NUIG at the FinSBD Task: Sentence Boundary Detection for Noisy Financial PDFs in English and French

Tobias Daudert* and **Sina Ahmadi**
Insight Centre for Data Analytics
National University of Ireland, Galway
{tobias.daudert, sina.ahmadi}@insight-centre.org

Abstract

Portable Document Format (PDF) has become the industry-standard document as it is independent of the software, hardware or operating system. Publicly listed companies annually publish a variety of reports and too take advantage of PDF. This leads to the rise in PDF containing valuable financial information and the demand for approaches able to accurately extract this data. Analyzing and mining information requires a challenging extraction phase, particularly with respect to document structure. In this paper, we describe a sentence boundary detection approach capable of extracting complete sentences from unstructured lists of tokens. Our approach is based on the application of a language model and sequence classifier for both the English and the French language. The results show a good performance, achieving F1 scores of 0.855 and 0.91, and placed our team in 3rd and 5th for the French and English language, respectively.

1 Introduction

At a time we face an information deluge, automated solutions tailored to different formats are crucial for the data interpretation. In industry, Portable Document Format (PDF) has become the standard document as it is independent of the software, hardware or operating system in use [DocumentCloudTeam, 2013]. Publicly listed companies annually publish a variety of reports and too take advantage of PDF. In addition to factual information and numerical data, such documents provide deeper knowledge which is conveyed through wording and linguistic structure [Thomas, 1997]. With the rise in PDF containing valuable financial information, the demand for approaches able to accurately extract this data is also growing. However, analyzing and mining information requires a challenging extraction phase reliant on the document structure. Sentence boundary detection is vital to understand the document structure. Hence, this is the focus of the FinSB task and this paper.

Although not considered one of the grand challenges in natural language processing (NLP), sentence boundary detection remains challenging particularly due to textual variation [Read *et al.*, 2012]. Sentence boundary detection (SBD) aims at determining where a sentence begins and ends, in detail, it is the task of binary classifying text into boundary point or non-boundary point after each character [Read *et al.*, 2012]. SBD plays an important role in structuring textual data. For example, machine translation needs correct sentence segmentation as it heavily impacts the translation performance [Walker *et al.*, 2001], and speech recognition requires segmented sentences for the processing in downstream tasks as well as to improve human readability [Liu *et al.*, 2005]. SBD is paramount for text extraction in PDF since a major "problem in the conversion of PDF documents is the detection of the boundaries of common textual units such as paragraphs, sentences and words" [Tiedemann, 2014]. Although SBD is being researched for almost 20 years, the majority of works focus on structured texts (e.g. WSJ corpus, Brown corpus) and little attention is given to SBD in PDFs. In particular, research dealing with sentence boundary detection in financial PDFs is non-existing, to the best of our knowledge. The only related work found was the paper by Loughran and McDonald which deal with the readability of 10-k reports, however, the authors do not target sentence boundaries in their FOG index [Loughran and McDonald, 2014].

In this paper, we define SBD as the ternary classification of a token to identify the *sentence beginning*, *sentence end*, and *other token*. Below, we outline that *other token* variations occur in the form of *in-sentence-token* or *out-of-sentence-token*. Thus, our classification goes a step further and does not only aim at boundary points (i.e. sentence beginning and end) but is also able to determine a sentence within a list of tokens from its beginning to its end. This becomes particularly important for cases in which a sentence does not follow another sentence (e.g. a headline followed by a sentence).

The paper is organized as follows: First, we present work related to this paper; second, we define the research problem, third, we explain our methodology to deal with sentence boundary detection for domain-specific texts in the English and French language; fourth, we present the results of the methodology application and analyze these; lastly, we conclude this work with a methodology and findings summary.

*Contact Author

Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2019), pages 108-114, Macao, China, August 12, 2019.

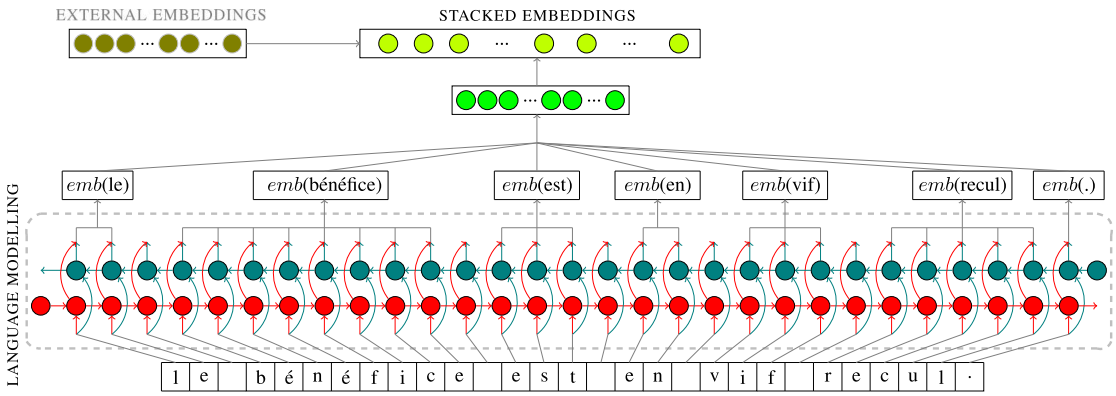

Figure 1: An illustration of our language modelling architecture. A bidirectional recurrent neural network, forward in red and backward in green, with LSTM mechanism retrieves the contextual embedding of each word at character-level. The produced embedding is then merged with an external embedding to create the stacked embeddings.

2 Related Work

Sentence boundary detection is a fundamental preprocessing step for the use of text in downstream tasks such as part-of-speech-tagging and machine translation. While rule-based approaches are the earliest method applied, we focus the related work on more advanced approaches, namely neural networks. The use of neural networks (NN) for sentence boundary detection dates as far back as 1994 [Cutting *et al.*, 1992].

Palmer and Hearst used a NN with two hidden units as an adaptable approach to overcome the restrictions of rule-based sentence boundary detection [Palmer and Hearst, 1994]. Their work utilised the part-of-speech (POS) surrounding sentence endings as an indicator. Since most POS tagger require available sentence boundaries, they inferred the POS based on the previous part-of-speech. When applied on a corpus of Wall Street Journal articles (WSJ), their work correctly disambiguated over 98.5% of sentence ending punctuation marks. Riley uses a decision-tree based approach to detect endings of sentences in the Brown corpus [Riley, 1989]. The maximum entropy approach by Reynar and Ratnaparkhi achieves an accuracy of 98.0 % on the Brown corpus and 97.5% on the WSJ corpus [Reynar and Ratnaparkhi, 1997]. In an effort to segment sentences in the output of vocabulary-speech-recognizers, Stolcke and Shriberg use a statistical language model to retrieve the probabilities of sentence endings [Stolcke and Shriberg, 1996]. They also mention the beneficial impact POS use can have. In a later work, Storcke et al. used decision trees to model a combination of prosodic cues aiming at the detection of events (i.e. sentence boundaries and disfluencies) [Stolcke *et al.*, 1998]. Dealing with a similar problem, Gotoh and Renals utilise n-gram language models to predict sentence boundaries from broadcast transcripts which have been converted to text [Gotoh and Renals, 2000]. Stevenson and Gaizauskas approach sentence boundary detection in automated speech recognition transcripts using a memory-based learning approach [Stevenson and Gaizauskas, 2000]. Other works used Hidden Markov Models (HMM) [Shriberg *et al.*, 2000] and Conditional Random Fields (CRF) [Liu *et al.*, 2005; Liu *et al.*, 2006]. Also in a machine translation setting, sentence boundary detection is important since it affects the translation quality. Walker et al. explore the use of three different algorithms to detect sentence boundary as pre-requisite of machine translation [Walker *et al.*, 2001]. They name the first algorithm, which is based on the Barcelona engine as part of the Power Translator, *The Direct Model*. The second algorithm is based on rules and employed as independent preprocessing, contrary to the first algorithm. The third algorithm is essentially a re-implementation of [Reynar and Ratnaparkhi, 1997]. Besides their results, which show the highest performance for the third algorithm, they also argue for its use as it is flexible in terms of adaption to other languages, fast in terms of training and delivers results only requiring a small corpus of labelled data, and straightforward in terms of feature selection. Kiss and Strunk propose a language-independent unsupervised SBD algorithm using Dunning's log-likelihood ratio on a tagged corpus [Kiss and Strunk, 2006]. Shriberg et al. compare different evaluation methods for the task of SBD [Liu and Shriberg, 2007]. Tomalin and Woodland compare two types of prosodic feature models for the task of sentence boundary detection [Tomalin and Woodland, 2006]. Particularly, they compare discriminatively trained Gaussian Mixture Models, CART-style decision trees, and task-specific language models with each other. Their results do not show a difference in performance between Gaussian mixture models and CART-style decision trees. An implementation of a multilingual sentence boundary detector is iSentenizer-μ [Wong *et al.*, 2014]. iSentenizer-μ first creates a binary decision tree, the authors call it offline training, on initially provided training data, which is continuously revised by an incremental tree learning algorithm whenever unseen data arrives.

3 Problem Definition

The goal of this shared task is to predict sentence boundaries from a list of words. Data is provided for two languages: English and French. In detail, it is provided a JSON file containing the fields $text$, $begin\_sentence$, and $end\_sentence$. $text$ contains the unsegmented text to be tagged, $begin\_sentence$ and $end\_sentence$ contain the indices of the beginning and

Parameter	Language Model	Sequence Classifier
hidden_size	2048	256
nlayers	1	1
mini_batch_size	100	32
epochs	2	100
sequence_length	250	-

Table 1: Parameter selection values for the language models and the sequence classifier training.

end of a sentence, respectively [Ait Azzi *et al.*, 2019]. In addition, a python script which applies two processing methods is given. On one hand, it splits the text into individual tokens using a white-space tokenizer, on the other hand, it creates a list of O tags replacing each O with a BS or ES in case its index is contained in the $begin\_sentence$ or $end\_sentence$ fields. After applying the python script to the file, we obtain two lists: The first list contains tokens while the second list contains tags. The used tags are [BS, ES, O] with BS representing the beginning of a sentence, ES the end of a sentence, and O other.

4 Methodology

In this section, we describe the approach designed to tackle the problem described in 3. It relies on two pillars: 1) The creation of two language models for each language to use as additional data, and 2) the training of two sequence classifier to tag the test data.

4.1 Language Modelling

One aim of language models is "to learn the joint probability function of sequences of words in a language" [Bengio *et al.*, 2003]. This makes them useful to our task as we can reformulate the sentence boundary detection challenge as a probabilistic problem in which we want to determine whether the following word in a string belongs to a sentence, given all previous words. Furthermore, recent developments in neural-network-based language models have shown relevant improvements, hence, we take advantage of such an approach for the extraction of word embeddings (e.g. [Devlin *et al.*, 2018; Peters *et al.*, 2018]).

The data provided in this shared task consists of PDFs containing financial prospectus, hence, we aimed at identifying corpora providing similar texts (i.e. organizational writing) for the training of the language models. For SBD in English texts, our choice fell on two corpora; the 10-k Corpus which contains 10-k reports filled by US companies between 1996 and 2006 [Kogan *et al.*, 2009], and the JoCo Corpus which consists of annual reports and company responsibility reports of a diverse set of companies (e.g. DJIA, FTSE100, DAX, S&P500, NASDAQ) collected between 2000 and 2015 [Händschke *et al.*, 2018]. Together, both corpora provide us with a diverse set of organizational writing from a period of 20 years in the English language. We further cleaned both corpora using multiple regular expressions to remove irregular breaklines and spacing, HTML tags, and JavaScript strings. Regarding SBD in French texts, we have not been able to to locate an appropriate corpus containing financial texts, especially organizational writing. Therefore, we created a novel corpus containing 2655 company reports, amounting to over 188 million tokens, from the 60 largest French companies by market capitalization, published between 1995 and 2018 [Ahmadi and Daudert, 2019].

The joint English corpus and CoFiF are then used to train two character-level language models using recurrent neural networks (RNN) for each language as illustrated in Figure 1. The language model is composed of two independent RNNs, in the forward direction and the backward direction, shown respectively in red and green in Figure 1. In the forward direction, the input sequence is fed in normal time order while in the backward direction, in reverse time order. The outputs of the two networks are concatenated at each time step. We used the publicly available NLP library Flair [Akbik *et al.*, 2018; Akbik *et al.*, 2019] for the language modelling. The training details are shown in table 1.

We conducted experiments to determine the quality of the trained language models by employing sentence perplexity calculations. The experimental results for the English language models are detailed below; the evaluation of the French language models is described in [Ahmadi and Daudert, 2019]. In both cases, we randomly selected 100 sentences from annual reports external to the corpora and rendered these meaningless by removing or replacing words, or characters. Having 100 correct sentences and 100 incorrect ones in place, we queried the language models for the sentence perplexity score for each sentence. The language model prediction is correct if it provides a lower perplexity score for the original sentence and a higher score for the modified sentence. Three sample sentences are shown in table 2. The language models are then used to extract word embeddings to use in the sequence classifier.

4.2 Sequence Classification

Given the similarity of the stated problem with part of speech (POS) tagging, we choose to re-train a sequence classifier also provided by Flair, as it has shown state-of-the-art performance on POS-tagging [Akbik *et al.*, 2018]. Instead of a list of tuples containing a word and the respective label, our sequence classifier requires segmented sentences. As the input for the sequence classifiers requires a TSV format, segmented sentences are separated by empty lines in the training data. Hence, we preprocess the training, development, and test data inserting an empty line after each \n. Furthermore, we conduct a second modification as part of our experiments; this modification includes manipulation of the labels. The originally provided data contains the labels [BS,ES,O] (section 3); we refer to their use as approach 1. However, we aim to provide further information to the classifier by introducing a fourth label [BS,ES,IS,O]. We refer to this as approach 2. The label IS stands for *in-sentence* and is determined during the preprocessing by labeling all words after begin-of-sentence and before end-of-sentence, as IS (i.e. in-sentence). Consider the following text *"October 2013 Distribution of this prospectus is not authorised ."*, where *October 2013* is part of the header. Approach 1 would label *Distribution* as $< BS >$ and . as $< ES >$; the remaining tokens are labelled as $< O >$. Whereas approach 2 would label *October* and

Sentence	Perplexity
GET SA's shares and the NRS issued by EGP have been listed on the London Stock Exchange since 2 July 2007.	2.9654
GET SA's shares and the NRS issued by EGP have been listed on the London Stock Exchange from 2 July 2007.	3.0157
The board of directors of GET SA has endeavoured to set up appropriate committees as envisaged by its internal procedures.	2.4056
The board of directors of SA has endeavoured to set up appropriate committees as envisaged by its internal procedures.	2.3441
In cooperation with the SNCF, Europorte 2, the rail freight subsidiary of Eurotunnel Group has just started its operational activity in the Frethun cross-Channel rail freight depot adjacent to the French end of the Tunnel.	3.9674
In cooperation with the , Europorte 2, the rail freight subsidiary of Eurotunnel Group has just started its operational activity in the Frethun cross-Channel rail freight depot adjacent to the French end of the Tunnel.	3.8726

Table 2: Six English sample sentences and their perplexity scores according to the character-level forward language model. The upper sentence of each pair is the original sentence, the lower sentence is the modified and wrong sentence.

2013 as $< O >$, *Distribution* as $< BS >$, . $< ES >$, and the remaining tokens as $< IS >$, since these occur between $< BS >$ and $< ES >$, to form a valid sentence. Although the ultimate goal is only to predict the BS and ES label, our intuition behind providing this additional knowledge to the classifier is that it might learn to differentiate between sentences and non-sentences (e.g. headlines) as a complete sequence, with BS being at the beginning of a sentence and ES at the end.

To fine-tune the sequence classifier parameters we split the development data into a development set and a test set by the ratio 70% / 30%; having a temporary test set available before the actual test data is released allowed us to experiment with the classifier. When the test data was released, only the last three sentences of the development data were used to test the final classifier and the remaining were included in the training set.

To train the classifier, the first step is to vectorize the data; to achieve this, we use the concept of stacked embeddings [Ammar *et al.*, 2016] and the embeddings from our language models. For the English data, we stack GloVe embeddings [Pennington *et al.*, 2014] with embeddings from the forward language model, and embeddings from the backward language model. Whereas for the French data, we stack fastText embeddings [Grave *et al.*, 2018] with embeddings from the forward language model, and embeddings from the backward language model [Ahmadi and Daudert, 2019]. As the pre-trained GloVe embeddings are only available in English, fastText was chosen in the French data vectorization. With the data prepared, the sequence classifiers were trained for each of the approaches and languages; the training parameters are presented in table 1.

5 Results

The results achieved are presented in two parts: we first provide an analysis of the created language models, and then report on the sequence classifier performance.

5.1 Language Model Evaluation

To evaluate the language model quality, we employed the sentence perplexity-based approach described in section 4.1. Although the sentence perplexity is not used directly to refine the sequence classifiers output, it influences the quality of the stacked embeddings which we employ to train the sequence classifiers. Thus, a good quality language model is imperative for the classification output. The character-level forward language model was tested on 117 random sentences extracted from an additional annual report. The model correctly identified 102 as original sentences and failed to detect 15; three sentence pair examples are shown in table 2, the top sentence is the original/correct version. A lower sentence perplexity score indicates a higher probability for the sentence to appear in this form. Considering these examples, in the first pair, we replaced *since* with *from* which rendered the sentence grammatically incorrect. The difficulty in the second example consists in knowing the structure of French company names, specifically that SA stands for *Société anonyme*, a company type; with the removal of *GET* the model failed to capture that this string/part of the company name is missing. However, we need to keep in mind that the English training data did not contain reports of French companies, thus, it is unlikely our language model has come across such names before. In the third example, we removed the company name *SNCF*. Although this mistake seems obvious to a human, the language model did not detect it. Looking closely at the wrong sentence, one can also understand it as *"In cooperation with the Europorte 2, the rail [...]"* and, hence, only see a misplaced comma.

5.2 Sequence Classification

The sequence classifiers are evaluated with the F1 score for the sentence boundary labels [BS, ES]. The results are shown in table 3. For the shared task in English, our approaches rank 5th and 12th out of 18 submissions; for the French task, our approaches rank 3rd and 4th out of 15 submissions. Comparing both approaches, the results for French and English are the same or higher in approach 1 than approach 2. This indi-

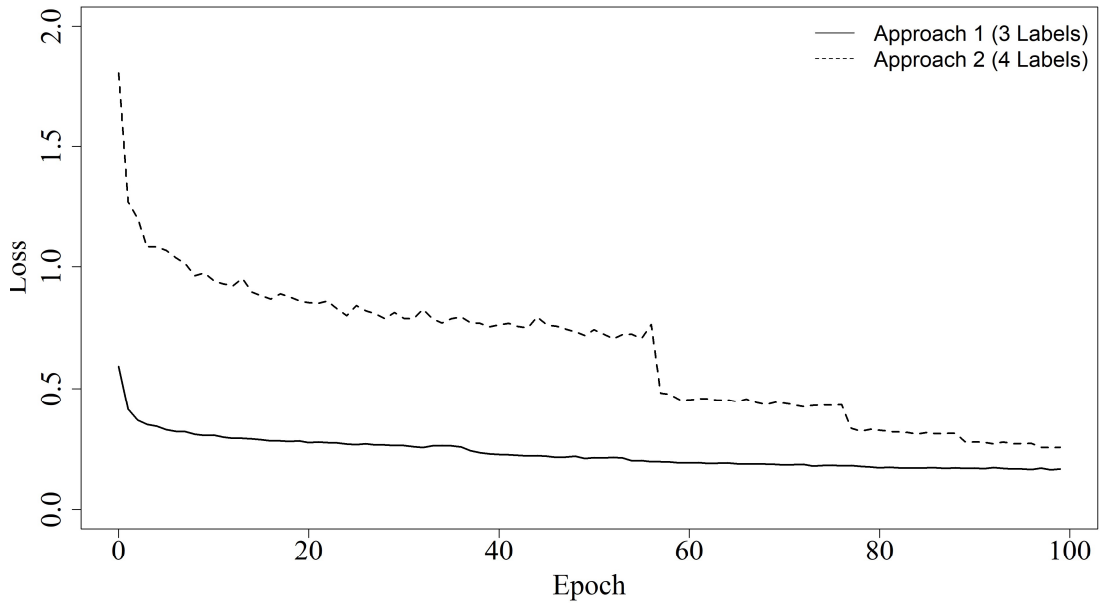

Figure 2: Training loss per epoch during the English sequence classifier training.

cates that the addition of a fourth label (IS) did not improve the classification. Figures 2 shows the training loss during the English classifier training. While the training loss for approach 1 steadily decreases, the decrease for approach 2 is rather unsteady. This volatility could also be an indicator for the difficulty in training the sequence classifier on data containing 4 labels. For the French data, the loss behaviour is similar.

The bottom graph in figure 2 shows a stabilizing loss towards the end of the training. This suggests that a prolonged training period could yield improved results. Nonetheless, a 4-label classification is inherently more difficult than a 3-label classification. The nuances between $< O >$ and $< IS >$ also suggest that additional training data is required.

Additionally, although we employ the same approaches to both languages, the F1 scores for French are generally higher than for English. We hypothesize this is due to three reasons: 1) the French language models are trained on data more similar to the test data, hence, providing better word embeddings for this particular task; 2) the stacked embeddings using fastText provide a better generalization than stacking embeddings with Glove; 3) French financial reports structure is stricter, making the sentence boundaries more predictable. We also have to consider natural languages differences between English and French which can have an effect on the performance of classification tasks. For all languages and approaches, the F1 scores for the end-of-sentence tag are higher than for the begin-of-sentence tag. We can also observe that approaches 1 and 2 achieve the same F1 score on French while approach 1 achieves different F1 scores for the end-of-sentence tag for English.

Language	Approach	F1 score		Mean F1 score
		BS	ES	
English	1 (3-labels)	0.81	0.9	0.855
	2 (4-labels)	0.81	0.85	0.83
French	1 (3-labels)	0.9	0.92	0.91
	2 (4-labels)	0.9	0.92	0.91

Table 3: sequence classifier evaluation results.The BS and ES tag represent begin-sentence and end-sentence.

6 Conclusions

In this paper, we described our approach to detect sentence boundaries in a corpus of unsegmented text. This approach is tested on English and French data. To this purpose, we utilize two powerful character-level language models, as well as two sequence classifier for each language. In addition, we target two approaches, one based on the original labels and another introducing a modified label set. Our results yield a good performance placing us at the 3rd rank of this shared task for French and 5th for English. Specifically, the submitted approach for French achieves an F1-score of 0.91 while the approach for English retrieves an F1 score of 0.855.

Our results suggest that fine-tuning the models by training two domain-specific language models and using these to retrieve word embeddings as input for the sequence classifier is key to the achieved performance. Furthermore, we believe that the use of embeddings from other domains (i.e. GloVe and fastText) also contributed to the performance as it avoids a narrow domain focus.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund.

References

[Ahmadi and Daudert, 2019] Sina Ahmadi and Tobias Daudert. Cofif: A corpus of financial reports in french language. In *The First Workshop on Financial Technology and Natural Language Processing (FinNLP 2019)*, Macao, China, 2019.

[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.

[Akbik *et al.*, 2018] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[Akbik *et al.*, 2019] Alan Akbik, Tanja Bergmann, and Roland Vollgraf. Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page to appear, 2019.

[Ammar *et al.*, 2016] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*, 2016.

[Bengio *et al.*, 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[Cutting *et al.*, 1992] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing*, 1992.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DocumentCloudTeam, 2013] DocumentCloudTeam. Top 10 reasons to use pdf instead of word, excel or powerpoint, 2013.

[Gotoh and Renals, 2000] Yoshihiko Gotoh and Steve Renals. Sentence boundary detection in broadcast speech transcripts. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[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.

[Händschke *et al.*, 2018] Sebastian GM Händschke, Sven Buechel, Jan Goldenstein, Philipp Poschmann, Tinghui Duan, Peter Walgenbach, and Udo Hahn. A corpus of corporate annual and social responsibility reports: 280 million tokens of balanced organizational writing. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 20–31, 2018.

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

[Kogan *et al.*, 2009] Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics, 2009.

[Liu and Shriberg, 2007] Yang Liu and Elizabeth Shriberg. Comparing evaluation metrics for sentence boundary detection. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–185. IEEE, 2007.

[Liu *et al.*, 2005] Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. Using conditional random fields for sentence boundary detection in speech. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 451–458. Association for Computational Linguistics, 2005.

[Liu *et al.*, 2006] Yang Liu, Nitesh V Chawla, Mary P Harper, Elizabeth Shriberg, and Andreas Stolcke. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech & Language*, 20(4):468–494, 2006.

[Loughran and McDonald, 2014] Tim Loughran and Bill McDonald. Measuring readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671, 2014.

[Palmer and Hearst, 1994] David D Palmer and Marti A Hearst. Adaptive sentence boundary disambiguation. In *Proceedings of the fourth conference on Applied natural language processing*, pages 78–83. Association for Computational Linguistics, 1994.

[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.

[Peters *et al.*, 2018] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[Read *et al.*, 2012] Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. Sentence boundary de-

tection: A long solved problem? *Proceedings of COLING 2012: Posters*, pages 985–994, 2012.

[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.

[Shriberg *et al.*, 2000] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, 32(1-2):127–154, 2000.

[Stevenson and Gaizauskas, 2000] Mark Stevenson and Robert Gaizauskas. Experiments on sentence boundary detection. In *Sixth Applied Natural Language Processing Conference*, 2000.

[Stolcke and Shriberg, 1996] Andreas Stolcke and Elizabeth Shriberg. Automatic linguistic segmentation of conversational speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 2, pages 1005–1008. IEEE, 1996.

[Stolcke *et al.*, 1998] Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Mari Ostendorf, Dilek Hakkani, Madelaine Plauche, Gokhan Tur, and Yu Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Fifth International Conference on Spoken Language Processing*, 1998.

[Thomas, 1997] Jane Thomas. Discourse in the marketplace: The making of meaning in annual reports. *The Journal of Business Communication (1973)*, 34(1):47–66, 1997.

[Tiedemann, 2014] Jörg Tiedemann. Improved text extraction from pdf documents for large-scale natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 102–112. Springer, 2014.

[Tomalin and Woodland, 2006] Marcus Tomalin and Philip C Woodland. Discriminatively trained gaussian mixture models for sentence boundary detection. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

[Walker *et al.*, 2001] Daniel J Walker, David E Clements, Maki Darwin, and Jan W Amtrup. Sentence boundary detection: A comparison of paradigms for improving mt quality. In *Proceedings of the MT Summit VIII*, volume 58, 2001.

[Wong *et al.*, 2014] Derek F Wong, Lidia S Chao, and Xiaodong Zeng. isentenizer-: Multilingual sentence boundary detection model. *The Scientific World Journal*, 2014, 2014.