aiai at the FinSBD-2 task: Sentence, List, and Item Boundary Detection and Items Classification of Noisy Financial Texts Using Data Augmentation and Attention Model

Ke Tian, Hua Chen & Jie Yang
Rakuten Inc, Japan
Jiangxi Normal University, China
South China University of Technology University, China
tianke0711@gmail.com, hua.chen@kyudai.jp, 1661075801@qq.com

Abstract

This paper describes the method that we submitted to the FinSBD2-shared task in IJCAI-2020 to detect the sentence, list, and item boundaries and classify the items from noisy unstructured English and French financial texts. We used the spatial and semantic information of text to augment each tokenized word of text as a fixed-length sentence, and we labeled each word sentence as different boundary types. Then, we proposed the deep attention model based on word embedding to detect the sentence, list, and items boundaries in noisy English and French texts extracted from the financial documents and classified the item sentences into different item types. The experiment shows that the proposed method could be an effective solution to deal with the FinSBD2-shared task.

1 Introduction

The sentence is the fundamental unit in the written language. Thus, using sentence boundary detection (SBD), which detects the beginning and end of the sentence, is the first step of many language tasks, for example Part-of-speech (POS) tagging, discourse parsing, and machine translation. Until now, research about SBD has been confined to formal texts, such as news and European parliament proceedings, which have high accuracy using rule-based machine learning and deep learning methods due to the perfectly clean text data [Tian et al., 2019b]. However, there is almost no SBD research to address the problem in noisy texts, which are extracted automatically from machine-readable files, such as the financial PDF file format. One type of financial file is prospectus documents. Financial prospectuses are official PDF documents in which investment funds precisely describe their characteristics and investment information. The most critical step of extracting any information from these PDF files is to first analyze the information to get noisy unstructured text data, clean the text to format the information, and finally, transform it into semi-structured text so that sentence and list boundaries are well organized [FinNLP-2020, 2020]. Therefore, using SBD is an essential step to process the noisy financial text. The FinNLP workshop in IJCAI-2019 is the first proposal of FinSBD-2019 shared tasks that detect sentence boundaries in the noisy text of finance documents [Ait Azzi et al., 2019]. However, these financial prospectuses documents also contain many visual demarcations that indicate a hierarchy of sections, including bullets and numbered texts. There are many sentence fragments and titles—and not just complete sentences; some are lists or item texts. The prospectuses often contain punctuation errors. To organize the dense information into a more easily read format, lists are often utilized. Therefore, detection of the list, item boundary, and items classification is also crucial in the processing of the noisy text of finance documents. The task of FinSBD2 in the second FinNLP-2020 extended last year's task to include the detection of the list, items boundary detection, and items classification of the noisy text of finance documents.

There are English and French datasets in the FinSBD2 task. Our proposed method used the deep attention model and data augmentation method to approach the English and French tasks. According to the final leader board, the result is that our method could be a possible solution to deal with the English and French tasks, respectively.

Section 2 explains the details of the FinNLP-2020 task. Section 3 describes our method. Section 4 shows experimental configurations and discusses the results. Then, Section 5 concludes this paper.

2 Task Description

The FinSBD2 tasks included two sub-tasks. One is to detect the start and end char index of the sentence, list, and items part in noisy English and French text of finance documents. Another task is to classify the text of the item into four items: item1, item2, item3, and item4. We take one of the English task JSON files as an example to describe the task data.

J1: {“text”: “ \nCredit Suisse Fund I (Lux) \nInvestment Company with Variable Capital established \nunder….” "sentence": [{"start": 2, "end": 28,"coordinates": {"lines": [{ "x0": 216.42, "x1": 425.83,"y0": 441.44, "y1": 452.16,

"page": 1}],"start": {"text": "C","x0": 216.42,"x1": 228.75, "y0": 441.44, "y1": 452.16, "page": 1}, "end": {

"text": ")", "x0": 419.49, "x1": 425.83, "y0": 441.44, "y1": 452.16, "page": 1}}}, “list”: [..], “Item”:[..], “Item1”:[…], “Item2”:[], “Item3”:[…], Item4:[…].

As shown in the English JSON file J1, "text" is the text from the English financial PDF document. "Sentence" is the sentence text. Moreover, the start and end char index of sentence text is provided. In addition, the rectangle coordinate (xmin: the left upper x coordinate, ymin: the left upper y coordinate, xmax: the right bottom x coordinate, ymax: the right bottom y coordinate) data of each character of text is provided for text spatial visualization. The list and item texts label data are the same as sentence text information. As submitted prediction JSON data, the start and end char index of the sentence, list, and items should be predicted respectively, and items include four item types: item1, item2, item3, and item4. There are 29 labeled JSON (6 English and 23 French) in the training data. The test data is composed of 2 English and 10 French prediction JSON data.

3 Method

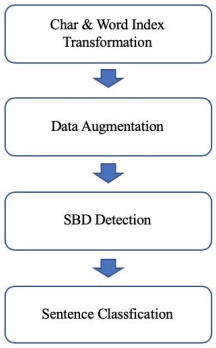

Figure 1: Procedure of proposed method

The whole processing procedure of the proposed method to deal with these two tasks is shown in Figure 1. Firstly, we transform the begin and end char index of the sentence, list, and items to the begin and end word token index in the word token text. Secondly, the spatial coordinate of the begin index of each word, the previous n word tokens, and the next n word tokens of each work token are used to augment each tokenized word to be a fixed sentence. After the fixed sentences are created, the fixed sentence could be labeled as four classes: begin word, end word, independent word, and other words. Then, the deep attention model is used to detect each fixed sentence. Thirdly, after the boundary label of each tokenized word in the text is detected, the begin and end char indexes of the sentence are detected. Then, the deep attention model is used to classify the predicted sentences into six types: sentence, list, item1, item2, item3, item4.

The details about char and word index transformation and data augmentation are described in Section 2.1. The sentence boundary detection and sentence classification are described in Section 2.2, and the ensemble result is presented in Section 2.3.

3.1 Data Augmentation

In the training JSON data, the begin and end char index of the sentence, list, and items are provided. We found that just char index information is not enough to predict the sentence boundary since the number of chars in training text is limited. In order to get more information for prediction boundary detection, it is better to obtain the begin and end word index of sentences. However, the char and word index transformation are not provided by the task. We have tried the spaCy library to do the word tokenization of English and French text. The spaCy library [spaCy, 2020] provides a function to obtain the begin char index of each word token in the text.

We observed that the end part of a sentence does not just use punctuation like '.' and ';' It includes some words like ')' and ':', which caused the ending part to be complicated. Like the ending part, the beginning part of the sentence also is not just words beginning with upper letters like 'The' and 'This.' It also includes symbol characters like '(' and '1.' Therefore, using only the rule to detect the beginning and ending of a sentence may not be effective. Besides, there are some words that also sentence that the beginning word and end word is the word itself.

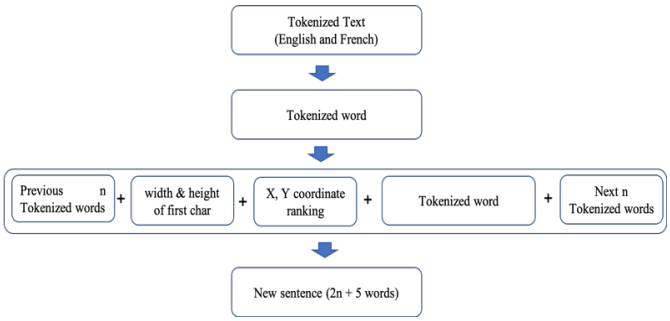

Figure 2: Procedure of data augmentation for boundary detection

We found that unusual beginnings and endings are identifiable by context. The surrounding words around beginning and ending words could help to identify the sentence boundary. Moreover, the size of the begin char is different from other chars. Most of the begin chars are upper letters, so the widths and heights are longer than that of lower letters. Moreover, the position of the beginning index is often located in the left part of the whole page. Based on these observations, the spatial coordinate, width, and height ranking of the begin char index of each word, the previous n word tokens, and the next n word tokens of each work token are used to augment each work token to be a fixed tokenized sentence. The procedure of data augmentation for sentence

boundary detection training and test data is shown in Figure 2.

We take the J1 sentence as an example to describe how to augment each word to be a fixed sentence. After using the spaCy library to token the text, we found that the J1 token text is as follows: [“\n”, 'Credit', Suisse', 'Fund', 'I', '(', Lux', ')', '\n', 'Investment', 'Company',……]

As each tokenized word, the previous n tokenized words following n tokenized words of each tokenized word, x and y coordinate ranking, width and height ranking of first char are taken to be concatenated into a new sentence. For example, take the "n is 5" as an example. As the first word is "/n" in the J1, there were no previous 5 words, so we added 5 "pre" words at the beginning of the sentence, and X, Y coordinate ranking, width, and height ranking are null. Therefore, the new sentence for "/n " is the T1 sentence. With the beginning word " Credit ," the new sentence is T2. As the end word “)”, the new sentence is T3. At the end of the J1 text, there were no next 5 words, so we added 5 "EOS" words at the end of the sentence. The labels of T1, T2, and T3 are "OS", "BS", and "ES," respectively, which are the same as the labels of the tokenized words "'\n,'" "Credit," and "'),'" respectively. Besides, there are some sentences in which the begin and end char is located in the same word. So the label of such a kind of word is labeled as “IS.” The train and test data in English and French use the same method to augment words for a fixed sentence. There are four labels (OS, BS, ES, IS) for the train and test data. Therefore, the goal of sentence boundary detection is converted to classify the labels of the new augmented sentence.

T1: ['pre', 'pre', 'pre', 'pre', 'pre', 'coor_x_null, 'coor_y_null', ' width_null',' height_null', ' \n', 'Credit', 'Suisse', 'Fund', 'I', '(]

T2: ['pre', 'pre', 'pre', 'pre', ' \n', 'coor_x_164', 'coor_y_364', ' width_11', 'height_6', 'Credit', 'Suisse', 'Fund', 'I', '(', 'Lux']

T3: ['Suisse', 'Fund', 'I', '(', 'Lux', 'coor_x_367', 'coor_y_364' , 'width_5', 'height_6', ')', '\n', 'Investment', 'Company', 'with', 'Variable']

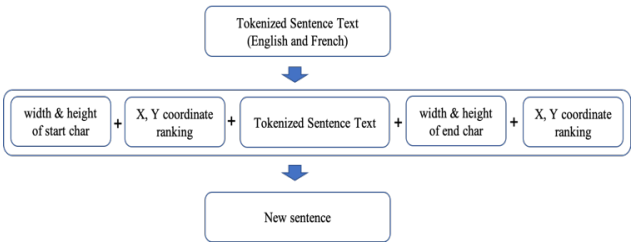

Figure 3: Procedure of data augmentation for sentence classification

Regarding the sentence classification data, the start and end char index of sentences, lists, and items are provided. Based on the char and word transformation, the tokenized word sentence text of the sentences, lists, and items could be inferred. The spatial information and width and height rankings of start char and end char are added to augment the training and test sentence texts. The procedure of data augmentation for sentence classification training and test data is shown in Figure 3.

We take one of the J1 sentence texts as an example to describe how to augment the J1 sentence to be a new sentence. As T4 sentence text. X & Y coordinate ranking, width & height ranking of first char and end char are taken to be concatenated into a new sentence named T5. The train and test data in English and French use the same method to augment the text of the sentences, lists, and items.

T4: ['Suisse', 'Fund', 'I', '(', 'Lux', ')]

T5: ['coor_x_164','coor_y_364','width_11','height_6','Credit' ,'Suisse','Fund','I','(','Lux',')','coor_x_374','coor_y_364','widt h_4','height_6']

Now the goal of the FinSBD2 task has changed to text classification. Word embedding is the foundation of deep learning for natural language processing. We use the new train, test text data to train the word embedding as the first SBD classification. In the recreated English text data, there are 329,908 recreated sentences with 9,894 unique token words from the training and test data. In the French text, there are 698,612 recreated sentences with 12,374 unique token words from the training and test data. Regarding sentence classification, there are 11,931 and 20,858 sentence texts of the sentences, lists, and items in English and French train and test data, respectively. The CBOW model [Tomas Mikolov, 2013] is taken to train word vectors for the English and French text data, and the word2vec dimension is set to 100.

3.2 Sentence Boundary Detection and Sentence Classification

To complete the task goal and get the submission data, we first classify the tokenized words into four classes: BS, ES, OS, and IS with augmented sentence classification. The start and end char index are based on the word index using spaCy. After we obtain the word label of text, the sentence text can be extracted based on boundary labels. Secondly, the extracted sentence texts are predicted to be sentence, list, and item labels. Finally, the submission data with sentence label and char index is complete.

Through the task train data, we observe that some keywords can help determine the category of a sentence. For example, "." and ")" indicate the ending part of the sentence. Thus, some keywords in the sentence have more importance in predicting the label of sentence text. Since the attention mechanism can enable the neural model to focus on the relevant part of your input, such as the words of the input text, the attention mechanism is used to solve the task [Ke Tian et al., 2019a]. In this paper, we mainly use the feed-forward

attention mechanism [Colin Raffel et al., 2015]. We put in the attention layer in the long short-term memory (LSTM) [Sepp Hochreiter et al., 1997] model, which has been provided adequately in the sentence boundary detection [Ke Tian et al., 2019b], as shown in Figure 4. In this paper, the attention-based LSTM is utilized for sentence boundary detection and sentence classification.

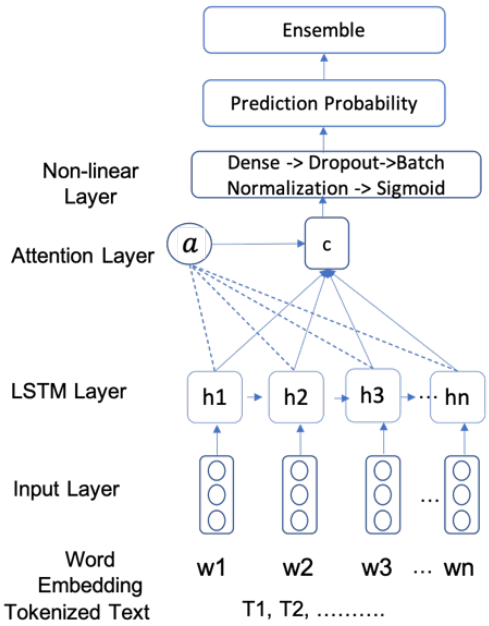

Figure 4: Attention-based LSTM model

Regarding the structure of the proposed model, as the LSTM layer, the embedding dimension and max word length of word embedding are set at 100 and 2n+1 (n is the number of words surrounding the tokenized word), respectively, for English and French boundary detection. As the sentence classification, embedding dimension is 100 for both English and French tasks. For max word length, the French text is 400, and the English text is 700. The embedding layer of the word embedding matrix as an input layer of LSTM, and the size of the output dimension is 300. We used the feed-forward attention mechanism as the attention layer. As the non-linear layer, the activation function is to dense the output of the attention layer to be 256 dimensions, and by using the dropout rate of 0.25, the output result after the dropout rate will be batch normalization. Finally, the sigmoid activation function is to dense the dimension of batch normalization input to be the length of the label as the final output layer for boundary detection and sentence classification.

3.3 Ensemble Result

As the model training stage, the 5-fold cross-validation is used to train the deep attention model for predicting the test data for boundary detection and sentence classification. We sum 5 folds of predicted probability and get the mean value of 5 folds for the final predicted probability result.

4 Experiment

4.1 Experiment Design and Implementation

In the experiment stage, the spaCy-based method and the proposed method were implemented to complete the boundary detection and sentence classification goals. Moreover, in the data processing stage, we have kept the upper letter of words to train the word embedding in the English and French text. In addition, we tested the different numbers of words surrounding each tokenized word. The numbers 10 are used to complete these tasks. The deep attention model in our paper was implemented using Keras deep learning libarary [Keras, 2019].

Based on the evaluation requirements of the FinSBD2 Task, the F-scores were taken to evaluate the performance of predicted sentence, list, and item boundaries, which are pairs of character indexes ("start" and "end"), using the proposed model in the paper.

4.2 Experiment Result and Discussion

Based on the result, the results of the deep attention model are shown in Table 1. As seen in Table 1, as the English task, the score of the spaCy based is worse than the proposed method score. Moreover, the final score of the aiai team in the final leader board is shown in Table 2.

Method	Lang	Subtask 1	Subtask 2
spaCy	French	0.199	0
	English	0.208	0
Proposed method	French	0.471	0.35
	English	0.413	0.203

Table 1: Experiment result

	English		French		
Team name	Subtask1	Subtask2	Subtask1	Subtask2	mean
PublishCovid19	0.937	0.844	0	0	0.445
aiai	0.413	0.203	0.471	0.35	0.359
Daniel	0.317	0	0.262	0	0.145
Subtl.ai	0.217	0	0	0	0.054
Anuj	0.126	0	0.025	0	0.038

Table 2: Final leader board ranking

The above result shows that the proposed method could be a possible solution to predict the beginning and end char index of the sentences, lists, and items in English and French text.

However, the score of the proposed method still needs to be improved compared with the ideal score. We surmised that the following reasons may cause the score to be low. Firstly,

the preprocessing of text is not done enough for boundary detection and sentence classification. Secondly, the parameter tuning (such as n words tuning and other parameters) is not often done due to busy schedules. Thirdly, we have only used the deep attention model—other models such as the Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018] and the Named Entity Recognition (NER) methods have not been tried in the current tasks. Fourthly, the limited amount of task training data may influence the performance of the proposed method. Finally, in the English word and char transformation, there are some words that do not match the relative char index.

5 Conclusions and Future Work

This paper primarily informs the aiai team how to tackle the FinSBD2-2020 shared tasks. There are two tasks—one is to predict the start and end char index of the sentences, lists, and items part in noisy English and French text of finance documents. Another one is to classify the items' text into four items: item1, item2, item3, and item4. We approach these two tasks as text classification problems using data augmentation and the deep attention model. The experimented result showed that the proposed model might effectively solve the goal of the task.

However, our method still needs to be improved to achieve better performance in the following directions. Firstly, it is better to do more parameter tuning in the current model to improve the accuracy of boundary detection and sentence classification. Moreover, we will explore different methods and models, such as the BERT model, to improve the boundary detection accuracy.

Acknowledgments

This work is financially supported by the Scientific Research Foundation of Jiangxi Normal University for the PhD (No. 0901/12019572).

References

[Ait Azzi et al., 2019] Abderrahim Ait Azzi, Houda Bouamor, and Sira Ferradans. The finsbd-2019 shared task: Sentence boundary detection in pdf noisy text in the financial domain. *In The First Workshop on Financial Technology and Natural Language Processing (FinNLP 2019)*, Macao, China, 2019.

[Colin Raffel et al., 2015] Colin Raffel and Daniel P. W. Ellis. Feed-forward networks with attention can solve some long term memory problems. https://arxiv.org/abs/1512.08756, 2015.

[FinNLP-2020 2020] FinNLP-2020. https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp2020/shared-task-finsbd-2. Accessed: May 2020.

[Ke Tian a et al., 2019a] Ke Tian and Zi Jun Peng. aiai at FinNum task: Financial numeral tweets fine-grained classification using deep word and character embedding-based attention model. The 14th NTCIR Conference, Tokyo, Japan, June 2019.

[Ke Tian et al., 2019b] Ke Tian and Zi Jun Peng. aiai at FinSBD task: Sentence Boundary Detection in Noisy Texts From Financial Documents Using Deep Attention Model. aiai at FinSBD task: Sentence Boundary Detection

[Keras, 2019] Keras. The python deep learning library. https://keras.io. Accessed: May 2019

[Sepp Hochreiter et al., 1997] Sepp Hochreiter and Jürgen Schmid- huber. Long short-term memory. Neural Computation, 9(8):1735-1780, 1997.

[spaCy 2020] spaCy. https://spacy.io. Accessed: May 2020

[Tomas Mikolov, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean. Distributed represen- tations of words and phrases and their compositionality. https://arxiv.org/abs/1310.4546, 2013.