FinSBD-2020: The 2nd Shared Task on Sentence Boundary Detection in Unstructured Text in the Financial Domain

Willy Au , **Bianca Chong** , **Abderrahim Ait Azzi** , **Dialekti Valsamou-Stanislawski**
Fortia Financial Solutions, France
{willy.au, bianca.chong, abderrahim.aitazzi, dialekti.valsamou}@fortia.fr

Abstract

In this paper, we present the results and findings of FinSBD-2020, the 2nd shared task on Sentence Boundary Detection in unstructured text of PDFs in the Financial Domain. This shared task was organized as part of the 2nd Workshop on Financial Technology and Natural Language Processing (FinNLP) of the conference IJCAI-PRICAI 2020. This second edition differs from its predecessor by introducing list structure extraction. Participating systems aimed at detecting boundaries of sentences, lists and list items by marking their beginning and ending boundaries in the text extracted from financial prospectuses. In addition, systems were also tasked to determine the hierarchy level of each item in its list. 5 teams from 4 countries participated (4 of which submitted a paper) in this shared task using different approaches.

1 Introduction

Sentences are arguably the most foundational units of language and are at the core of Natural Language Processing (NLP) architectures. Sentence Boundary Detection (SBD) has become fundamental as the first pre-processing step to high-level tasks in any Natural Language Processing (NLP) application, parsing textual data from a string of characters into linguistic segments (sentences).

Issues around SBD have not received much attention and most research has been confined to clean texts in standard reading formats such as the news and limited datasets such as the WSJ corpus [1] or the Brown corpus [2].

The first FinSBD [3] task aimed to provide further research on the issue of noise in machine-readable formats such as PDFs. The financial sector is one of many that uses PDFs as an integral form of documentation. Most PDF-to-text conversion tools introduce noise in the form of missing, erroneous, unordered characters and obstructing texts (from tables, page footers and page headers). Moreover, financial documents often use full-stop punctuation in various ways. For example, section numbers and enumerated items (e.g. "1.", "2.") and abbreviations (e.g. "S.A.", "LTD.") all contain periods. Consequently, applying out-of-the-box SBD tools can often yield inaccurate sentence boundaries (i.e., Stanford sentence segmenter[4], spaCy[5] , NLTK[6]). In order to function optimally, these tools require tinkering with inner heuristics based on punctuation, syntax and sometimes semantics.

However, what the first FinSBD did not address were the obstacles concerning text appearance and physical position within a document, especially in structuring text into units in the form of lists. Lists are a visual hierarchy of information that organizes data-rich documents into more easily read blocks. Simple lists containing two or three enumerated items may be restructured into sentences (Figure 1), but the notion of a "sentence" becomes lost when lists contain multiple sentences, paragraphs, or lists within the list (Figure 2). Boundaries of such structures are often undetectable with simple rule-based approaches that depend on sentence-ending punctuation.

Existing tools of SBD are unreliable when given unstructured text. They do no account for text position within a document page where visual information allows us to understand whether the text belongs to the document structure (e.g. page footers, page headers, footnotes, etc.) or structured information (e.g. lists, tables, titles, etc.). Extraction of such unstructured text results in incomplete sentences or multiple sentences embedded in a sentence. This hinders the performance of NLP application (i.e. POS tagging, information extraction, machine translation, etc.) which expects a well-formatted and grammatical sentence of which boundaries are clear [7].

For this reason, this year's task differs from its predecessor by introducing boundary detection on lists and list items, including a subtask for identifying each item's level in its list. Understanding the position of text in structures such as lists is essential for SBD in segmenting characters not only into sentences but into semantic units.

In this shared task, we first focus on extracting well-segmented sentences, lists and and list items from text originating from financial prospectuses by detecting and marking their beginning and ending boundaries. Secondly, we focus on determining the depth level of each item in its list. These prospectuses are official PDF documents in which investment funds precisely describe their characteristics and investment modalities.

In this paper we report the results and findings of the FinSBD-2020 shared task. The shared task was organized as part of The Second Workshop on Financial Technology

and Natural Language Processing (FinNLP) [1], collocated with IJCAI-PRICAI-2020. A total of 5 teams from 4 countries submitted runs and contributed 4 system description papers. All system description papers are included in the FinNLP workshop proceedings and cited in this report.

This paper is structured as follows: Section 2 describes previous work on SBD, Section 3 describes the task, Section 4 describes the shared task data, Section 5 describes the participants and their proposed systems, Section 6 describes the results and discussion and finally, Section 7 finishes the paper with conclusions.

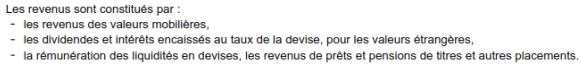
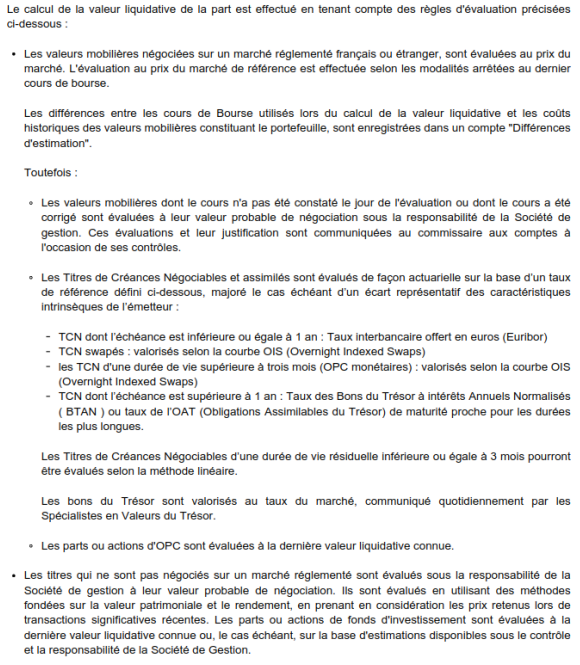

Les revenus sont constitués par :
- les revenus des valeurs mobilières,
- les dividendes et intérêts encaissés au taux de la devise, pour les valeurs étrangères,
- la rémunération des liquidités en devises, les revenus de prêts et pensions de titres et autres placements.

Figure 1: Simple list


Le calcul de la valeur liquidative de la part est effectué en tenant compte des règles d'évaluation précisées ci-dessous :

- Les valeurs mobilières négociées sur un marché réglementé français ou étranger, sont évaluées au prix du marché. L'évaluation au prix du marché de référence est effectuée selon les modalités arrêtées au dernier cours de bourse.

  Les différences entre les cours de Bourse utilisés lors du calcul de la valeur liquidative et les coûts historiques des valeurs mobilières constituant le portefeuille, sont enregistrées dans un compte "Différences d'estimation".

  Toutefois :

  - Les valeurs mobilières dont le cours n'a pas été constaté le jour de l'évaluation ou dont le cours a été corrigé sont évaluées à leur valeur probable de négociation sous la responsabilité de la Société de gestion. Ces évaluations et leur justification sont communiquées au commissaire aux comptes à l'occasion de ses contrôles.

  - Les Titres de Créances Négociables et assimilés sont évalués de façon actuarielle sur la base d'un taux de référence défini ci-dessous, majoré le cas échéant d'un écart représentatif des caractéristiques intrinsèques de l'émetteur :

    - TCN dont l'échéance est inférieure ou égale à 1 an : Taux interbancaire offert en euros (Euribor)
    - TCN swapés : valorisés selon la courbe OIS (Overnight Indexed Swaps)
    - les TCN d'une durée de vie supérieure à trois mois (OPC monétaires) : valorisés selon la courbe OIS (Overnight Indexed Swaps)
    - TCN dont l'échéance est supérieure à 1 an : Taux des Bons du Trésor à intérêts Annuels Normalisés (BTAN) ou taux de l'OAT (Obligations Assimilables du Trésor) de maturité proche pour les durées les plus longues.

  Les Titres de Créances Négociables d'une durée de vie résiduelle inférieure ou égale à 3 mois pourront être évalués selon la méthode linéaire.

  Les bons du Trésor sont valorisés au taux du marché, communiqué quotidiennement par les Spécialistes en Valeurs du Trésor.

  - Les parts ou actions d'OPC sont évaluées à la dernière valeur liquidative connue.

- Les titres qui ne sont pas négociés sur un marché réglementé sont évalués sous la responsabilité de la Société de gestion à leur valeur probable de négociation. Ils sont évalués en utilisant des méthodes fondées sur la valeur patrimoniale et le rendement, en prenant en considération les prix retenus lors de transactions significatives récentes. Les parts ou actions de fonds d'investissement sont évaluées à la dernière valeur liquidative connue ou, le cas échéant, sur la base d'estimations disponibles sous le contrôle et la responsabilité de la Société de Gestion.

Figure 2: Complex list

2 Previous Work on SBD

SBD has been largely explored following several approaches that could be classified into three major classes: (a) rule-based SBD, using hand-crafted heuristics and lists [8]; (b) machine learning approaches such as Naïve Bayes and Support Vector Machine (SVM) based models as reviewed in [2], decision tree classifiers [9] and the Punkt unsupervised model [10]; and more recently (c) deep learning methods [11]. Most of these approaches give fairly accurate results and prove to be highly accurate for most domain language data (e.g. clean collections of news articles). However, these systems are based on a number of assumptions [8] that do not hold for noisy, unstructured text extracted automatically from PDFs.

Read et al., [7] proposed a survey of publicly-available SBD systems such as CoreNLP, tokenizer, RASP and others. They evaluated several systems on a variety of datasets and report a performance decrease when moving from corpora with formal language to those with less formal language. Such designing and implementation customized to different domains has attracted the attention of several researchers. Griffis et al. [12] evaluated popular off-the-shelf NLP toolkits on the task of SBD for a set of corpora in the clinical domain. López and Pardo [13] tackle SBD on informal user-generated content such as web reviews, comments, and posts. Rudrapal et al,. [14] presented a study on SBD in a social media context. SBD from speech transcriptions has also gained much attention due to the necessity of finding sentential segments in transcripts created by automatized recognition. Carlos-Emiliano et al [15] tackled the problem of SBD as binary classification applied on an expansive written dataset (French Gigaword), an ASR transcription corpus. They focused on deep learning methods such as Convolutional Neural Networks to handle the task.

There are few papers that directly explore the problem of segmenting lists with items. Savelka et al. [16] treated the SBD problem in Adjudicatory Decisions [2], legal documents with a complex structures similar to prospectuses. They also conducted an analysis on tagging strategies of sentences and list of items, proposing several tags with which they then applied rule-base SBD systems such as OpenNLP and other trainable systems such as CRF (Lafferty et al., 2001; Liu et al., 2005; Okazaki, 2007) that prove to perform better on this kind of task. George Sanchez[17] worked on the same dataset, and explored the use of Punkt unsupervised model [10], CRF, and BiLSTM algorithm. They treated the problem as a sequence labeling task to predict the beginning and the end of sentences.

3 Task Description

The FinSBD-2020 shared task is an extension of FinSBD-2019, with the addition of list and list items boundaries. We have included lists and items due to their unique structure and common occurrence in financial documents. The first subtask consists of detecting the boundaries of three types of text segments: sentences, lists and list items. The second subtask requires distinguishing the hierarchy depth level of each item in its list. Each item can be assigned a depth level of 1, 2, 3 or 4.

The shared task provided a corpus of annotated data allowing supervised approaches. The annotated prospectuses were split into a train set and a hidden test set used for evaluating submitted systems. We provided one JSON file for each PDF (i.e. Figure 3) with the following keys:

- **text**: whole text extracted from the document
- **sentence**: boundaries of sentences
- **list**: boundaries of list
- **item**: boundaries of list items

[1] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp2020/

[2] https://github.com/jsavelka/sbd_adjudicatory_dec/tree/master/data_set

```
{
  "text": "... The risk management team of the Management Company may impose stricter criteria in terms of financial \nguarantees received and thereby exclude certain types of instruments, certain countries, certain issuers or even \ncertain securities. \n \nc) Level of financial guarantee \n \n The Management Company has put in place a policy which requires a level of financial guarantee based on the \ntype of transactions as follows: \n\uf0a7\n securities lending transactions: 105% of the value of the assets transferred; \n\uf0a7\n repurchase agreements and reverse repurchase agreements: 100% of the value of the assets transferred; \n\uf0a7\n OTC financial derivative instruments: In OTC financial transactions, some sub-funds may cover operations by \nmaking cash margin calls in the currency of the sub-fund in accordance with the restrictions laid down in \nclause 7.1 of the present Prospectus as regards the counterparty risk. ...",

  "sentence": [ ... {"start": 66281, "end": 66546}, {"start": 66550, "end": 66582} ... ],
  "list": [ ... {"start": 66587, "end": 67246} ... ],
  "item": [ ... {"start": 66753, "end": 66830}, {"start": 66836, "end": 66937}, {"start": 66943, "end": 67246} ... ],
  "item1": [ ... {"start": 66753, "end": 66830}, {"start": 66836, "end": 66937}, {"start": 66943, "end": 67246} ... ],
  "item2": [ ... ],
  "item3": [ ... ],
  "item4": [ ... ]
}
```


Figure 3: Example of a truncated JSON of a simple list with items of depth level 1 (actual JSON is composed of many more boundaries and the whole text from the document)

- **item1**: boundaries of list items of depth 1
- **item2**: boundaries of list items of depth 2
- **item3**: boundaries of list items of depth 3
- **item4**: boundaries of list items of depth 4

A boundary consists of a pair of integer indexes marking the starting and ending character of well-formed text segments (see Figure 3). Item depth level is the hierarchy level of the list to which the item belongs. Subtask 1 focuses on the detection of the boundaries of *sentence*, *list* and *item*. Subtask 2 focuses on detecting the boundaries of *item1*, *item2*, *item3* and *item4*. Note that *item* boundaries are equal to the union of all boundaries of items of different levels, $item = item1 \cup item2 \cup item3 \cup item4$. Therefore, subtask 2 can be formulated as classifying *item* boundaries into 4 classes.

One important detail is that the provided text was not pre-tokenized whereas in the first FinSBD, the text was already pre-tokenized at the word level. We hoped to encourage diverse approaches by not constraining participants to one type of word tokenization. Boundary indexes correspond to the character index and systems should predict pairs of character indexes as boundaries. Coordinates of each character were also provided in a separate JSON file for each PDF to encourage the creation of multi-modal system exploiting both textual and positional information. Each character's coordinates is referenced by its index in the text.

4 Shared Task Data

Next, we discuss the corpora used for the English and French subtasks.

4.1 Corpus annotation

For FinSBD-2019, annotated data for SBD was created by using Pdf2text and Brat tools [3]. Due to the many limitations of Brat (e.g. lack of visual cues, dependency between annotations and Brat), we decided to use a new annotation tool, tagtog[3], which allowed direct annotations on visualized PDF pages. This tool displays each document in its entirety and provides an ergonomic web interface that allows the annotator to select text directly on the PDF document. The visual component of PDFs with unique structures, graphs and images provides annotators valuable information that would otherwise not be available via Pdf2text and Brat. As a result, annotations were no longer dependent on PDF-to-text conversion tools. In addition, the annotation guidelines also had to be reworked to obtain better data with respect to a more linguistic approach.

Financial prospectuses were available both online in PDF format and directly from fund managers. We built a medium-sized dataset consisting 8 English (66 pages on average) and 33 French (26 pages on average) prospectuses.

Three bilingual (English and French) annotators were used to annotate these documents according to SBD's new annotation guidelines. The guidelines define what constitutes a "sentence" in financial documents and more specifically the different types of units and their boundaries that can be found in the text.

Guidelines Our guidelines consisted of 4 types of units: "sentences", "lists", "nesting lists" and "items".

A "sentence" was defined as a set of words that represents a complete and independent thought. A "sentence" usually contains a subject and a predicate along with independent and/or dependent clauses. A "sentence" could also be nominal and verbal groups that took form as a title of a passage of text.

A "list" was defined as an introduction followed by "items" of the same category which are read in a vertical manner. Lists composed of several other embedded lists and levels of items were considered "nesting lists". This nesting of lists can reach

[3] https://www.tagtog.net

up to four depth levels. It is important to mention that the typographic occurrence of bullets did not determine whether an enumeration was a list, as it could also simply be a sequence of independent "sentences".

In the corpora, we focused on extracting well-defined sentences, lists and items by detecting their boundaries and discarding non-phrases (figures, images, footers, page headers, etc.).

The annotated corpus was converted into the FinSBD-2020 labels mentioned in Section 3. Annotated sentence boundaries within item of list were not given to participants in order to simplify the shared task.

A total of 41 documents were annotated. To create the ground-truth, each document was annotated independently by two analysts, then reviewed and corrected by a third analyst.

Annotation Challenges Data annotation may have varied due to interpretations of the following ambiguities:

1. Lists were visually distinguished by bullets and numbers, but not always. Some lists did not contain visual indicators (bullets or numbers) of items and appeared to be only sentences.
2. Groups of sentences were sometimes found with bullets that had no semantic relationship and therefore did not make up part of a list.
3. The distinction between a title and items of a list were not always clear. Titles were visually distinguishable from text passages by differences in font types and sizes, bold or italics, underlined words, etc. These titles sometimes appeared to be items in a list.
4. The use of colons was inconsistent. Colons were sometimes used at the end of titles, followed by either grammatically complete or incomplete sentences. The boundaries of a "sentence" in this case becomes unclear.
5. Human errors in the documents such as missing punctuation, incorrect punctuation or incorrect grammar required that each annotator independently interpret what the intended text was.

4.2 Corpus Description

In this section, we provide an analysis of the data used for both subtasks in English and in French.

In Table 1, we report some statistics about the dataset. #Prospectuses indicates the number of financial prospectuses used in each set; #Page the total number of document pages in each the set; and finally the number of occurrence each classes #sentence, #list, #item (subtask1) and #item1, #item2, #item3 and #item4 (subtask2).

We also report the percentage of segments ending with a punctuation mark ("?", "!", ";", ".", ":") as well as percentage of segments starting with an uppercase letter to support the claim that SBD cannot solely rely on capital letters and punctuation. In the shared task data, only 75% up to 86% of text segments ended with a punctuation mark and only 28% up to 57% began with a capital letters. This is significantly lower than the numbers reported in FinSBD-2019 and is due to the introduction of list items which boundaries are more subtle than those of sentences. Hence, FinSBD-2020 presented a non-trivial problem, which had the potential to be solved by novel SBD systems that would leverage richer features, such as syntactic and semantic cues from the text, and features related to the position of the text in its page.

	English		French	
	train	**test**	**train**	**test**
# Prospectuses	6	2	23	10
# Page	350	180	624	224
# sentence	8070	2450	13164	4748
# list	249	69	494	173
# item	1111	332	1722	638
# item1	1029	272	1548	570
# item2	78	60	150	60
# item3	4	0	21	8
# item4	0	0	3	0
% Punct. as end	76%	86%	76%	75%
% Uppercase start	45%	28%	56%	57%

Table 1: Distribution of the Training and Testing sets used in the English and French corpora.

5 Participants and Systems

	# team submissions
subtask 1 EN	6
subtask 2 EN	2
subtask 1 FR	4
subtask 2 FR	1

Table 2: Statistics on the participation in the French and English subtasks.

A total of 18 teams registered in the shared task, of which 5 teams who participated and 4 who submitted a paper to describe of their method. The participants came from 8 different countries and belonged to 18 different institutions. The shared task brought together private and public research institutions including Rakuten, Flipkart Pvt Ltd, Subtl.ai. and Sorbonne University (see Table 3 for more details).

In table 2, we show the details on the submissions per task. One team who submitted boundaries did not send a paper describing their approach.

Participating teams explored and implemented a wide variety of techniques and features. In this section, we give a short summary of the methods proposed by each participating team (for further details, all papers appear in the proceedings of the FinNLP 2020 Workshop).

Team	**Affiliation**
PublishInCovid19	Flipkart, India
aiai	Rakuten, Japan
Daniel	Sorbonne University, France
Subtl.ai	Subtl.ai, India

Table 3: List of the 4 teams that participated and submitted papers in subtasks English and French of the FinSBD Shared Task.

PublishInCovid19 [18] This team formulated the boundary prediction problem as a sequence labeling task on overlapping windows of words. They first simplified the annotations by removing the recursiveness and hierarchy of items inside lists. Each word was given a beginning or ending label for a type of segment. Then, they compared two neural architectures, namely BiLSTM-CRF and BERT, trained on predicting the boundaries of sentences and simplified items. They used window sizes of 300 and 512 with overlapping of 20 words. Boundaries were post-processed by heuristics to correct missing beginning and ending boundaries. In the second phase, they identified the hierarchy and the recursive relation between items through a rule-based method applied on item boundaries predicted in the first phase. This allowed reconstitution of lists boundaries. The rules were based on visual cues like left-indentation and bullet-style of items. Their submitted system was the BiLSTM-CRF for both subtasks in English which achieved the highest score in the shared task.

aiai [19] This team approached the task as a two-stage text classification problem using two LSTM models with attention. First, they trained a multi-label boundary classifier to determine if a word is inside, outside, starting or ending a text segment. Each word is classified using a window of words, with additional features based on the word position and characters' width and height. They tested different window sizes and chose 21 as the most optimal (10 words before + 10 words after + current word). In a second stage, using these boundaries, they extracted candidate text segment and trained a second multi-label classifier to determine if the segment was a sentence, an item or a list. For both stages, the team used their own trained word embedding using CBOW on the shared task data. They managed to submit their system in English and French for both subtasks.

Daniel [20] This team decided not to use the provided textual representation. Instead, they utilized the "pdf2xml" converter to extract both text content and structural information, from which they extracted PDF structures in a top-down fashion, from higher-level to lower-level structures (i.e. the table of contents, tables, page headers and footers). Therefore, they were able to ignore text from table of contents, tables, page headers and footers. Finally, they created a set of heuristics based on bullet points, text position and font characteristics to identify lists, lists items and paragraphs. They exploited font features thanks to the use of the "pdf2xml" converter. Sentences were extracted from paragraphs by identifying end-sentence punctuation. The team submitted their system in English and French for the first subtask but not for the second subtask.

Subtl.ai [21] This team proposed an architecture combining a two-stage deep learning approach with heuristics. They first used a vocabulary to identify candidate words which could be sentence boundaries. Each candidate word was then represented by two windows of one-hot vectors derived from POS tags from 7 words located before and after. This was used to train a binary classifier, composed of 2 LSTM models, to determine if the candidate word was a true sentence boundary. From these boundaries, candidate segments of words were extracted for training a second LSTM with attention model, the input of which were pre-trained Glove word embeddings. This model was used to determine whether a segment was a true sentence. Multiple segments were merged if the concatenated sequence was classified as a true sentence. The team submitted their system in English for the first subtask and did not complete the second subtask.

6 Results and Discussion

In this section, we describe the evaluation metrics used in the shared task and we give an analysis of the results obtained for the various submitted systems.

Evaluation Metric Participating systems were ranked based on the macro F1-score of each subtask for each language obtained on a blind test set. A predicted boundary was considered to be true if both starting and ending indexes were correct. Consequently, this metric was more severe than the one used in FinSBD-2019 where a boundary could be considered true even if the corresponding starting or ending boundary was false. For each document, the F1-score was computed by label. Then, the scores of *sentence*, *list* and *item* were averaged as an F1-score of subtask 1 and those of *item1*, *item2*, *item3* and *item4* were averaged as an F1-score for subtask 2. Finally, the mean over all documents was taken as the macro-averaged F1-score to rank systems in each subtask by language.

We provided a starting kit[4] with an evaluation script and a baseline based on spaCy [5] for detecting only boundaries of sentences. Interestingly, low F1-scores of our baseline showed that applying out-of-the-box spaCy's SBD does not yield optimal results for our documents.

Table 4 and Table 5 reports the results by team obtained from FinSBD-2020 in English and French.

	English	
	subtask1	**subtask2**
PublishInCovid19	0.937	0.844
aiai	0.413	0.203
Daniel	0.317	0
Subtl.ai	0.217	0
our baseline	0.208	0
Anuj	0.126	0

Table 4: Ranking of teams according to macro-averaged F1-score for each subtask in English (0 means no submission).

Discussion As stated in Section 2, most previous work on SBD relied on unsupervised approaches based on heuristics derived from punctuation, letter capitalization, abbreviations and so on. This is mainly due to a lack of annotated data on unstructured text from documents. Through FinSBD, thanks to the introduction of annotated boundaries, we offered the opportunity of supervised approaches to tackle SBD given unstructured and noisy text. Each team had a unique approach in solving this problem and all teams who submitted a paper outperformed our baseline.

Most teams trained a supervised system on word-level labels they created by pre-processing the provided character-level

[4] https://github.com/finsbd/finsbd2

	French	
	subtask1	**subtask2**
PublishInCovid19	0	0
aiai	0.471	0.350
Daniel	0.232	0
Subtl.ai	0	0
our baseline	0.161	0
Anuj	0.025	0

Table 5: Ranking of teams according to macro-averaged F1-score for each subtask in French (0 means no submission).

	ranking score
	mean F1-score
PublishInCovid19	0.445
aiai	0.359
Daniel	0.145
Subtl.ai	0.054
our baseline	0.092
Anuj	0.038

Table 6: Ranking of teams by averaging the F1 scores obtained on each subtask for each language.

labels. This allowed application of transfer learning by using existing embeddings and architecture that expect words as input. Each word was assigned a class which served as start or end segments. Moreover, training a word-level model was computationally cheaper than character-level, the latter of which no team attempted.

There were two main approaches. The best performing one, proposed by *PublishInCovid19* [18], was sequence labeling: one multi-label architecture was trained to classify in one-go all words from a window into all different types of boundaries. The second approach, proposed by *aiai* [19] and *Subtl.ai* [21], was a two-stage classification architecture. A first stage model determined whether a word is a boundary given a window of surrounding words. Boundaries are then used in a second stage to create candidate segments, which were then classified by a second model into different types of segment: sentence, list or item. Separating the task into boundary detection and segment-type classification did not yield improvement over sequence labeling.

aiai [19] and *Subtl.aiai* [21] experimented with LSTM-based models with an attention mechanism in order to exploit dependencies between words for SBD for their classification tasks. *PublishInCovid19* [18] also based his model on LSTM layers, but with classic sequence labeling elements such as a CRF layer, bi-directionality and pre-trained word embeddings. Larger windows (300 and 512 words) [18] proved to be quite effective compared to smaller windows (7 and 21 words) [19] [21] for detecting both boundaries and their type. This was due to long dependencies between boundaries, especially of lists, which can span hundreds of words. There were also long dependencies between different types of segments, between lists and items for example, that large windows are better at detecting. Interestingly, *PublishInCovid19* [18] reported no significant improvement using large pre-trained language model such as transformers, i.e. BERT, compared to a BiLSTM-CRF with pre-trained word embedding. They respectively scored 0.956 and 0.959 weighted F1 scores in a sequence labeling setting, meaning there is little difference between both models. It is possible that there was a lack of sufficient data in order to leverage large transformers. In addition, transformer pre-training already depends on some type of sentence segmentation, which makes transformers ill-suited for predicting sentence boundaries.

All teams resorted in some extent to the use of heuristics based on text position, text appearance and/or punctuation to improve their SBD. *PublishInCovid19* [18] used a set of post-processing rules to resolve erroneous boundaries predicted by his models. *Daniel* [20] was the only team that used solely unsupervised rule-based approaches in their SBD system. Based on positional and syntactical heuristics, they explored a top-down pipeline for structuring PDFs into table of content, tables, page headers and footers and finally paragraphs and lists. Furthermore, other heuristics allowed them to extract clean segments in paragraphs and lists and exclude unwanted text from tables, page headers and footers. Their work possibly suggests that SBD of a document will only be solved once PDF structuring is. In FinSBD-2020, annotated boundaries excluded tables, page headers and footers and table of content.

For future work, it would be interesting to confirm if some of the submitted systems, *Subtl.ai* [21] and *PublishInCovid19* [18], experimented only in English, would perform as well on the French data where list items reaches up to depth level 4 (only 3 in English). Finally, *PublishInCovid19* expressed interest in exploring the idea of multi-modality by exploiting text, its position and its appearance equally in an end-to-end trainable system. In submitted systems, visual and positional features were only used in heuristics or as features complementing word-level representation during supervised training.

7 Conclusions

This paper presents the setup and results for the FinSBD-2020 Shared Task on Sentence Boundary Detection in Unstructured text in the Financial Domain, organized as part of The Second Workshop on Financial Technology and Natural Language Processing (FinNLP) of the conference IJCAI-2020. A total of 18 teams from 8 countries registered of which 4 teams participated and submitted papers in the shared task with a wide variety of techniques.

All supervised approaches were based on LSTM. The most successful method was based on a BiLSTM-CRF applied in a sequence labeling setting. The best average F1 scores on the FinSBD English subtasks were 0.937 for subtask 1 and 0.844 for subtask 2. And the best average F1 scores on the FinSBD French subtasks were 0.471 for subtask 1 and 0.35 for subtask 2. Despite high performance, especially for English, SBD is far from being completely resolved, particularly for list segmentation.

The diversity of both public and private institutions that participated in FinSBD-2020 illustrates that the issue of SBD

remains an area that requires further research and development especially concerning analysis of documents of unstructured formats. Achieving higher accuracy in sentence extraction that builds better NLP-based solutions proves to be a shared interest among a wide variety of fields.

Acknowledgments

We would like to thank our dedicated data and language analysts who contributed to building the French and English corpora used in this Shared Task: Sandra Bellato, Marion Cargill, Virginie Mouilleron and Aouatef Djillani.

References

[1] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[2] Dan Gillick. Sentence boundary detection and the problem with the us. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 241–244. Association for Computational Linguistics, 2009.

[3] 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 *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 74–80, August 2019.

[4] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[5] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[6] Edward Loper and Steven Bird. Nltk: the natural language toolkit. In *ETMTNLP '02: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational*, pages 63–70, 2002.

[7] Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. Sentence boundary detection: A long solved problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[8] Gregory Grefenstette and Pasi Tapanainen. What is a word, what is a sentence?: problems of tokenisation. 1994.

[9] 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.

[10] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525, 2006.

[11] Marcos V Treviso, Christopher D Shulby, and Sandra M Aluisio. Evaluating word embeddings for sentence boundary detection in speech transcripts. *arXiv preprint arXiv:1708.04704*, 2017.

[12] Denis Griffis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert M Lai. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. *AMIA Summits on Translational Science Proceedings*, 2016:88, 2016.

[13] Roque López and Thiago AS Pardo. Experiments on sentence boundary detection in user-generated web content. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 227–237. Springer, 2015.

[14] Dwijen Rudrapal, Anupam Jamatia, Kunal Chakma, Amitava Das, and Björn Gambäck. Sentence boundary detection for social media text. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 254–260, 2015.

[15] Carlos-Emiliano Gonzalez-Gallardo and Juan-Manuel Torres-Moreno. Sentence boundary detection for french with subword-level information vectors and convolutional neural networks. 02 2018.

[16] Jaromír Savelka, Vern R. Walker, Matthias Grabmair, and Kevin D. Ashley. Sentence boundary detection in adjudicatory decisions in the united states. 2017.

[17] George Sanchez. Sentence boundary detection in legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 31–38, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[18] Janvijay Singh. Publishincovid19 at the finsbd-2 task: Sentence and list extraction in noisypdf text using a hybrid deep learning and rule-based approach. In *The Second Workshop on Financial Technology and Natural Language Processing of IJCAI 2020*, 2020.

[19] Ke Tian, Hua Chen, and Jie Yang. aiai at the finsbd-2 task: Sentence, list, and itemsboundary detection and items classification of financial textsusing data augmentation and attentionmodel. In *The Second Workshop on Financial Technology and Natural Language Processing of IJCAI 2020*, 2020.

[20] Emmanuel Giguet and Gaël Lejeune. Daniel at the finsbd-2 task: Extracting lists and sentences from pdf documents, a model-driven approach to pdf document analysis. In *The Second Workshop on Financial Technology and Natural Language Processing of IJCAI 2020*, 2020.

[21] Aman Khullar, Abhishek Arora, Sarath Chandra Pakala, Vishnu Ramesh, and Manish Shrivastava. Subtl.ai at the finsbd-2 task: Document structure identification by paying attention. In *The Second Workshop on Financial*

Technology and Natural Language Processing of IJCAI 2020, 2020.