Team 5: Acronym Extraction & Disambiguation

Snehal Kumar, 2019101003
Shrey Gupta, 2019101058
K V Aditya Srivatsa, 2018114018
Mukund Choudhary, 2018114015

01

02

03

04

Table of contents

01
02
03
04

01

Introduction

Brief on the task and the datasets used.

01

02

03

04

The Tasks

Acronym Extraction: to identify acronyms and the related long forms from the documents in the dataset given

Acronym Disambiguation: to select the correct meaning of an ambiguous acronym in a given sentence of the dataset from full-forms given in a dictionary

01

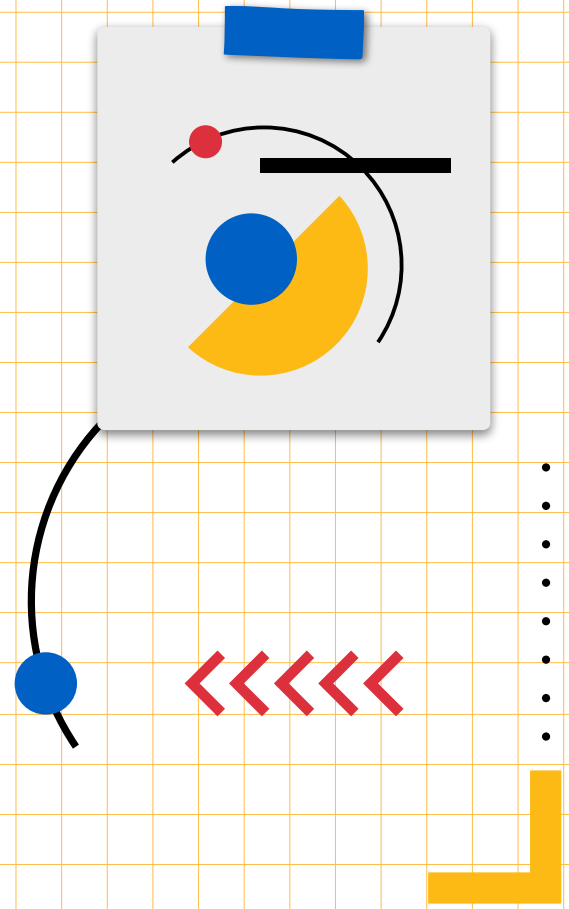
02

03

04

Dataset Description - AE

- Dataset consists of English sentences of scientific and legal domains. It has approximately 4000 paragraphs each in both domains
- Each datapoint contains:
  - Sentence
  - Index span of acronyms in the sentence
  - Index span of long forms in the sentence



01

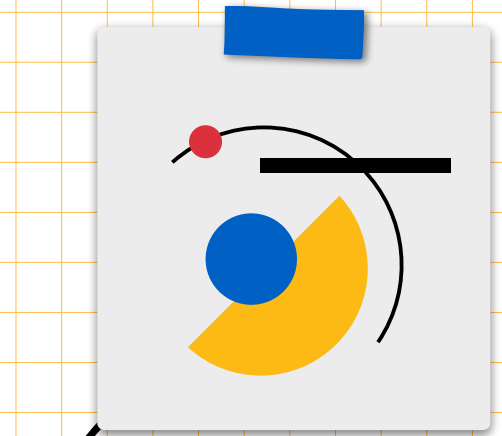
02

03

04

Dataset Description - AD

- Dataset consists of English sentences of scientific and legal domains.
- Contains diction.json which has 457 acronyms in the Scientific domain & 273 in the Legal domain
- Each datapoint contains:
  - Sentence
  - Acronym in sentence
  - Long form of Acronym



01

02

03

04

Trials

Describes our initial attempts for the task

AE

Dataset Experiments

With and without symbols, brackets etc.

Model Experiments

BERT: Legal & Sci Sci-Bert

Trainable Params MLP

Hidden layers, # neurons in them, class balancing...

Best params: 1 Hidden Layer with 200 neurons

01 02 03 04

AD

Dataset Experiments

With and without augmentation.

Model Experiments

BERT: Legal
Sci-Bert: Cased & Uncased

Fine Tuning

Masked-Language Modeling

01

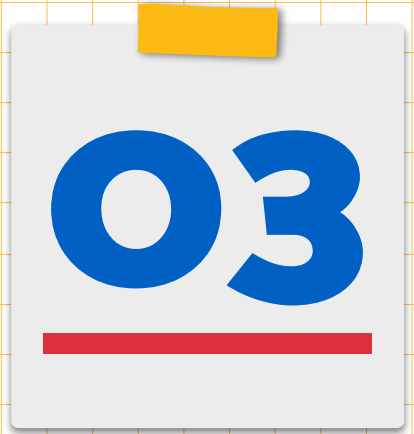
02

03

04

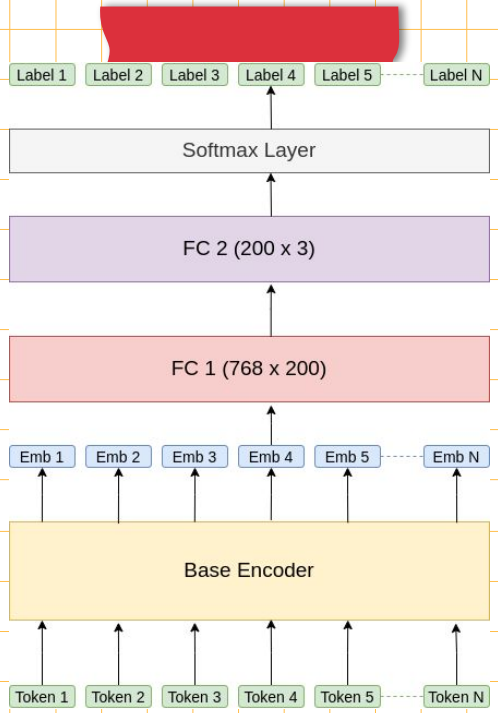
Architecture

Describes best AD & AE models



AE

- BERT-Sklearn Library's BertTokenClassifier
- MLP head consisting of 3 Fully Connected (FC)
  - FC1 (768 x 200)
  - FC2 (200 x 3)
  - ReLU activation with a dropout of 0.1 as well.
- SciBert model (Base Encoder)
  - Scibert-base-cased model
  - Encoder layers use GeLU activation, with a dropout of 0.1
- Training
  - Original BIO Dataset
  - 5 epochs, with a batch size of 16
  - Initial learning rate of 1e-4 and a validation split ratio of 0.1.



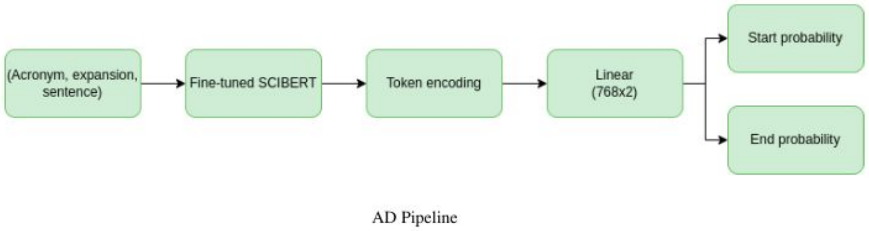
01

02

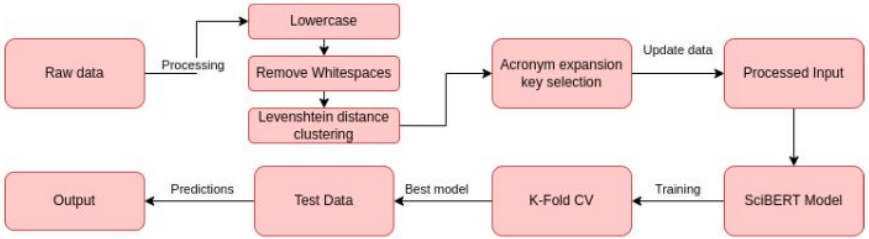
03

04

AD



AD Pipeline



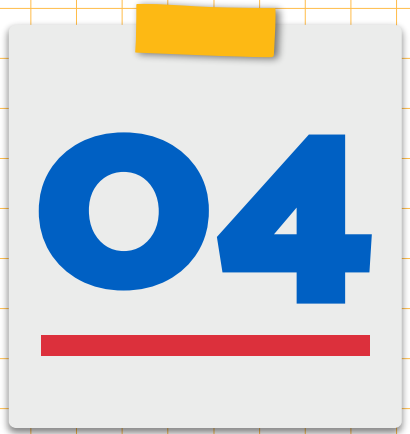
AD Final Architecture

- Data Pre-processing
  - Lowercasing
  - Whitespace removal from expansions
  - Levenshtein distance matrix for all pairs and clustering based on threshold
- SciBert model (Base Encoder)
  - Fine-tuned Scibert-base-uncased model
  - Encoder layers use GeLU activation, with a dropout of 0.1
- Training
  - Processed data
  - 10 epochs, with a batch size of 32
  - Initial learning rate of 2e-5 and a 5 fold cross validation

01 02 03 04

Results & Analysis

Final Statistics and why they are so



AE

Dataset / Model	BERT			SciBert		
(domain_flavor)	F1	P	R	F1	P	R
leg_vanilla	75.63	75.6	75.66	**75.70**	**76.10**	**75.30**
leg_nosym	70.84	71.31	70.37	72.58	73.13	72.04
leg_nosym_plusbrac	73.95	74.62	73.28	74.67	75.41	73.95
leg_nosym_plusbrac_equal	74.27	75.50	73.09	73.93	74.89	72.99
sci_vanilla				**82.74**	**81.58**	**83.94**
sci_nosym				79.35	79.3	79.4
sci_nosym_plusbrac				81.70	80.93	82.49
sci_nosym_plusbrac_equal				82.57	81.20	84.00

01

02

03

04

AD

dataset	model	scibert uncased		
	flavor	F1	P	R
legal	finetuned + data preproc	**89.52**	**90.92**	**88.17**
	data preproc	88.39	91.63	85.38
	finetuned	74.79	83.38	67.81
	vanilla	74.79	83.38	67.81
scientific	finetuned + data preproc	**92.33**	**94.13**	**90.6**
	data preproc	92.33	94.13	90.6
	finetuned	77.95	84.31	72.49
	vanilla	77.95	84.31	72.49

dataset	model	others		
	flavor	F1	P	R
legal	finetuned + data preproc			
	data preproc	88.28	90.81	85.89
	finetuned			
	vanilla	54.33	77.22	41.91
scientific	finetuned + data preproc			
	data preproc	91.25	93.74	88.88
	finetuned			
	vanilla	78.95	86.37	72.7

01

02

03

04

Analysis: Dataset Quality

AE

"United Nations Population Fund", "DP", and "FPA", were marked as O.

AD

"mation", "nition", and "Noun Phras" which are typos or incomplete words.

Noise

A lot of random symbols, repetitions, lower and upper cases...

01

02

03

04

Analysis: AE

Mid Phase

ISSUE	DETAILS
DATA QUALITY	Words like “misc” and “ad hoc” are now valid words. Special Symbols hindered model performance
SYMBOLS	There were sequences that could have an acronym which were marked by model but the data didn't.

01

02

03

04

Analysis: AE, data

End Phase

RESULT	WHAT HAPPENED
BETTER	*"? Abbre"viation features (ABB) : For ... checked whether,"* model couldn't ID *Abbreviation* in long form, but on the edited data, it could.
NEUTRAL	*"3 Experimental Results ... the Hong Kong City University(CityU) corpus ... Segmen -"*, our model marked *City University* as the long form, while expected annotation was only *City*.
WORSE	*"Output (0<x<1) Figure 3 Neural network architecture (**DA =descriptor array** of 20 items)"* our model didn't understand the importance of short forms equating to long ones.

Analysis: AE, models

End Phase

PROS	WHY SCI-BERT WAS BETTER ON LEGAL?
INTUITIVE MAPPING	*"Coordinating Committee for Geoscience Programmes in East and Southeast Asia (CCOP) ... Agency (IAEA)"*, here the BERT model skipped the word *Coordinating*.
ORPHAN SHORTS	*":... the eTIR international system ... Service) : the eTIR international system will ..."*, here the BERT model skipped short tags for *eTIR*.

01

02

03

04

Analysis: AD

Mid Phase

ISSUE	WHAT HAPPENED
TOO VAGUE	words like *"critical"* were shortened , for these model predicted similarly vague, but different words like *"creation"*.
TOO SPECIFIC	pairs like *"Semantic Role"*-*"Semantic Relation"* where both domain specific short forms could fit the context and had very close meanings.
VARIETY	for cases like *"Support Vector Machines"*, prediction was *"Vector Machines"*, doesn't fit the tokens *"S, V, M"*. The data also had variations like plural form etc. of the same long form.

Analysis: AD

End Phase

DATA	NOTES
LEGAL	most were a result of not being grouped by the augmentation Levenshtein process like: *dermal* & *dermal toxicity*. Very few like: *inland transport committee* & *international trace centre*, were actually wrong.
SCIENTIFIC	In contrast to legal, majority were wrong, like: *hierarchical dirichlet tree* & *hindi dependency treebank*, a few like: *longest common sub* & *longest common subsequence*, were the result of Levenshtein misgrouping.

01

02

03

04

Thanks!

01

02

03

04