New Energy and Industrial Technology Development Organization
JAPAN

Improving Financial Terminologies Recognition regarding to Morphological Inflections

Presenter: Ziwei XU

Date: 06 / 06 / 2023

Site: JSAI (Kumamoto, JP)

Authors: Ziwei Xu, Rungsiman Nararatwong,
Natthawut Kertkeidkachorn,
Ryutaro Ichise

NATIONAL INSTITUTE OF
ADVANCED INDUSTRIAL SCIENCE
AND TECHNOLOGY (AIST)

OUTLINES

I. Problem

II. Task Description

III. Dataset Preparation

IV. Performance Evaluation

V. Result

1. Problem

- recognize financial terminologies from text
- in terminologies, a proper name might have diverse expressions, like abbr. and morphological inflections

Example (terminologies & their occurrences in text)

- Given the string similarity score by Levenshtein Distance

Terminology: IPO

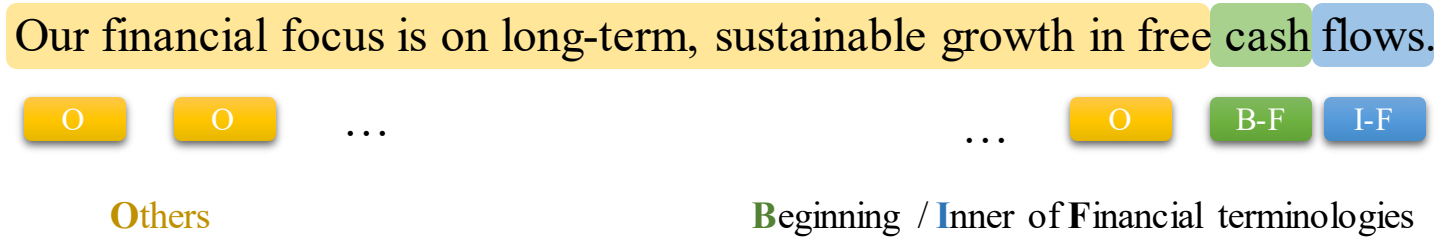
- ❑ IPOs (0.86)
- ❑ Initial Public Offering (0.23)
- ❑ Initial public offerings (0.07)
- ❑ Initial public offer (0.09)
- ❑ Initial public offers (0.08)
- ❑ pre-IPO (0.6)

Terminology: non-operating income

- ❑ nonoperating income (0.95)
- ❑ non operating income (0.86)
- ❑ non operating revenue (0.73)
- ❑ operating incoming (0.79)

2. Task Description

Initial Task: to detect financial terminologies from text (BIO scheme)



Practical Task: to classify tokens from given sentences

General methodology

- train a token classifier to distinguish the 3 labels for given text.

Scope focus

- apply the pre-trained language models, i.e. XLM-Roberta, is more beneficial to reflect semantic/syntactic relations between text

3. Dataset Preparation

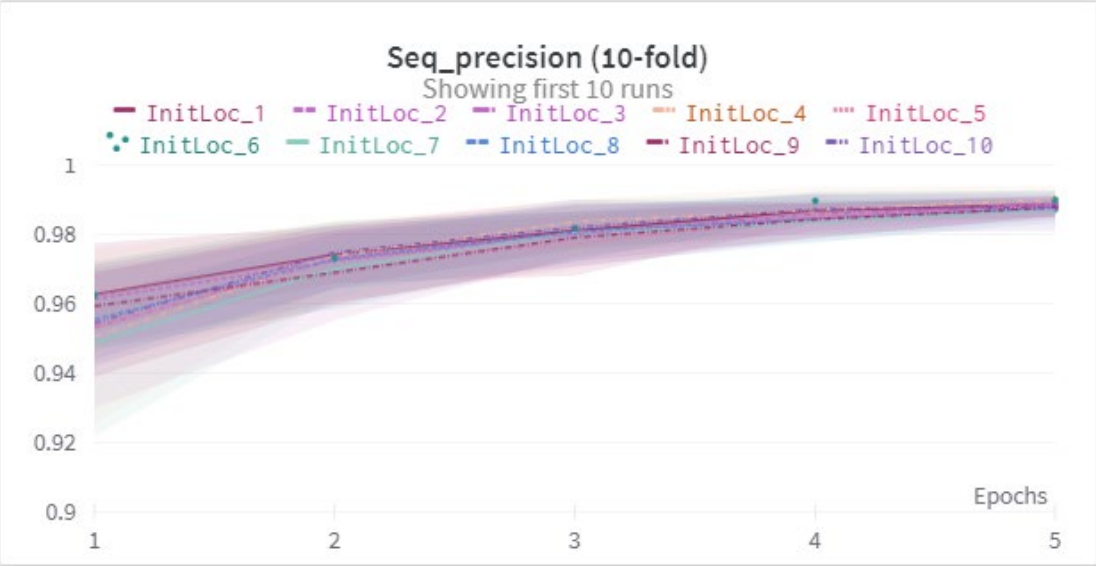
Resources

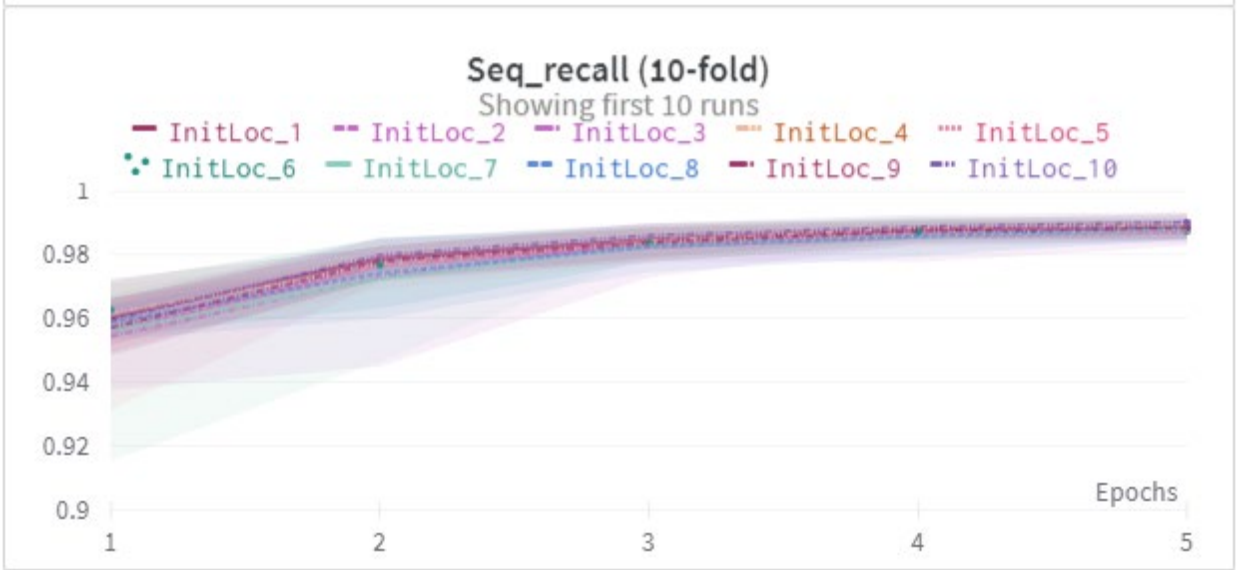
- Financial Reports: S.E.C. 10-k report (497K sentences)
- Financial Terminologies: Investopedia financial dictionary (6259terminologies)

Steps	Amounts of Unique Terminologies		
pre-processing	#investo.	6259	
	#abbr.	1192	
		Train/Val	**Prediction**
model	#sents	$497k \times \frac{9}{10}$	$497k \times \frac{1}{10}$
	#investo.	1317	1317
	#new	-	1623

Table 1: The statistics of terminologies from different resources.

4. Performance Evaluation





Experiments:

1. deploy **10-fold cross-validation strategy**, where 9/10 reports are used for model training and validation and the rest 1/10 reports perform as the unseen text for model prediction.
2. apply the **sequence labeling metrics – seqeval**, which reflects the recognition of an entire terminology, but not the separate tokens

Analysis:

- At epoch 5, we observed the rather good performances (approximately 0.99) in all metrics.
- Our models **remember well** which are terms of ground truth and which are not.

5. Results

- investigate **whether** those models have **sufficient prediction capability** on financial keywords.

Table 2: The predicted terminologies from the model.

	Investo.	Prediction
k.	tier 1 capital	tier i capital
	lease payments	lease payment
	ground lease	ground leases
	homeowners insurance	homeowners' insurance
	non-operating income	non operating income
	non-operating income	nonoperating income
	adjusted ebitda	adjusted ebitdar
	adjusted ebitda	adjusted ebitdda
n.	-	goodwill tax
	-	goodwill allocation
	-	convertible preference shares
	-	knight frank
	-	profit sharing liability
	-	hazard ratio

k.nown financial terminologies from Investopedia.

n.ew terminologies that were not existing in the training dataset but were **somehow meaningful and surprising prediction words.**

Thanks for your listening.