evaluation. As seen in email 2, one of the most important keyword 'Free' is not captured in the title, mainly due to the fact that it's not written specifically in the body of email. It may be inferred from 'will be provided'. However, that would require the machine to have a much deeper understanding for the semantics.

The third email showcases the machine's ability to capture important phrase from structured text (calendar entry). In the fourth example, the original email body is an informal greeting message with only one sentence of fifteen words. TextRank directly copies the whole sentence without any modification. The attention model's prediction for email 4 reveals issues with words that are out of the vocabulary. One of the ways to alleviate this issue would be the use of a larger training dataset, therefore having a more vocabulary, since the current vocabulary built from training is only around 32k. Another way would be to build a custom vocabulary that includes not only general words, but also slang, acronyms, common typos of words and emoticons. Introducing a better data cleaning strategy that corrects typos may also help. This highlights some future challenges in obtaining human-level summarizations for informal texts.

Email 1: we have received and executed assignment and assumption agreement effective as of february 1, 2001, by and among midcoast energy resources, inc., midcoast marketing, inc. and enron north america corp., whereby midcoast energy resources, inc. assigned its transactions to midcoast marketing, inc. copies will be distributed. samantha m. boyd sr. legal specialist enron north america, corp. 1400 smith, eb3802a houston, tx 77002 phone: (713) 853-9188 fax: (713) 646-3490 email: samantha.boyd@enron.com
Subject: midcoast - assignment
TextRank: and enron north america corp. legal specialist enron north america, corp. 1400 smith, eb3802a houston, tx 77002 phone: (713) 853-9188 fax: (713) 646-3490 email: samantha.boyd@enron.com
Attention: assignment of merrill lynch agreement

Email 2: thanks for all your hard work and happy birthday! lunch will be provided on friday, april 13, by tim belden and chris calger to everyone on the floor as a thanks for all you've done for enron this month. we'll also celebrate this month's birthdays by having cookies for everyone.
Subject: free lunch on friday, april 13
TextRank: we'll also celebrate this month's birthdays by having cookies for everyone.
Attention: lunch friday, april 13

Email 3: calendar entry: appointment description: 1st group s. central team luncheon w/ shelley date: 5/10/2001 time: 11:30 am - 1:00 pm (central standard time) chairperson: outlook migration team detailed description :...
Subject: 1st group s. central team luncheon w/ shelley
TextRank: calendar entry: appointment description: 1st group S.
Attention: 1st group s. central team luncheon w/ shelley

Email 4: hi jeff, kimberley, happy holidays. have fun! see you guys next year :-) -deepak
Subject: happy holidays!
TextRank: hi jeff, kimberley, happy holidays. have fun! see you guys next year :-) -deepak
Attention: <unk> [a]

[a] <unk> stands for unknown token

Figure 1: Examples subject lines from the test set, with comparisons of the original subject line and the title produced by our models.

Collaboration statement

Everybody on the team contributes equally to the project and report writing.

Yuntian Ye: TextRank4Keywords, Summa, and other baseline tests

Ying Jin: data cleaning, attention based encoder-decoder model

Shuyu Wang: human-evaluation

Danfeng Li: Github,

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv e-prints* abs/1409.0473. https://arxiv.org/abs/1409.0473.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *CoRR* abs/1602.03606. http://arxiv.org/abs/1602.03606.

Joost Bastings. 2018. The annotated encoder-decoder with attention.

Giuseppe Carenini, Raymond T. Ng, and Xiaodong Zhou. 2007. Summarizing email conversations with clue words. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '07, pages 91–100. https://doi.org/10.1145/1242572.1242586.

Giuseppe Carenini, Raymond T. Ng, and Xiaodong Zhou. 2008. Summarizing emails with conversational cohesion and subjectivity. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, pages 353–361. http://aclweb.org/anthology/P08-1041.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*. Springer-Verlag, Berlin, Heidelberg, ECML'04, pages 217–226. https://doi.org/10.1007/978-3-540-30115-8_22.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *CoRR* abs/1801.10198. http://arxiv.org/abs/1801.10198.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of EMNLP 2004*. Association for Computational Linguistics, Barcelona, Spain, pages 404–411. https://www.aclweb.org/anthology/W04-3252.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *CoRR* abs/1509.00685. http://arxiv.org/abs/1509.00685.

Mahmood Yousefi-Azar and Len Hamey. 2017. Text summarization using unsupervised deep learning. *Expert Systems with Applications* 68:93 – 105. https://doi.org/https://doi.org/10.1016/j.eswa.2016.10.017.

David M. Zajic, Bonnie J. Dorr, and Jimmy J. Lin. 2008. Single-document and multi-document summarization techniques for email threads using sentence compression. *Inf. Process. Manage.* 44:1600–1610.