PolyU_CBS-AF at FinSBD: Sentence Boundary Detection of Financial Data with Domain Knowledge Enhancement and Bilingual Training

Mingyu Wan[1] Rong Xiang[2] Emmanuele Chersoni[1] Natalia Klyueva[1] Kathleen Ahrens[3] Bin Miao[4] David Broadstock[4] Jian Kang[4] Amos Yung[3] Chu-Ren Huang[1]

[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
[2]Department of Computing, The Hong Kong Polytechnic University
[3]Department of English, The Hong Kong Polytechnic University
[4]School of Accounting and Finance, The Hong Kong Polytechnic University

12 August 2019

Task Overview

- **Purpose:**
  - to detect sentence boundaries (the begins and ends) from noisy texts of financial prospectuses.
- **Challenges:**
  - documents encoded in non-text formats, such as tables, images, etc.
  - conventional punctuation does not necessarily mark sentence boundary.
- **Our Approach:**
  - Machine Learning: a supervised task of three-class predictions + Feature Engineering: domain-specific and document-specific observation + Rule-based Validation: also know as domain knowledge enhancement.
- **Performance:**
  - English: 0.835 (avg. F-score of BS and ES)
  - French: 0.86 (avg. F-score of BS and ES)

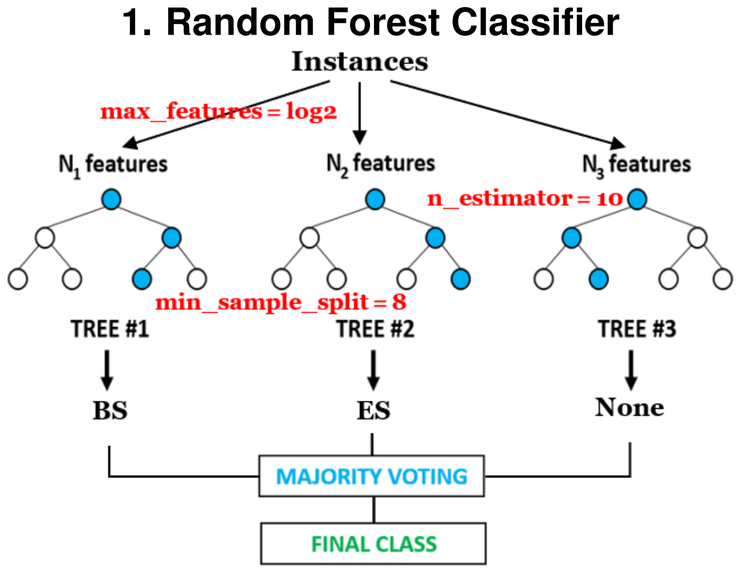
Feature Engineering

A fused feature set of 24 dimensions, including:

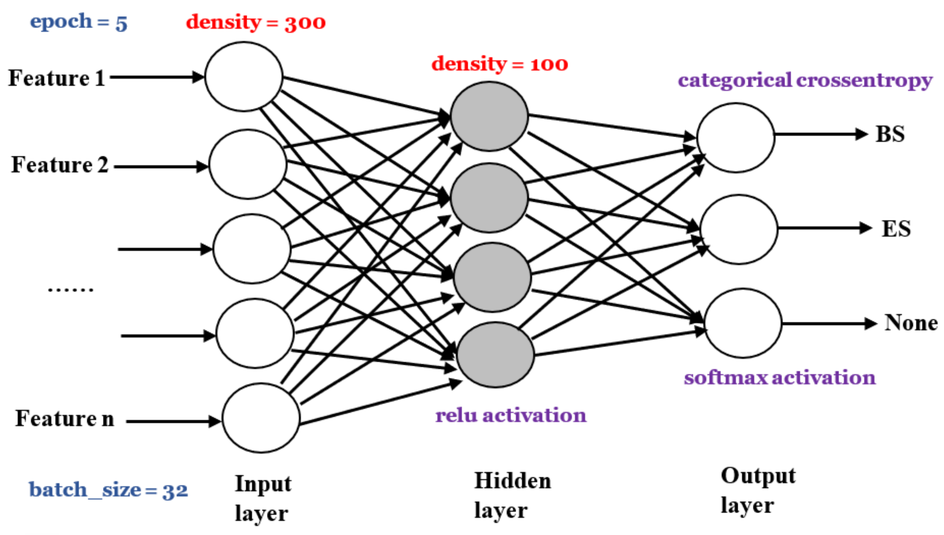
- **Two sets of punctuation:**
  - PUNC_SET1 = [‘.’]
  - PUNC_SET2 = [‘?’, ‘!’, ‘;’, ‘%’, ‘-’, ‘/’, ‘”’, ‘\’, ‘)’, ‘(’, ‘∗’, ‘⊡’, ‘<’, ‘≥’, ‘>’, ‘≤’, ‘•’, ‘e’, ‘$’, ‘£’, ‘“’, ‘©’, ‘®’]
- **Initially capitalized words**, *e.g.* “._Enter_END The_BEGIN sales”;
- **Acronyms**, *e.g.*“UBS_BEGIN” “co” or “kiid”;
- **Digital numbers**, ambiguity resolution as in “10.3”;
- **Letters or Roman numbers**, *e.g. e.g. A-Z, a-z* or *I, II, . . . , XII*;
- **POS tags**, structural cues.

Classifiers

1. Random Forest Classifier



2. Neural Network



Domain Knowledge Enhancement

Algorithm 1 Keyword Extraction

Input: dataset
Output: keyword_dict

```
for i in len(dataset) do
    curword = dataset[i].
    nxtword = dataset[i+1].
    if “END” in curword then
        if “BEGIN” in nxtword then
            continue
        else
            add keyword to keyword_dict.
            update frequency in keyword_dict.
        end if
    end if
end for
refine keyword_dict with length threshold.
refine keyword_dict with frequency threshold.
return keyword_dict
```

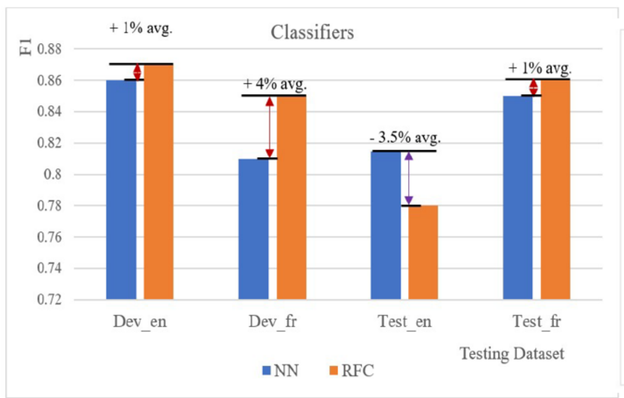
Algorithm 2 Rule-based validation

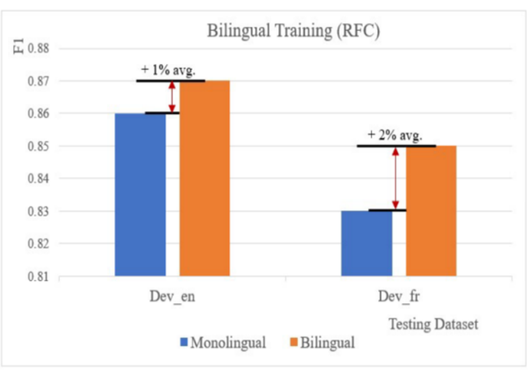
Input: dataset, raw_pred, keyword_dict
Output: updated_prediction

```
for keyword in keyword_dict.key do
    for i in len(dataset)-len(keyword) do
        if word sequence match keyword then
            update raw_pred with NO_BOUNDARY
        end if
    end for
end for
return raw_pred
```

- The two steps can correct false positive predictions caused by non-text sections, such as the titles, dates, tables, figures, etc., which fail to fall into the traditional category of sentence boundary.

Results and Comparison





- RFC outperforms NN in ¾ tasks
- NN outperforms RFC in the English Test set

Features\F1	BS[a]	ES[b]	Mean	δ(%)[c]
Punc1	0.70	0.82	0.76	baseline
+Punc2	0.71	0.83	0.77	1 ↑
+Cap	0.72	0.86	0.79	2 ↑
+Acro	0.73	0.87	0.80	1 ↑
+Dig	0.73	0.89	0.81	1 ↑
+Lett	0.74	0.90	0.82	1 ↑
+POS	0.78	0.92	0.85	3 ↑
+Enter/All	0.80	0.92	0.86	1 ↑

[a] Beginning boundaries
[b] Ending boundaries
[c] F1 improvement in percentage

Table 1: Performance of feature mining in the English Dev set with RFC

- **POS** vector is the most useful feature.

- Bilingual training consistently outperforms monolingual training with 1-2 % F gain.

	F1	NO[a]	YES[b]	δ(%)
Dev	BS	0.83	0.83	0
	ES	0.86	0.87	1 ↑
	Mean	0.845	0.85	0.5 ↑
Test	BS	0.81	0.84	3 ↑
	ES	0.88	0.88	0
	Mean	0.845	0.86	1.5 ↑

[a] Without keyword validation
[b] With keyword validation

Table 2: Performance of RFC in the French Dev and Test sets in terms of keyword validation

- Knowledge enhancement is **significantly** helpful.

Adaptation to Test Set

Test Set	F1-Mean	Rank
Dev_en_rfc1	0.875	—
Test_en_rfc1	0.78*	16◇
Test_en_rfc1_adapted	**0.835⋆**	**10⋇**
Dev_fr_rfc1	0.85	—
Test_fr_rfc1	0.86*	8◇
Test_fr_rfc1_adapted	0.86⋆	8⋇

* Result without adaptation to the test set
◇ Rank without adaptation to the test set
⋆ Result with adaptation to the test set
⋇ Rank with adaptation to the test set

Table 3: Performance of RFC *w.r.t.* knowledge adaptation

- Our approach works significantly better with adaptation to the English test set.
- Our system is robust for the French dataset with or without adaptation to the test set.

Conclusions

- NN does not show advantage over statistical classifiers, but it outperforms RFC in terms of unknown features;
- Syntactic information may be helpful for sentence detection;
- Our system is highly effective with minimal training data;
- Future studies will further verify the effectiveness of this feature design and knowledge enrichment.

Contact Person: WAN, Mingyu Clara, Email: mingyu.wan@polyu.edu.hk. Website: https://www.researchgate.net/profile/Mingyu_Wan/.

Acknowledgements: This work is partially supported by the GRF grant (No. 15608618) and the PDF project (No. 4-ZZKE) at PolyU HK.