AI_Blues at FinSBD Shared Task: CRF-based Sentence Boundary Detection in PDF Noisy Text in the Financial Domain

Ditty Mathew , Chinnappa Guggilla

Applied Intelligence Labs - Accenture Operations
Accenture Solutions Pvt Ltd, Bangalore, India
{ditty.mathew,chinnappa.guggilla}@accenture.com

Abstract

This paper reports the team AI_Blues's participation in the FinSBD 2019 shared Task on 'Sentence Boundary Detection in PDF Noisy Text in the Financial Domain'. Sentence detection from noisy text is a challenging task. We modeled the sentence boundary detection problem as a sequence labeling problem using Conditional Random Field (CRF) approach for English and French language financial texts. We proposed to use punctuation embeddings as an additional feature along with the basic language specific features and obtained 84.5%(F1) and 86.5%(F1) accuracies in the English and French language shared task datasets respectively.

1 Introduction

The task of Sentence Boundary Detection (SBD) is to identify the sentence segments within a text. In Natural Language Processing (NLP), the *sentence* is the foundational unit and extracting sentences or detecting the boundary of sentences from a noisy text is a challenging task. Any imperfect sentence boundary detection system can affect the morphologic, syntactic, semantic and discourse analysis in text processing. The punctuations such as '.', '?' and '!' are commonly used as sentence boundaries. However, the usage of punctuation '.' is ambiguous [Grefenstette and Tapanainen, 1994]. It can be used along with decimals, email addresses, abbreviations, initials in names, etc.

Despite the important role of sentence boundary detection in NLP, this area has not received enough attention so far. The existing approaches for this task are confined to formal texts and to the best of our knowledge no studies have been conducted in noisy texts for this task. In FinSBD shared task, the focus is to detect the beginning and ending boundaries for extracting well segmented sentences from financial texts. These financial texts are PDF documents in which investment funds precisely describe their characteristics and investment modalities. The noisy unstructured text from these PDF files was parsed by the shared task organizers and the task is to transform them into semi-structured text by tagging the sentence boundaries in two languages - English and French. For example: consider the English sentence "*Subscriptions may only be received on the basis of this Prospectus.*". Here the word *Subscriptions* is tagged as the beginning and the period[1] '.' is tagged as the ending of the sentence in the given corpus. We have modeled the sentence boundary detection problem as a sequence labeling problem. The tokenized text is the input and the output is the corresponding labels. The labels assigned to the tokens are 'BS', 'ES' and 'O' to mark the beginning of the boundary, ending of the boundary and non-boundary token respectively.

We propose a Conditional Random Field (CRF) [Lafferty *et al.*, 2001] model to predict the label sequence of the input text. The rules for detecting sentence boundaries can be captured as features of CRF and learns the conditional probability of the label sequence given the observation sequence of features. We report the related work in Section 2 and briefly discussed the idea of conditional random field in Section 3. In Section 4, we explain the proposed part of speech and punctuation embeddings-based clustering features for the CRF model in this task. Section 5 presents the data sets, experiments, evaluation and its results. Section 6 summarizes the error analysis and discussion which is followed by conclusion and future work in Section 7.

2 Related Work

In the literature, the approaches attempted for sentence boundary detection task fall into three categories - rule-based approach, supervised machine learning approach and unsupervised approach. The rule-based SBD uses hand-crafted rules and heuristics. Mikheev [2002] proposed a rule-based approach which disambiguates the occurrence of period/full stop by determining whether it decides the sentence boundary or not. This method identifies the abbreviations by looking at local contexts and the repetitions of individual words in the document. It then applies this information to detect sentence boundary by applying a small set of rules.

Recent research in sentence boundary detection focuses on machine learning techniques. Riley [1989] presented a decision tree classifiers in determining whether the instances of full stops mark sentence boundaries. This approach uses features such as probabilities of words being sentence final or initial, word length, and word case. Satz is an approach proposed by Palmer and Hearst [1997] which uses decision tree

[1] we use the term period or full stop interchangeably to refer the punctuation '.'

Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2019), pages 130-136, Macao, China, August 12, 2019.

or a neural network to disambiguate the role of punctuation mark in a sentence by using the prior distributions of word class surrounding the possible end-of-sentence punctuation mark as features. A maximum entropy learning is proposed by Reynar and Ratnaparkhi [1997] which disambiguates the potential sentence boundary tokens such as '.', '?', '!'. This model learns the contextual features of such ambiguous punctuations by considering the token preceding and following a sentence boundary.

Kiss and Strunk [2006] proposed an unsupervised sentence boundary detection system called Punkt. This method detects abbreviations, initials and ordinal numbers by using collocation information as evidence derived from unannotated corpora. A large development corpus of the Wall Street Journal is used to derive the collocation information. This method is proposed for sentence boundary detection in multilingual sentences.

Closest to our proposed approach is the work on token and sentence splitters using conditional random field in biomedical corpus [Tomanek *et al.*, 2007]. This model captures features such as - i) token size, ii) sentence boundary tokens such as full stop, question mark, and exclamation mark, iii) canonical form of word based on the usage of capital letter, small letter, digit and other characters, iv) orthographical features such as HasDash, AllCaps, InitalCap, and hasParenthesis, and v) abbreviations. In our proposed model, we study the contextual behaviour of punctuations and formation of rules based on the combined usage of sentence boundary tokens. Evang et al. [2013] proposed a sentence segmentation method using CRF which considers characters as basic units for labeling. However, in FinSBD shared task, tokens are the basic units for labeling.

With respect to the sentence boundary detection of the French language, Maegaard, and Spang-Hanssen [1973] described a method to segment the French sentences into principal clauses and subordinate clauses by using only a few kinds of linguistic signs in the text. Gonzalez et al. [2018] proposed a convolutional neural network [Kalchbrenner *et al.*, 2014] based approach for detecting sentence boundaries of French speech texts which tackle the task as a binary classification task.

3 Conditional Random Field

Conditional Random Field (CRF) [Lafferty *et al.*, 2001] is a probabilistic method for structured prediction and it computes the conditional probability of the label sequence given the observation sequence. The conditional probability of the label sequence $Y = y_1, y_2 \ldots, y_T$ given the observation sequence $X = x_1, x_2, \ldots, x_T$ is given as

$$P(Y/X) = \frac{1}{Z(X)} \exp \sum_{t=1}^{T} \sum_{k=1}^{F} \lambda_k f_k(x_t, y_t) \qquad (1)$$

where $f_k(x_t, y_y)$ is a feature function and its value may range from -∞ to +∞, but typically they are binary. Each feature function f_k is associated with a weight λ_k which is learned during training. $Z(X)$ is the normalization factor to make the probabilities sum up to 1 and it is defined as

$$Z(X) = \sum_{Y} \exp \sum_{t=1}^{T} \sum_{k=1}^{F} \lambda_k f_k(x_t, y_t) \qquad (2)$$

The conditional distribution discussed in Equation 1 is a linear chain CRF which includes the features only for current word. We used richer features of the input x_i such as prefixes (x_{i-1}, x_{i-2}), suffixes (x_{i+1}, x_{i+2}) and surrounding words of current word and their corresponding label sequences.

We have modeled the sentence boundary detection problem as a sequence tagging problem. When applying CRF to SBD problem, a sequence of tokens in a text is considered as the observation sequence and the label sequence is the corresponding sequence of labels. Each token is labeled with respect to its position in the sentence. If the token is positioned at the beginning of a sentence, its label is 'BS'. If the token is positioned at the end of the sentence, then the label is 'ES'. The remaining tokens are labeled as 'O'. In this way we used the English and French training corpus tagged with 3-tags for building the sentence boundary detection model using CRF.

4 Sequence Representation and Features for CRF

4.1 Preprocessing - Tokenization

The input text to the CRF model is tokenized in a way that the punctuations are considered as tokens. For example, consider the following text and its tokens from English data.
Text: GAM Star (Lux) Prospectus
Tokens: ['GAM', 'Star', '(', 'Lux', ')', 'Prospectus']

4.2 Sequence Representation

We consider an average of five sentences as a single unit for sequence representation in CRF modeling. The optimal size for the sequence representation is determined using the best performance of the CRF model on the development set during the training phase. In the case of test set, we have used the entire document as a sequence for CRF prediction. We have applied the same schema for both the English and French language SBD tasks.

4.3 Basic Features

In CRF, each token is represented by a set of features. In addition, the features of *k* preceding and *k* following tokens as n-grams are included for each token. The feature selection plays a crucial role in CRF. We propose the following basic set of surface and orthographic features for sentence boundary detection task. We also used part of speech (POS) syntactic feature in addition to other features and denoted as 'basic features' in the rest of the sections.

- Token: The token itself is considered as a feature. This feature captures the co-occurrence properties of tokens when we consider the preceding and following tokens.
- Length of token: The length of the token is considered as a feature.
- IsUpper: This is a binary feature which is set to 1 if all characters of the token are in upper case otherwise it is set to 0.

- IsLower: This is a binary feature which is set to 1 if all characters of the token are in lower case else 0. This feature is based on the assumption that a sentence may not start with a lower case character word.
- IsTitle: This is a binary feature which is set to 1 if the first character of the token is in upper case and the remaining characters are in lower case.
- PosTag: The part of speech tag of the token is used as a feature. This feature captures the role of parts of speech such as verb, noun, prepositions, etc in determining the sentence boundaries.
- Token Name: This feature assigns a name to the token based on its nature such as whether it is a word, punctuation, digit, etc. The assignments of token names for the corresponding token types are given in Table 1. This feature captures the characteristics of tokens.

Token type	Token Name
Word	NN
Punctuation	<Name of the punctuation>
Digit	NUMBER
Roman Numeral	ROMAN
Alphabet	ALPHABET
If token has number and character	ALPHANUMERIC

Table 1: Token names based on token type

- Lexical combinations: We consider features which check for the combined usage of tokens in the sentence boundaries. These features are listed in Table 2. These features obtain the contextual features of the potential sentence boundary tokens such as ‘.’, ‘?’, ‘!’ by considering the token preceding and following a sentence boundary. In table 2, we list the features only for the token ‘.’ as the financial corpus given in this shared task does not use ‘?’ and ‘!’ as sentence boundary. These features are binary features which are set to 1 if the pattern occurs for the token t_i.

$t_i =$ ‘.’, $t_{i+1} =$ A word begin with upper case
$t_i =$ ‘.’, $t_{i+1} =$ A word begin with upper case, $t_{i+2} =$ A word begin with upper case
$t_i =$ ‘.’, $t_{i+1} =$ ‘(’, $t_{i+2} =$ digit/alphabet/roman numeral $t_{i+3} =$ ‘)’ $t_{i+4} =$ A word begin with upper case
$t_i =$ ‘.’, $t_{i+1} =$ digit/alphabet/roman numeral $t_{i+2} =$ ‘)’ or ‘.’ $t_{i+3} =$ A word begin with upper case
t_{i-1} is a word begins with upper case $t_i =$ ‘.’, $t_{i+1} =$ ‘(’, $t_{i+2} =$ digit/alphabet/roman numeral $t_{i+3} =$ ‘)’ $t_{i+4} =$ A word begin with upper case

Table 2: Lexical combinations as features for CRF

4.4 Punctuation Embeddings as Clustering Feature

The word embeddings [Mikolov *et al.*, 2013a] have been proved effective in capturing contextual features and linguistic regularities [Mikolov *et al.*, 2013b]. After analyzing the corpus, we observed that the punctuation collocations in the sentences would contribute to identify the sentence boundaries. For example, in a sentence if ‘-’, ‘,’ and ‘.’ occur in combination with the other content words, we could say that all these punctuations together can help in identifying the right boundary of the sentence. Since punctuations are important in deciding the sentence boundaries, we use the information in punctuation embeddings as a feature. As the embedded vector is a high dimensional vector, we cannot directly use word embeddings as a CRF feature. Hence, we represent the punctuations using embedded vectors and cluster the embedded vectors of punctuations using *k*-means clustering algorithm [Hartigan and Wong, 1979]. The punctuations in each cluster are assigned a distinct value based on its cluster assignment and this value is used as the feature. The tokens which are not punctuations are grouped into a different cluster. We use pre-trained glove embeddings [Pennington *et al.*, 2014] and fastText embeddings [Grave *et al.*, 2018] for English and French texts respectively.

5 Experiments and Results

In this section, we describe the data sets, features used in the CRF method and evaluation of results.

5.1 Datasets

In FinSBD shared task, data sets are provided for English and French language [Ait Azzi *et al.*, 2019]. As the corpus is related to finance domain, the text contains various data elements such as formatting indicators, titles, subtitles, sections. Each of these elements, in turn, contains various types of vocabulary including special symbols, numerals, currencies, named entities. The data sets are in JSON format which contains i) the text to detect sentence boundaries and this text is already tokenized using NLTK, ii) begin_sentence which contains the indexes of tokens in the text that mark the beginning of well-formed sentences in the text, iii) end_sentence which contains the indexes of tokens in the text that mark the end of well-formed sentences in the text. The dataset statistics for English and French languages are given in Table 3. The average sentence lengths of English text in train, development, and test data sets are 30.82, 31.27 and 35.32 respectively; and that of French text are 26.71, 28.54 and 26.49 respectively. In particular to the FinSBD shared task dataset, we observed that i) the headings, bullet and numbering points etc. are considered as sentences, ii) some non-boundary tokens in sentences

Language	Dataset	No of tokens	No of sentences
English	Train	904,057	22,342
	Dev	49,859	1,384
	Test	56,952,	1,265
French	Train	827,852	22,636
	Dev	119,008	3,141
	Test	106,577	2,981

Table 3: Dataset Statistics

begin with upper case letter, iii) some tokens are fully given in upper case, iv) the punctuation ‘-’ is used as the beginning token of bullet points more frequently.

5.2 Experiment Setup

We extract the basic features and punctuation cluster features of tokens as discussed in Section 4. The parts of speech tag is tagged using python NLTK[2] package for English and stanford pos tagger[3] is used for French. To obtain the punctuation embeddings for English text, we use pre-trained glove embeddings [Pennington *et al.*, 2014] of 100 dimension and clustered them using *k*-means clustering. We experimented with different values of *k* ranges from 2 to 10 over the development data. We observed that the punctuations are clustered based on its contextual behaviour and the clusters resulted when *k*=7 are given in Table 4. For French, we use pretrained fastText embeddings [Grave *et al.*, 2018] of 300 dimension which is trained on Wikpedia data to obtain the punctuation clusters. The clusters are listed in Table 5.

Cluster No	Puncuations
1	, - . ” : & ’ ; $
2	( )
3	[ ]
4	? !
5	< >
6	+ * =
7	# @ — / %

Table 4: Punctuation Clusters obtained from English text

Cluster No	Puncuations
1	” ’
2	( ) [ ] . , ; :
3	? !
4	< >
5	+ * = - /
6	# @ — &
7	$ %

Table 5: Punctuation Clusters obtained from French text

Our CRF model is compared with the baselines such as punkt sentence tokenizer [Kiss and Strunk, 2006] and sentence boundary detector proposed by Tomanek et al [2007]. The punkt tokenizer divides a text into a list of sentences by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences. We use the punkt tokenizer model implemented in NLTK for English and French languages to report the results. The sentence boundary detector by Tomanek et al. [2007] is a conditional random field model with the following set of features.

- The token and its length.
- A binary feature which checks whether the token is a sentence boundary symbols such as full stop, question mark, and exclamation mark.
- Canonical word form which is constructed by applying the transformation rules such as i) replace capital letters by ‘A’, ii) replace lower case letters by ‘a’, iii) replace digits by ‘0’ and, iv) replace all other characters by ‘-’.
- Features such as HasDash, AllCaps, InitalCap, hasParenthesis.
- A binary feature which is set to 1 if the token is contained in a list of abbreviations.
- Local context features of neighboring tokens in the window [-1,1]

5.3 Experiments

We use python CRFsuite [Korobov and Peng, 2014] to model a linear chain CRF and it is trained using the gradient descent algorithm. The parameters such as “all_possible_transitions” and “all_possible_states” are set as True. We tuned the regularization parameters using separate development set for each model. The context window *k* to fetch the features of surrounding prefix and suffix tokens is selected based on the F1-score over the development data. The value *k* is chosen for English and French is 6. We use a paragraph as a sequence for training and the entire text in the test data as a single sequence for testing. As the prediction of boundary tokens require the characteristics of preceding and following tokens, the paragraph is considered over sentence as a sequence in the training phase. The text in the test data is input as a sequence as the paragraph information of test data is not available. During training, the sentences are constructed using the beginning and ending information given in the train data and a paragraph is considered as five consecutive sentences. The number of sentences in the paragraph is chosen based on the F1-score over the development data.

Language	Evaluation Measure	Label		Average
		BS	ES	
English	Precision	90	93	91.5
	Recall	85	90	87.5
	F1-Score	87	92	89.5
French	Precision	88	89	88.5
	Recall	86	90	88
	F1-Score	87	89	88

Table 6: Results obtained for the development data for the submitted model

Language	Evaluation Measure	Label		Average
		BS	ES	
English	Precision	77	83	80
	Recall	88	92	90
	F1-Score	82	87	84.5
French	Precision	89	90	89.5
	Recall	80	85	82.5
	F1-Score	85	88	86.5

Table 7: Results submitted to FinSBD shared task for the test data

5.4 Evaluation and Results

F1-score is used as the evaluation measure in the FinSBD shared task and we also report the precision and recall for evaluation. The precision, recall and F1-score are averaged for ‘BS’ and ‘ES’ labels and this average score is used for reporting the best model. The precision, recall, and F1-score obtained for English and French text over the development data are given in Table 6. The average F1-score obtained for English text is 89.5% and that of French text is 88% over the development data. The predicted results on the given test set are reported in Table 7. The submitted results are predicted using the CRF model which is trained using all the features discussed in Section 4. However, we later figured out that we used only a subset of all the features in the feature prepossessing step of the prediction module that we used for generating

[2] https://www.nltk.org/

[3] https://nlp.stanford.edu/software/tagger.html

Method	Evaluation Measure	Label BS	ES	Average
Punkt [Kiss and Strunk, 2006]	Precision	58	73	65.5
	Recall	71	89	80
	F1-score	64	81	72.5
CRF [Tomanek *et al.*, 2007]	Precision	77	81	79
	Recall	83	89	86
	F1-score	80	85	82.5
Basic Features	Precision	82	84	**83**
	Recall	85	94	89.5
	F1-Score	83	89	86
Basic + Punctuation Cluster	Precision	82	84	**83**
	Recall	86	94	**90**
	F1-Score	84	89	**86.5**

Table 8: Evaluation of English gold standard test set with all the features (corrected results post the shared task submission)

Method	Evaluation Measure	Label BS	ES	Average
Punkt [Kiss and Strunk, 2006]	Precision	54	39	46.5
	Recall	61	82	71.5
	F1-score	57	53	55
CRF [Tomanek *et al.*, 2007]	Precision	77	81	79
	Recall	82	87	82.5
	F1-score	80	84	82
Basic Features	Precision	89	90	89.5
	Recall	87	90	88.5
	F1-Score	88	90	89
Basic + Punctuation Cluster	Precision	90	92	**91**
	Recall	88	90	**89**
	F1-Score	89	91	**90**

Table 9: Evaluation of French gold standard test set with all the features (corrected results post the shared task submission)

results on given test data. We fixed this mistake later and reported the corrected results in Tables 8 and 9. The highest accuracy values for precision, recall, and averaged F1 measures are specified in bold font as shown in Tables 8 and 9.

In Table 8, we report the evaluation scores of baselines such as punkt [Kiss and Strunk, 2006] and CRF model [Tomanek *et al.*, 2007], and our proposed CRF models using basic features and basic + punctuation-based cluster features (See Section 4) for the English gold standard test set given in FinSBD shared task. We can observe that the CRF model using basic and punctuation-based cluster features are performing better than all other methods and, this model scores the highest F1-score for both 'BS' and 'ES' labels. The CRF model which uses basic features performs better than other baselines and scores the highest F1-score for 'BS'. While the Punkt unsupervised model scores an F1-score of 81% for 'ES' label, the sentence beginning performing very poorly. All CRF based models report greater than 80% F1-score for the 'BS' labels and it indicates that the sequential labeling of tokens gains more information on sentence beginning. We performed a paired t-test to check the statistical significance of the improvements of proposed CRF models over the baselines [Tomanek *et al.*, 2007; Kiss and Strunk, 2006] and observed that the improvements are statistically significant with p-value less than 0.05.

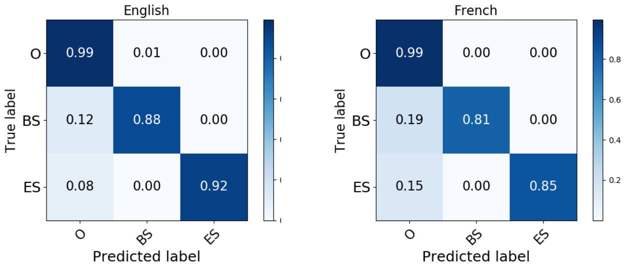

Figure 1: Confusion matrices obtained from the submitted results in the FinSBD shared task

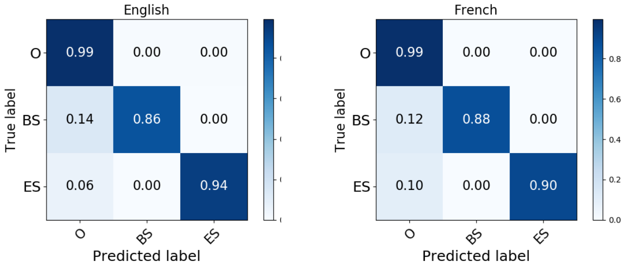

Figure 2: Confusion matrices obtained from post-submission results

The evaluation of French gold standard test data is reported in Table 9. The proposed CRF models are compared with the baselines and the CRF model which uses basic features. The CRF model which uses basic and punctuation cluster features is performing better than all other models in the gold standard test data. This model identifies beginning and ending of the sentence with F1-scores 89% and 91% respectively. The CRF model which uses basic features also performs much better than baselines. The CRF model by Tomanek [2007] identifies the 'BS' and 'ES' labels with F1-scores 80% and 84% respectively. The punkt model performs very poorly for French text in identifying both beginning and ending of the sentences. The improvements of the proposed models over the baselines [Tomanek *et al.*, 2007; Kiss and Strunk, 2006] are statistically significant with p-value less than 0.05 in paired t-test.

6 Error Analysis and Discussion

We have computed the confusion matrices on the shared task test results and for the post shared task submission results for English and French SBD tasks for the 3 tags - 'BS', 'ES' and 'O' tags. These confusion matrices are illustrated in Figures 1 and 2 for English and French data sets respectively. These confusion matrices are actually normalized by its values to avoid the skewed 'O' tag distortion for compact presentation of matrices.

In English SBD task, the average precision of 'BS' and 'ES' tags is relatively less when compared to the average recall of these tags as these tags are mostly confused with ‘O’ tag as shown in Figure 1. We also observe that the ‘BS’ and ‘ES’ tags in English shared task are not confused with each other as they were with ‘O’ tag. Same is seen in the confusion matrices of the improved post shared task submission results as shown in Figure 2.

In the case of French SBD task, the average precision of ‘BS’ and ‘ES’ tags is higher when compared to the average recall as these tags were mostly confused with ‘O’ tag as shown in the Figure 1. We also observe a negligibly small number of cases, ‘BS’ and ‘ES’ tags were got confused. The same behavior could be observed in the improved results of post-shared task submission as shown in Figure 2.

6.1 Error analysis of English text

In the case of English SBD task, we manually examined those sentences that are misclassified as ‘O’ (45% of the total errors) and found that they tend to contain short sentences, sentences starting with lower case words, hyphens, bullet points, named entity tokens (for example: GAM Star); numbering/currency tokens with brackets (for example, (a), a), etc.). Following are some examples of these cases.

- includes but is not limited to:
- – of issues linked to “emerging-country risks”.
- a) to e) are raised to a maximum of 20% for investments in Shares
- The European Union.

Some example sentences where the end tag ‘ES’ is predicted as ‘O’ tag.

- Cash collateral received may only be
- – of issuers domiciled in emerging countries , or
- The minimum capital is equivalent in US Dollar to EUR 1,250,000.00.
- available to distributors who have entered into arrangements with the GAM Group.

We also observed a few sets of errors in the ground truth test data especially when sentences start with bullet points and having currency numbers and other punctuation symbols. In rest of the 55% errors where the ’O’ tag is predicted as ’BS’ and ’ES’, it is found that punctuations such as ‘-’, ‘:’ and ‘.’ inside the title, as non-boundary tokens are misclassified as ‘ES’ and the tokens followed by such non-boundary tokens are also misclassified as ‘BS’ label.

6.2 Error analysis of French text

In case of French SBD task, we manually examined those sentences that are misclassified as ‘O’ (65% of the total errors) and observed that the short sentences occurring in the title, bullet points, structured segments containing currency numbers, etc are attributed to major errors. The following are some sentences for the prediction error where the beginning token is predicted as ‘O’ tag.

- 52,45 Euros.
- * le solde, s’il existe, est réparti entre les Parts A et B comme suit:
- CAMGESTION, une société de gestion appartenant au groupe BNP Paribas.
- (a) une Personne non Eligible et,

Examples of sentences where the ending token is predicted as ‘O’ tag are given below.

- Le FCP est exposé, entre:
- Fonds commun de placement de droit français (FCP)
- Siège social : 1, boulevard Haussmann - Paris 75009
- Tous les jours ouvrés jusqu ’ à 11:00, heure de Paris.

The first sentence end with ‘:’, second and third sentences are isolated sentences and they are not ending with fullstop.

The sentences where both beginning and ending tokens are predicted as ‘O’ are mostly titles and short sentences. Some example sentences are given below.

- FIA soumis au droit français
- Éligibilité : PEA
- Néant.
- Parts G : 300000€

In rest of the 35% errors where the ‘O’ tag is predicted as ‘BS’ and ‘ES’, it is found that punctuations such as ‘-’, ‘:’ and ‘.’ inside the title, as non-boundary tokens are misclassified as ‘ES’ and the tokens followed by such non-boundary tokens are also misclassified as ‘BS’ label.

Same pattern of errors are observed in the results of post shared task submission as shown in tables 8 and 9. Properly handling these type of sentences would require modeling with complex features such as combinations of punctuation, indentation and formatting indicators, currency symbols with nonlinear chain CRFs and advanced deep sequential neural networks such as bi-directional LSTMs.

7 Conclusion and Future Work

We presented the experimental results of FinSBD Shared Task: CRF-based Sentence Boundary Detection in PDF Noisy Text in the Financial Domain. There are 2 tasks for English and French financial texts. We modeled SBD as a sequential modeling approach and obtained 84.5% (F1) on English data set and 86.5%(F1) in French task using basic features in combination with punctuation-based embeddings and syntactic POS tags. After correcting the bug in the prediction code, we actually observed 86.5% (F1) and 90%(F1) in English and French SBD tasks respectively.

Sentence boundary detection in noisy pdf texts poses challenges, such as working with semi-structured text containing format indicators such as bullets, numerals, financial numbers and specialized vocabularies such as named entities. One of the future directions is to explore proximity specific meta structural features with better sequence representation in the dynamic CRFs to capture long range dependencies. Experimenting with hybrid CRF and bi-directional long short-term deep memory networks would also be our future work for getting the improved results.

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.

[Evang *et al.*, 2013] Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos. Elephant: Sequence Labeling for Word and Sentence Segmentation. In *EMNLP 2013*, 2013.

[González-Gallardo and Torres-Moreno, 2018] Carlos-Emiliano González-Gallardo and Juan-Manuel Torres-Moreno. Sentence Boundary Detection for French with Subword-level Information Vectors and Convolutional Neural Networks. *arXiv preprint arXiv:1802.04559*, 2018.

[Grave *et al.*, 2018] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[Grefenstette and Tapanainen, 1994] Gregory Grefenstette and Pasi Tapanainen. What is a Word, What is a Sentence?: Problems of Tokenisation. 1994.

[Hartigan and Wong, 1979] John A Hartigan and Manchek A Wong. Algorithm AS 136: A K-means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[Kalchbrenner *et al.*, 2014] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A Convolutional Neural Network for Modelling Sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[Kiss and Strunk, 2006] Tibor Kiss and Jan Strunk. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*, 32(4):485–525, 2006.

[Korobov and Peng, 2014] M Korobov and T Peng. Python-crfsuite, 2014.

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[Maegaard and Spang-Hanssen, 1973] Bente Maegaard and Ebbe Spang-Hanssen. Segmentation of French Sentences. In *COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*, volume 2, 1973.

[Mikheev, 2002] Andrei Mikheev. Periods, Capitalized Words, etc. *Computational Linguistics*, 28(3):289–318, 2002.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[Palmer and Hearst, 1997] David D Palmer and Marti A Hearst. Adaptive Multilingual Sentence Boundary Disambiguation. *Computational Linguistics*, 23(2):241–267, 1997.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[Reynar and Ratnaparkhi, 1997] Jeffrey C Reynar and Adwait Ratnaparkhi. A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19. Association for Computational Linguistics, 1997.

[Riley, 1989] Michael D Riley. Some Applications of Tree-based Modelling to Speech and Language. In *Proceedings of the Workshop on Speech and Natural Language*, pages 339–352. Association for Computational Linguistics, 1989.

[Tomanek *et al.*, 2007] Katrin Tomanek, Joachim Wermter, and Udo Hahn. Sentence and Token Splitting based on Conditional Random Fields. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, volume 49, page 57, 2007.