Extractive Summarization Using Multi-Task Learning with Document Classification

Masaru Isonuma,Toru Fujino, Junichiro Mori, Yutaka Matsuo and **Ichiro Sakata**
The University of Tokyo, Japan
{isonuma, jmori, isakata}@ipr-ctr.t.u-tokyo.ac.jp,
tr.fujino@scslab.k.u-tokyo.ac.jp, matsuo@weblab.t.u-tokyo.ac.jp

Abstract

The need for automatic document summarization that can be used for practical applications is increasing rapidly. In this paper, we propose a general framework for summarization that extracts sentences from a document using externally related information. Our work is aimed at single document summarization using small amounts of reference summaries. In particular, we address document summarization in the framework of multi-task learning using curriculum learning for sentence extraction and document classification. The proposed framework enables us to obtain better feature representations to extract sentences from documents. We evaluate our proposed summarization method on two datasets: financial report and news corpus. Experimental results demonstrate that our summarizers achieve performance that is comparable to state-of-the-art systems.

1 Introduction

With rapid increase in the volume of textual data that are available both online and offline, the need for automatic document summarization that can be implement in practical scenarios is increasing (Li et al., 2016; Chopra et al., 2016; Takase et al., 2016). Among the several summarization systems, extractive summarization approaches (Erkan and Radev, 2004; McDonald, 2007; Wong et al., 2008) are widely used. These techniques identify and subsequently concatenate relevant sentences automatically from a document to create its summary while preserving its original information content. Such approaches are popular and widely used for practical applications because they are computationally cost-effective and less complex. Extractive summarization approaches based on neural network-based approaches (Kågebäck et al., 2014; Cao et al., 2015; Yin and Pei, 2015; Cao et al., 2016) have advanced rapidly. Recently, an attentional encoder-decoder for extractive single-document summarization was proposed and its application to the news corpus was demonstrated (Cheng and Lapata, 2016; Nallapati et al., 2017).

The neural network-based approaches rely heavily on large amounts of reference summaries for training neural models, and consequently, for tuning a large number of parameters. The reference summaries are manually or semi-automatically created in advance. Some existing studies employ parallel corpora as artificial reference summaries (Woodsend and Lapata, 2010; Cheng and Lapata, 2016). However, preparing such large volumes of reference summaries manually is sometimes costly. Particularly, it is infeasible for humans to create hundreds of thousands of reference summaries in cases where summarization requires domain-specific or expert knowledge. Such cases include financial reports, financial and economic news (Filippova et al., 2009), and scientific articles (Parveen et al., 2016).

A fundamental requirement in extractive summarization is the identification of salient sentences from a document, i.e., sentences that represent key subjects mentioned in the document. Such subjects are often described in the form of topics, categories, sentiments, and other meta-information about a document. Sometimes they are extracted from external information related to document contents. Once one knows the subjects of a document beforehand, a straightforward strategy in extractive summarization is to select sentences that are relevant to the subjects. Importantly, subjects should be inferred from sentences identified from

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2101–2110
Copenhagen, Denmark, September 7–11, 2017.

the document. For example, assume that we are about to summarize a financial report of a company with knowledge from external information sources that the company has strong earnings. In this case, we might select sentences that explain factors affecting increase of earnings so that a reader of the summary can intuitively understand the company's financial situation.

The key idea is that we regard the subjects of a document as pseudo-rough reference summaries. Then, if we are able to estimate the subjects with small amounts of documents and the external information in them, the identification of salient sentences from a document can be supported by sentence features that have been learned from document subject estimation. As a result, smaller amounts of actual reference summaries are only needed as mutually learning feature representations for both subject estimation and sentence identification from pseudo-rough reference summaries.

As described earlier, we focus on single document summarization with small amounts of reference summaries, and propose a general framework for summarization that is useful for extracting sentences from a document along with its external related information. Particularly, we formalize estimation of the above-described document subjects as a document classification task and solve document summarization in the framework of multi-task learning for sentence extraction and document classification.

Our proposed summarization framework comprises two components: one designed for sentence extraction, which selects sentences relevant to the subjects of an input document, and one for document classification, which predicts the subject of the input document. In the multi-task learning framework, document classification supports sentence extraction by learning common feature representations of salient sentences for summarization. We use recurrent neural network encoder–decoder as sentence extractor and document classifier.

We evaluate our proposed summarization method on two datasets: the NIKKEI, the leading financial news publisher in Japan and a financial report corpus; and the New York Times Annotated Corpus (Sandhaus, 2008). The results of experimental evaluations demonstrate that our summarizers achieve a performance that is comparable to those of state-of-the-art systems.

The contribution of this paper is two-fold. First, we propose a general framework for single document summarization with small amounts of reference summaries, which is important for practical implementation of summarization techniques. Second, we propose a multi-task learning method with curriculum learning that supports sentence extraction from a document while solving document classification. Here, we assume that a document is classifiable into certain subjects, which comprise the meta-information of the document. Furthermore, sentences for a summary are extracted in relation to the subjects.

2 Problem Statement and Data Preparation

In this section, we define the task of sentence extraction for document summarization as addressed in this paper. We specifically examine documents that satisfy the following requirements. (1) Reference summaries are few. (2) The document is associated with a list of subjects, $\{a_1, a_2, \cdots, a_m\}$ $(a_i \in \{0, 1\})$, that includes topics, categories, sentiments, and other meta-information. $a_j = 1$ denotes that the document is classified into the subject j. Given a document D consisting of a sequence of sentences $\{s_1, s_2, \cdots, s_n\}$, we aim to extract k sentences in relation to a document subject a_j, which is expected to be included in the summary ($k < n$) of the document. We predict both a subject a_j for the document and a label $y_i \in \{0, 1\}$ for each sentence within the document, which indicates that the i-th sentence should be extracted.

In this study, we use two datasets for our sentence extraction task: the NIKKEI financial report corpus and New York Times news corpus.

For the financial report corpus, we used financial reports published every quarter during 2013–2016 by Japanese exchange listed companies. The reports explained the economic activity and the factors affecting revenue or profits for the quarter. For the reference summary, which is the gold standard summary used for training a classifier that predicts a sentence label, we use financial news articles published by the NIKKEI [1]. The NIKKEI publishes articles summarizing financial reports of each company. It covers approximately 10 % of all the reports: 3911 reports from 2013 to 2016. The

[1] http://www.nikkei.com/

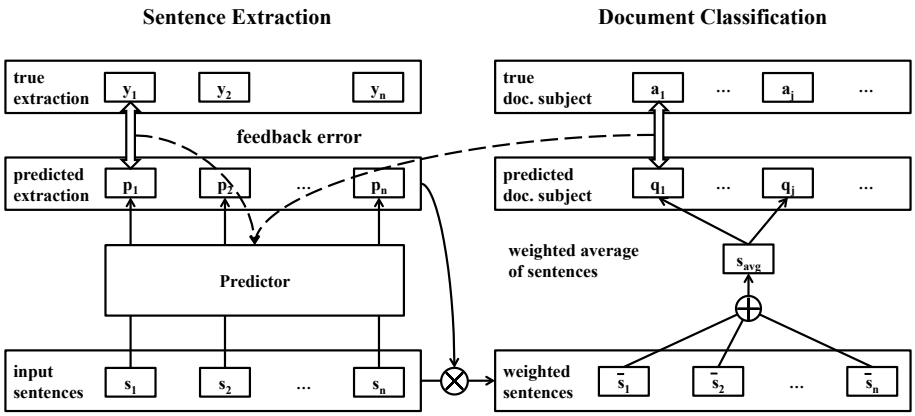

Figure 1: Proposed multi-task learning framework for sentence extraction with document classification

language of the reports and the articles is Japanese. For the document subject of a financial report, we used profit and revenue information as a subject a_j $(j = 1, 2)$, which indicates its profit and revenue increase compared to an earlier term.

Our second dataset, the New York Times Annotated Corpus (NYTAC), is a collection of articles from the New York Times. The gold standard summaries are attached to some of the articles. As the subject of a document article, we use already annotated category of the news from its metadata such as *Business* and *Arts*.

For the task of sentence extraction, the gold standard labels indicating sentences that should be extracted are needed. To attach the labels on sentences that maximize the Rouge score with respect to gold summaries, we introduce a greedy approach (Cheng and Lapata, 2016; Nallapati et al., 2017). We first select one sentence that has a maximum Rouge score with respect to the entire gold standard summary. We add it to the reference summary set and select sentences incrementally until no candidate sentence improves the score when added to the current summary set. The labels of sentences in the summary set are set as $y = 1$. The greedy approach is efficient because the computational costs associated with the identification of a global optimal summary set are too large.

The labels are attached by computing ROUGE-1 (Lin, 2004). ROUGE-1 and ROUGE-2 are reported as best for emulating evaluation by humans (Owczarzak et al., 2012). For financial reports, words apart from nouns, verbs, adjectives, and adverbs are removed for computing appropriate ROUGE scores. The accuracy between the labels attached by ROUGE-1 and humans is 81%.

3 Summarization Model

This section introduces our novel summarization framework. Figure 1 presents the proposed multi-task learning framework for sentence extraction with document classification. The left half of the figure shows the common sentence extraction part. The right half depicts a novel sentence extraction by document classification. We assume that a document is classifiable into certain subjects that represent meta-information of the document, and assume that sentences for a summary are extracted in relation to the subjects. Therefore, solving document classification supports sentence extraction from a document with multi-task learning of both tasks.

In Fig. 1, s_i denotes the embedding of sentence i. Furthermore, y_i denotes whether the sentence should be extracted and a_j is a subject of a document, which includes topics, categories, sentiments, or other meta-information. The predictor component computes $p_i \equiv p(y_i = 1 \mid D)$, the probability of sentence i extraction. Our proposed method estimates p_i by learning both sentence extraction and document classification.

We now explain how learning document classification supports sentence extraction. In Fig. 1, s_{avg} is the weighted average of s_i in terms of the probability of sentence extraction. It means that s_{avg} includes much more information about sentences with higher extraction probability. The probability that the document is related to a subject $q_j \equiv p(a_j = 1 \mid D)$ is estimated by s_{avg}. The error is larger if the contents of extracted sentences do not correspond with the document subject. By feeding back this error to the predictor, the model learns to extract sentences related to

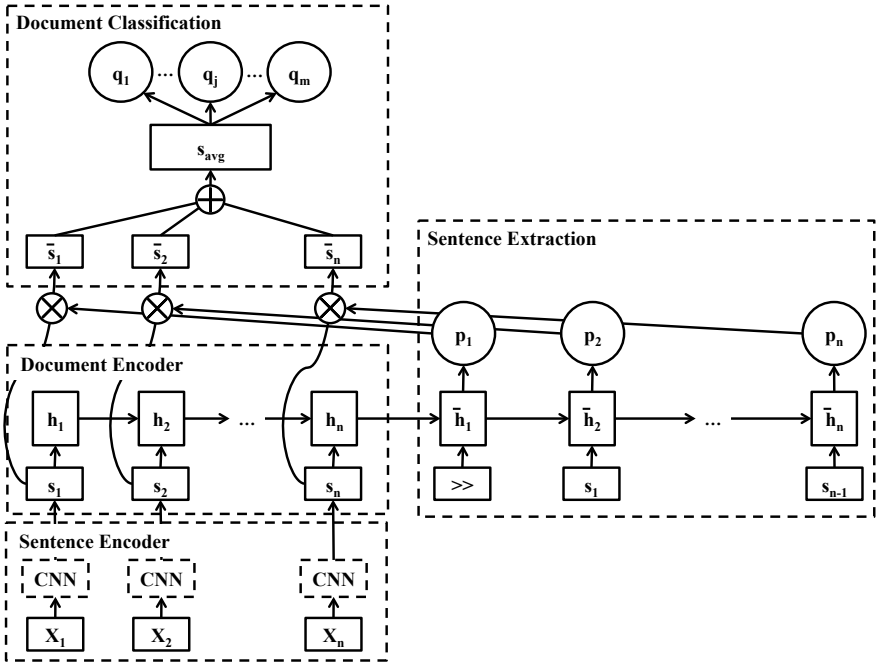

Figure 2: Sentence extraction model using LSTM-RNN with multi-task learning

the document subject. For example, in the case of a financial report, profit information indicating profit and revenue increase compared with an earlier term is used as a document subject. Positive sentences are expected to be extracted so that the extracted sentences reflect good financial results if the profit and revenue increase.

Figure. 2 shows the entire model. With respect to the predictor component in the proposed model, we use an encoder–decoder architecture modeled by recurrent neural networks (Kim et al., 2016) based on recent neural extractive summarization approaches (Cheng and Lapata, 2016; Nallapati et al., 2017). However, our summarization framework is applicable to all models of sentence extraction using distributed representation as inputs. We explain four sub-modules of the summarization model: sentence encoder, document encoder, sentence extraction, and document classification.

3.1 Sentence Encoder

We use Convolutional Neural Network (CNN) to obtain a sentence embedding from word embeddings. The training speed of single-layer CNN is high. It is effective for sentence-level classification such as sentiment analysis (Kim, 2014). Actually, CNN is suitable for use because our model requires a high computational cost. Sentence embeddings are used for both sentence extraction and document classification.

Let $x_i \in \mathcal{R}^d$ denote the embedding of the i-th word in the sentence, and $x_{i:i+q-1} \in \mathcal{R}^{dq}$ represent a concatenated vector that represents a sequence of q words. Convolutional filter $w \in \mathcal{R}^{dq}$ is applied as

$$s_w^i = f(w \cdot x_{i:i+q} + b), \tag{1}$$

where f is a nonlinear function such as the hyperbolic tangent and b is the bias. Max pooling over time is applied to obtain a single feature s_w representing the sentence under filter w.

3.2 Document Encoder and Extractor

The LSTM-RNN Encoder–Decoder model is used for sentence extraction. First, on the encoder part, all sentences of a document are input into the hidden layers of RNN. LSTM assigns the input gate, forget gate, output gate, and memory cells as activation functions of RNN.

In Fig. 2, $h_n \in \mathcal{R}^k$ is the output of the encoder part, for which information of all sentences is input. The extraction probability is estimated based on the encoder part output. The hidden layer of the decoder part $\bar{h}_t \in \mathcal{R}^k$ is updated by LSTM equal to the encoder part. The initial value $\bar{h}_0$ is h_n. The input is the prior sentence s_{t-1} multiplied by the extraction probability p_{t-1}. Therefore, more information about sentences that are likely to be extracted is input to the hidden layer. Based on each

hidden layer, the extraction probability of sentence t is computed as shown below.

$$h_t = \mathrm{LSTM}(s_t, h_{t-1}) \tag{2}$$

$$\bar{h}_t = \mathrm{LSTM}(p_{t-1} \cdot s_{t-1}, \bar{h}_{t-1}) \tag{3}$$

$$p_t = \sigma(w_y \cdot [h_t : \bar{h}_t] + b_y) \tag{4}$$

Here, $w_y \in \mathcal{R}^{2k}$ represents the weight vector, b_y stands for the bias, and : is the concatenation operator of vectors. By concatenating the hidden layers of the decoder part with the encoder part, the extraction probability is computed by referencing the information of input sentences more directly. It seems reasonable to pay attention to the input sentence directly when deciding whether the sentence should be extracted or not.

3.3 Document Classification

Using the embeddings and estimated extraction probability of sentences, the probability distribution of input document subjects is estimated. The probability that a document is classified into the subject j, q_j, is computed as shown below.

$$q_j = \sigma(w_a \cdot s_{avg} + b_a) \tag{5}$$

$$s_{avg} = \frac{\sum_t p_t \cdot s_t}{\sum_i p_i} \tag{6}$$

Here, w_a signifies the weight vector and b_a is the bias. Additionally, s_{avg} represents the weighted average of sentence embeddings. Each sentence is weighted by the estimated extraction probability. The predictor computes the probability distribution of a document subject from s_{avg}, which means that the predictor pays more attention to sentences that are likely to be extracted.

3.4 Multi-Task Learning with Curriculum Learning

This section presents an explanation of the procedure followed to train the summarization model. The model parameters are updated to maximize the likelihood of all sentence labels and document subject labels. This is equivalent to minimization of the following error terms.

$$E_y(\theta) = -\sum_{t=1}^{n} (y_t \log p_t + (1-y_t)\log(1-p_t)) + \lambda_\theta \|\theta\|^2 \tag{7}$$

$$E_a(\theta) = -\sum_{j=1}^{m} (a_j \log q_j + (1-a_j)\log(1-q_j)) + \lambda_\theta \|\theta\|^2 \tag{8}$$

In these equations, $\|\theta\|^2$ is the L2 norm, and λ_θ is the regularization term. L2 regularization is introduced to avoid overfitting.

Multi-task learning is generally complicated because the parameter is optimized simultaneously for sentence extraction and document classification. We introduce curriculum learning (Bengio et al., 2009) to overcome this difficulty. Curriculum learning is a learning method that aims to improve the performance of a complicated model or data. The model starts by learning a simple model or data, and gradually adapts to more complicated ones.

We introduce two kinds of curriculum learning for multi-task learning. We apply baby step curriculum learning (Cirik et al., 2016), which demonstrates the effectiveness of the LSTM-RNN architecture. In this, the dataset is categorized based on the difficulty and added to the order of ease.

We divide the dataset into three subsets based on the combination of document type and objective function. The first subset has documents with an attached reference summary. The model is trained for optimizing sentence extraction. The second uses the same documents as the first. However, the objective function is document classification. The last one has documents with no reference summary. Only the likelihood of document subjects as pseudo-rough reference summaries is maximized in the last dataset. For sentence extraction task, it is more difficult to train from the last dataset than the first dataset because information related to document subjects are more truncated than the reference summary. The second dataset is the bridge between the first and the last.

For document classification, sentences are weight-averaged by the estimated sentence extraction probability p_t. In the second dataset, sentences are weighted not only by p_t, but also the true label y_t. p_t in Eq.(6) is replaced by $\bar{p}_t$ as follows.

$$\bar{p}_t = \kappa p_t + (1-\kappa) y_t \tag{9}$$

Here, κ is the mixing rate of extraction probability and the true label. At the beginning of training,

p_t is not predicted accurately. This will affect document classification adversely, therefore κ is set to nearly zero so that the true label y_t is used for document classification. By using the true label, training for sentence extraction and document classification does not mutually interfere. As training progress, κ gets larger and document classification supports sentence extraction.

We believe that our basic idea of curriculum learning, with some modifications depending on the task applied, can be applied for other kinds of multi-task learning in general.

4 Experimental Setup

4.1 Dataset

The NIKKEI Financial Report Corpus

For training and evaluation, we used financial reports published from April 2013 through September 2016. The reports used for evaluation and validation were published in the last and the second to last quarter. All other reports were used for training. The numbers of reports used for training, validation, and evaluation were $12,262$, 191, and 183. In the training dataset, $8,725$ reports with no reference summary were included and used only for training of document classification, which predicts document subjects. As for document subjects, we used subjects of two kinds as $a_j \in \{0,1\}$ $(j = 1,2)$, indicating that the profit and revenue increases compared with the prior term. $a_j = 1$ denotes that the value increases.

New York Times Annotated Corpus

For the experiment using NYTAC, we evaluated our model using different amounts of reference summary. The numbers of articles used for both validation and evaluation were 200. For training, we prepared 125, 250, and 500 articles. For training of document classification, we used 3000 articles for which the reference summary was not attached. As document subjects, we used the category of a news article $a_j \in \{0,1\}$ $(j = 1,2,\cdots,C)$ as a subject. Each subject corresponded to a news article category, such as "Business" and "Arts." $a_j = 1$ denotes that a document is classified into the category j. C is the number of categories, which is 26 in our experiment.

4.2 Implementation Details

The word embeddings were pre-trained using Skip-gram (Mikolov et al., 2013) on all $1,043,064$ articles in the Japanese version of Wikipedia. The dimensions of word embedding were 200. Those of the hidden layer in LSTM-RNN were 400. For CNN, the list of kernel sizes was $\{1,2,3,4,5,6\}$. The number of feature maps was 50. Adadelta (Zeiler, 2012) was used for updating parameters. The initial learning rate was 10^{-6}. The hyper parameters were optimized using grid search. We extracted three sentences with the highest scores in the manner described in an earlier report (Cheng and Lapata, 2016).

4.3 Baselines

For the NIKKEI financial report dataset, we used LEAD, which extracts the leading three sentences of a document as a baseline. We also built a baseline classifier LREG using logistic regression and human engineered features. The features were sentence length, position in the document, number of entities, nouns, verbs, adverbs, and adjectives in the sentence. We also added the sentiment of sentence to the features. For the financial report summarization, sentiment information is important because positive/negative sentences are frequently included in the summary when the revenue increases/decreases. The sentiment is computed by the frequency of words that appear in the articles when the revenue increases/decreases. For both datasets, we assigned NN-SE(Cheng and Lapata, 2016) as the baseline. The difference between NN-SE and our model is the introduction of multi-task learning and curriculum learning. The hyper-parameters are the same as those of our model. Through comparison with NN-SE, we can validate the effectiveness of the proposed framework.

5 Results

5.1 Results obtained from the NIKKEI Financial Report Corpus

Table1 presents the results for financial reports using F-measure. The precision and recall are calculated based on binary classification setup. LEAD, LREG, and NN-SE are used as the baselines. The proposed neural multi-task learning model, NN-ML, is significantly inferior to NN-SE and LREG. However, NN-ML-CL, the proposed model with curriculum learning, is superior to all

Table 1: F-measure evaluation (%) on financial reports

Models	F-measure	Precision	Recall
LEAD	42.1	39.1	50.4
LREG	60.5	**67.6**	66.5
NN-SE	59.9	58.0	68.8
NN-ML	55.2	52.1	64.6
NN-ML-CL	**60.6**	54.6	**74.9**

Table 2: F-measure evaluation (%) depending on the amount of reference summary

Models	125	250	500	1000	2000
NN-SE	55.2	56.1	58.0	58.0	58.1
NN-ML-CL	**58.1**	**58.4**	**59.4**	**59.3**	**59.2**

other models. This result shows that merely introducing multi-task learning does not positively influence on sentence extraction. However, curriculum learning overcomes the difficulty of multi-task learning; thus, document classification has positive effects on sentence extraction.

We confirmed the relation between the effectiveness of our model and the amount of reference summaries. We compared NN-ML-CL (our model) and NN-SE in several cases for which the amounts of reference summaries were 125, 250, 500, 1000, and 2000; the results are shown in Table 2. As observed, NN-ML-CL is superior to NN-SE in all cases. The margin between NN-ML-CL and NN-SE grows as the amount decreases, which means that document classification is more effective in cases with fewer reference summaries.

We also reported the results of human evaluation for summaries generated by the respective systems. Referring to the gold summary, participants ranked the generated summaries generated by four systems: NN-ML-CL(our system), NN-SE, LEAD, and LREG. The judging criteria was informativeness, which indicates how a generated summary covers information in the gold summary. From the test documents, we remove summaries for which the same sentences were extracted by different systems and randomly sampled 20 documents. 6 persons participated in the evaluation.

Table 3 presents the distribution of ranking and the average. Our NN-ML-CL model is ranked first in more than half the tests and markedly surpasses other models. Comparison with NN-SE verifies the effectiveness of multi-task learning for human evaluation.

Table 3: Ranking distributions (%) and the average evaluated by humans

Models	1st	2nd	3rd	4th	Ave.
LEAD	21.7	20.0	28.3	30.0	2.67
LREG	20.0	28.3	26.7	25.0	2.45
NN-SE	31.7	21.7	16.7	30.0	2.57
NN-ML-CL	51.7	20.0	21.7	6.7	**1.83**

Table 4: ROUGE scores (%) for various amounts of reference summaries in NYTAC

Model	Ref.	ROUGE-1	ROUGE-2
NN-SE	125	17.2	12.1
	250	16.8	11.3
	500	18.0	12.5
NN-ML-CL	125	18.1	12.7
	250	18.3	12.7
	500	18.5	12.9

5.2 Results on NYTAC

Table 4 presents our results for NYTAC using ROUGE-1 and ROUGE-2. In all cases, NN-ML-CL outperforms NN-SE on both metrics. When the amount of reference summary is 250, the margin between NN-ML-CL and NN-SE is the largest on each metric. For cases with 125 and 500 reference summaries, improvement is observed, but the margin is smaller than in the case for financial reports.

5.3 Discussion

In this section, we discuss how document classification contributes to the improvement of sentence extraction performance on the financial report dataset.

As mentioned above, NN-ML, the model that uses multi-task learning, exhibits a performance that is worse than that of NN-SE, the model that does not use multi-task learning. One possible explanation would be that it is difficult to optimize the parameters to maximize the likelihood of both sentence extraction and document classification simultaneously. If the learning task of sentence extraction does not proceed well enough, the task of document classification may also not work well. The reason is that the classification task relies on the estimated probability for sentence extraction in our proposed summarization framework.

However, curriculum learning improves the performance of the model with multi-task learning. By introducing curriculum learning into the framework, we are able to start training the model

Table 5: Rates (%) of extracted sentences corresponding with a document subject

	Correspond	Not	Others
NN-SE	**40.0**	40.0	20.0
NN-ML-CL	**85.7**	14.3	0.0

Table 6: Example of gold summary and sentences extracted using NN-ML-CL and NN-SE

Sentences extracted by NN-ML-CL
The rapid progress of the strong yen adversely influenced financial results, but the growth of revenue in areas such as Europe, Asia, and Oceania increased the revenue to 535 billion yen.
The demand for air conditioners increased because of the intense July heat.
Sentences extracted by NN-SE
As for fluorine resin, the demand for semiconductors rose steadily. However, competing Chinese companies gained power, and revenue for electrical wire use declined.
The revenue of parts for guided missiles increase year-on-year, but the revenue of medical equipment decrease.
News Article (Gold summary)
In Southeast Asia and Europe, high-end models of air conditioners sold well. In China and US, revenue rose and overcame the adverse influence of strong yen. The corporate tax ratio reduction also supported business performance. The revenue of air-conditioners, a leading product of Daikin, rose 9% in Southeast Asia. The revenue network in Vietnam and Indonesia expanded. revenue of air-conditioners rose at a higher pace than market scale. In China, the revenue of air conditioners for business use recovered. High-end models were also selling well.

only for sentence extraction. Then, the training for document classification is started gradually. Eventually, it contributes to the improvement of the performance of sentence extraction through the multi-task learning approach.

From the results for the financial report corpus, we confirmed that the contents of sentences extracted by our model corresponded with revenue and profit changes. Before validation, the sentences were categorized as corresponding to or not corresponding to others. We compared the results of sentence extraction with NN-ML-CL and NN-SE and checked the category distribution of sentences extracted using NN-ML-CL or NN-SE.

Table 5 shows that 85.7% of sentences extracted using NN-ML-CL correspond to changes of revenues and profits. However, only 40.0% of sentences extracted by NN-SE correspond to these parameters, which indicates that document classification supports extraction of sentences related to the revenue and profit change, and contributes to the improvement.

Table 6 shows sentences extracted from financial reports published by Daikin, Ltd., the leading air-conditioner manufacturer in Japan. During this term, the air conditioner revenue increased; moreover, revenues and profits increased considerably year-on-year. NN-ML-CL extracted sentences that mention the good revenue performance of air-conditioners in Asia and Europe, which is the same as that in the gold summary. In contrast, NN-SE extracts sentences mentioning the bad revenue performance of fluorine resin and medical equipment, which are not described in the gold summary. NN-SE is badly affected owing to training on past reports and articles. Our model extracts sentences with words that appear frequently in a positive context. Therefore, sentences related to good revenue performance are extracted.

There are two main ways of applications for our summarization approach with document classification. In the first case, the text collection has explicitly annotated document labels, which includes the collection of news articles with their category information, product reviews with their rating, scholarly paper abstracts with their discipline information, etc. In the second case, a document label can be acquired from external information sources about the text collection. For financial reports, the information about financial situation of a target company is extracted from the financial statement, which in turn can be used for a label of document classification.

6 Related Work

Based on the recent advances of neural network-based approaches (Kågebäck et al., 2014; Cao et al., 2015; Yin and Pei, 2015; Cao et al., 2016), an attentional encoder-decoder for extractive single-document summarization and its application to the news corpus was proposed (Cheng and Lapata, 2016; Nallapati et al., 2017). Although we employ an encoder-decoder architecture in the predictor component of our summarization framework, the framework can be applied to all models of sentence extraction using distributed representation as inputs, including recently advanced other attention-based encoder-decoder networks (Wang et al., 2016; Yang et al., 2016)

(Cheng and Lapata, 2016; Nallapati et al., 2017) argue that a stumbling block to applying neural network models to extractive summarization is the lack of training data and documents with

sentences labeled as summary-worthy. To overcome this, several studies have used artificial reference summaries (Sun et al., 2005; Svore et al., 2007; Woodsend and Lapata, 2010; Cheng and Lapata, 2016) compiled by collecting documents and corresponding highlights from other sources. However, preparing such a parallel corpus often requires domain-specific or expert knowledge depending on the domain (Filippova et al., 2009; Parveen et al., 2016). Our summarization uses document-associated information as pseudo rough reference summaries, which enables us to learn feature representations for both document classification and sentence identification with smaller amounts of actual reference summaries.

Neural networks based multi-task learning has recently proven effective in many NLP problems (Liu et al., 2015, 2016; Firat et al., 2016; Dong et al., 2015). Aiming at single document summarization with relatively small amounts of reference summaries, we demonstrated document summarization in the framework of multi-task learning with curriculum learning for sentence extraction and document classification. This enabled us to obtain better feature representations to extract sentences from documents.

7 Conclusion

In this paper, we proposed a general framework for extractive summarization using document subjects. Our key idea is to use a multi-task learning method that supports sentence extraction while enabling document classification, assuming that a document can be classified into certain subjects, and sentences for a summary are extracted in relation to the subjects.

This framework enables single document summarization with relatively small amounts of reference summaries since document subjects can be used as pseudo-rough reference summaries. Our proposed method can be widely applied for actual documents attached with meta-information such as product reviews, sports news and so on.

Experimental results showed that our model is less effective on the news corpus. For higher performance, more information such as the embeddings of news descriptors for document classification must be used.

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR1513, Japan and the New Energy and Industrial Technology Development Organization (NEDO).

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*, pages 41–48.

Ziqiang Cao, Wenjie Li, Sujian Li, Furu Wei, and Yanran Li. 2016. Attsum: Joint learning of focusing and summarization with neural attention. In *COLING*, pages 547–556.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*, pages 2153–2159.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *ACL*, pages 484–494.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL-HLT*, pages 93–98.

Volkan Cirik, Eduard Hovy, and Louis-Philippe Morency. 2016. Visualizing and understanding curriculum learning for long short-term memory networks. *arXiv preprint arXiv:1611.06204*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL*, pages 1723–1732.

Günes Erkan and Dragomir R Radev. 2004. Lexpagerank: Prestige in multi-document text summarization. In *EMNLP*, volume 4, pages 365–371.

Katja Filippova, Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2009. Company-oriented extractive summarization of financial news. In *EACL*, pages 246–254.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. Extractive summarization using continuous vector space models. In *the second Workshop on Continuous Vector Space Models and their Compositionality in EACL*, pages 31–39.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *AAAI*, pages 2741–2749.

Wei Li, Lei He, and Hai Zhuge. 2016. Abstractive news summarization based on event semantic link network. In *COLING*, pages 236–246.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out Workshop in ACL*, volume 8.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In *IJICAI*, pages 2873–2879.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *NAACL-HLT*, pages 912–921.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, pages 557–564.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop in ICLR*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, pages 3075–3081.

Karolina Owczarzak, John M Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Workshop on Evaluation Metrics and System Comparison for Automatic Summarization in ACL*, pages 1–9.

Daraksha Parveen, Mohsen Mesgar, and Michael Strube. 2016. Generating coherent summaries of scientific articles using coherence patterns. In *EMNLP*, pages 772–783.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium*, 6(12).

Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. 2005. Web-page summarization using clickthrough data. In *ACM SIGIR conference on Research and development in information retrieval*, pages 194–201.

Krysta Marie Svore, Lucy Vanderwende, and Christopher JC Burges. 2007. Enhancing single-document summarization by combining ranknet and third-party sources. In *EMNLP-CoNLL*, pages 448–457.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on abstract meaning representation. In *EMNLP*, pages 1054–1059.

Y Wang et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *EMNLP*, pages 606–615.

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *COLING*, pages 985–992.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *ACL*, pages 565–574.

Z Yang et al. 2016. hierarchical attention networks for document classification. In *NAACL-HLT*, pages 1480–1489.

Wenpeng Yin and Yulong Pei. 2015. Optimizing sentence modeling and selection for document summarization. In *IJCAI*, pages 1383–1389.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.