D Task: etection for Noisy sh and French

Tobias Daudert and Sina Ahmadi

Insight Centre for Data Analytics

Introduction

Portable Document Format (PDF) has become the industry-standard document as it is independent of the software, hardware or operating system. This leads to the rise in PDF containing valuable financial information and the demand for approaches able to accurately extract this data. In this paper, we describe a sentence boundary detection approach capable of extracting complete sentences from unstructured lists of tokens.

Goals

- Sentence boundary prediction from an unsegmented list of words
- Label the sentence beginning (BS), sentence end (ES), and the remaining positions (O)

Methodology

Our approach relies on the creation of two language models (LM) to use as additional data and on the training of two sequence classifier to tag the test data. The following steps are taken:

1. Given the provided task data, we identify similar corpora:
   - English: 10-k corpus [Kogan *et al.*, 2009], JoCo [Händschke *et al.*, 2018]
   - French: CoFiF [Daudert and Ahmadi, 2019]
2. We train a forward and a backward character-level LM for each language using recurrent neural networks;
3. Using the concept of stacked embeddings, we use these LM embeddings with Glove (for English) and FastText (for French) to vectorize the task data;
4. We re-train a state of the art POS-tagger with the [BS,ES,O]-tagged text using a ratio of 70% / 30%;
5. We further experiment by adding a fourth label, in-sentence (IS).

For the training and fine-tuning, we split the provided data by the ratio 70% / 15% / 15% into a training set, development set, and test set; for the final training run, we use all the data.

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

Table: Parameter selection values.

Evaluation

- The language models are evaluated using the sentence perplexity: We randomly select 100 sentences unseen by the LM during training in each, English and French; both sets are duplicated and all duplicates are rendered wrong; finally, the sentence perplexity for all 200 sentences is calculated. The LM prediction is correct if the perplexity is higher for the original sentence.
- The sequence classifiers are evaluated using the F1 score for the sentence boundary labels [BS, ES].

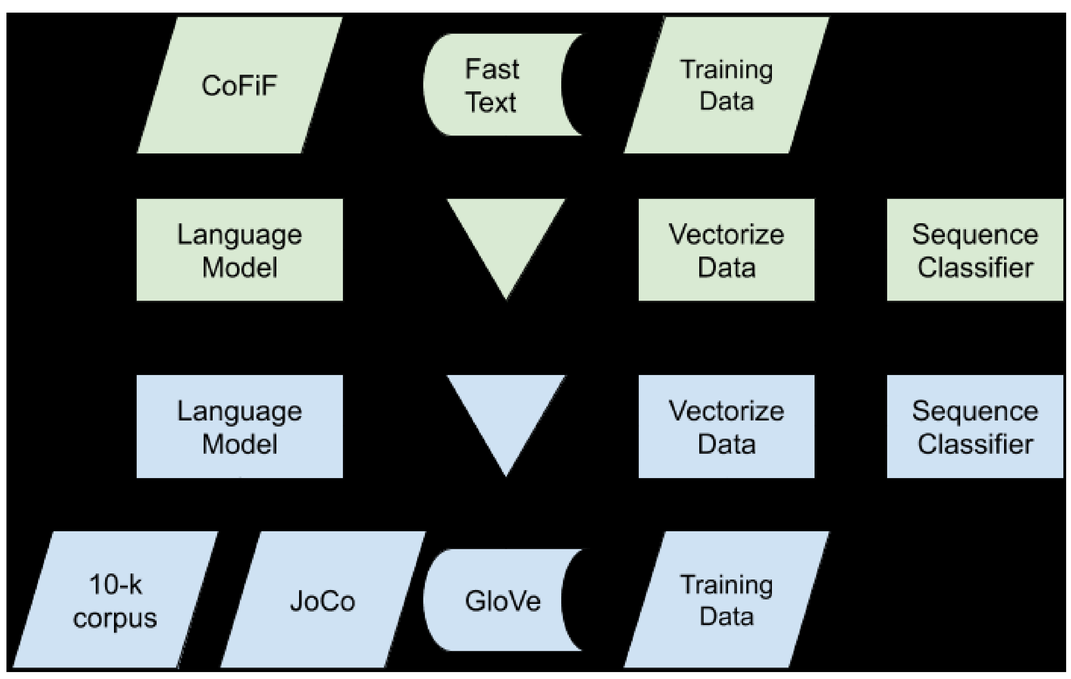

Figure: Model architecture for French in green and English in blue.

Results

- The character-level forward LMs are tested on random sentences extracted from an additional annual report. The English model is tested on 117 sentences and correctly identified 102. The French model is tested on 100 sentences and identified all the modified correctly.
- The sequence classifier evaluation results are presented below. The BS and ES tag represent begin-sentence and end-sentence.

Language	Approach	F1 score BS	F1 score ES	Mean F1 score
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

Conclusion

The experiments show that the addition of a fourth label (IS) does not improve the classification. The contrast in the training loss during the training of the English classifiers for approach 1 and 2 reveals an increased difficulty in learning the in-sentence label (IS).
Overall, our 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 task, respectively. This further suggests the reliability of our approach for different languages.

References

Tobias Daudert and Sina Ahmadi. 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.

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.

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.

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