MILAB at PragTag-2023: Enhancing Cross-Domain Generalization through Data Augmentation with Reduced Uncertainty

Yoonsang Lee[1*], Dongryeol Lee[2*], Kyomin Jung[2, 3]

[1] College of Liberal Studies, Seoul National University, [2] Dept. of ECE, Seoul National University,
[3] ASRI, Seoul National University

EMNLP 2023

Task Definition

- **Pragmatic Tagging of Peer Reviews:** Given a peer review data from **5 distinct domains**, classify each sentence into **6 predefined labels**.
- Three task conditions: ***Full***, ***Low*** (20% data of *Full* condition), and ***Zero*** distinguished by the number of training data.

A very good attempt to present the Indian COVID-19 scenario by the authors. ← Strength
I congratulate them on their work. However a few queries: ← Structure
Recap → - The data analysis has been performed on 1161 patients.
To project it for such a large population has limited scope. ← Weakness
Other → - In COVID, most of the patients recover in due course.
If possible, the SEIR model could have been used for a better picture. ← Todo

Main Challenges

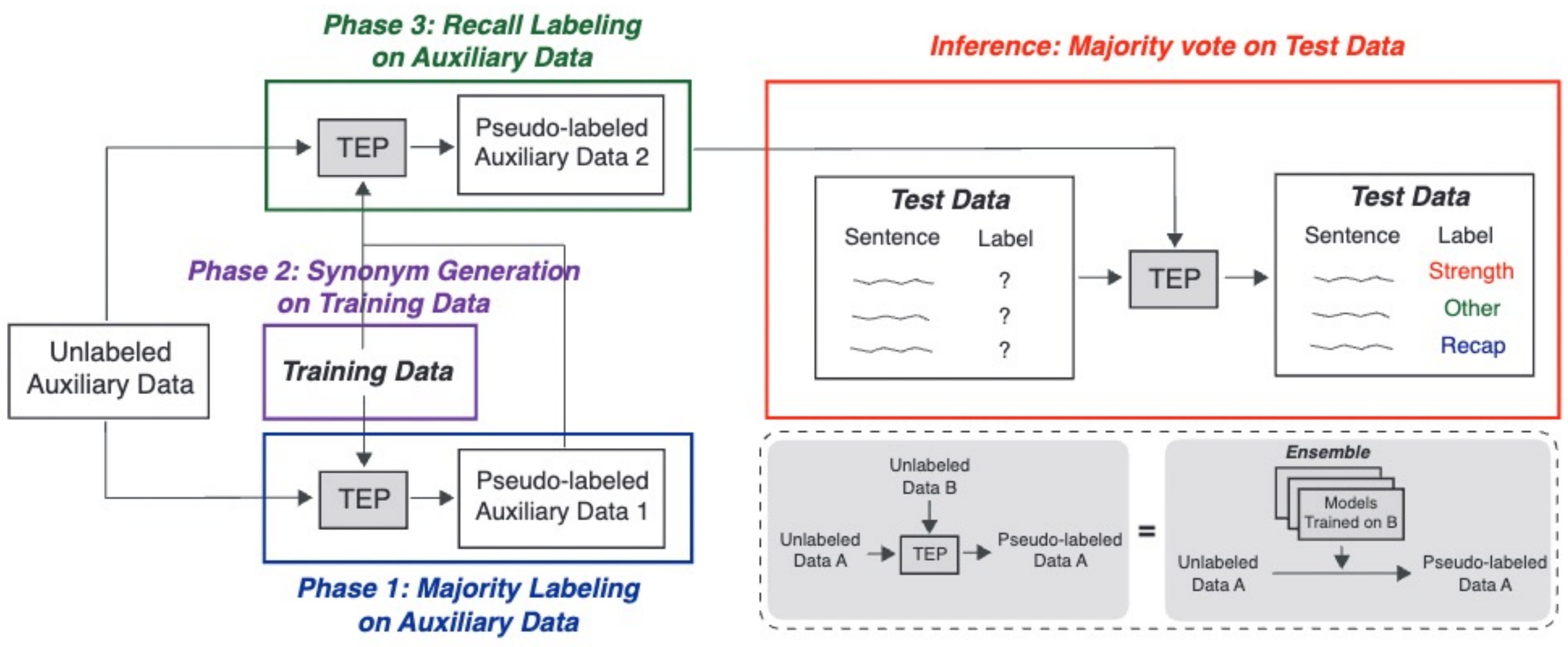
Cross-Domain: The proposed task is designed for a multi-domain scientific corpus, where certain domains may employ specific terminologies or require a unique evaluative perspective.

Low-resource: The distribution of data varies across 1) domains and 2) labels, which introduces a data imbalance problem.

	Full						
Domain	**Strg.**	**Weak.**	**Strc.**	**Rec.**	**Td.**	**Oth.**	**Total**
scip	46	73	70	52	115	105	461
iscb	30	93	53	77	173	70	496
rpkg	67	85	64	69	132	89	506
diso	43	81	61	76	135	79	475
case	34	45	53	72	126	58	388
Total	220	377	301	346	681	401	2326

- 5 domains: science policy research (scip), bioinformatics (iscb), R package (rpkg), disease outbreak (diso), medical case reports (case)
- 6 labels: Strength (Strg.), Weakness (Weak.), Structure (Strc.), Recap (Rec.), Todo (Td.), Other (Oth.).

Method



Our proposed method to handle cross-domain low-resource processing of peer reviews consists of three phases: **(1) Majority Labeling on Auxiliary Data**, **(2) Synonym Generation on Training Data**, and **(3) Recall Labeling on Auxiliary Data**.

(1) Majority Labeling on Auxiliary Data

- Given a labeled training dataset, train five BERT based classifier using different models and hyperparameters.
- Then unlabeled auxiliary dataset is labeled using ensemble of five classifiers. We compare Majority-vote and Consensus methods.

(2) Synonym Generation on Training Data

- Given a labeled training dataset, utilize a synonym generator to generate additional labeled data.
- To secure the quality of augmented dataset, calculate BERTSCORE between original and augmented data, and only sample top-k data.

(3) Recall Labeling on Auxiliary Data

- For each pragmatic tag, select the model with the highest recall.
- Then models label the sentences in descending order of their recall scores.
- After labeling the distinct tags, any residual sentences are designated as "*Other*".

Results

Majority Labeling Model

- We compare majority-vote and Consensus methods with different combination of training data.

	Majority	Consensus
F1000raw	**0.8454**	0.8333
F1000raw+ARR	0.8263	0.8251

Recall Labeling Model

- Recall score of best performing models for each label. We label auxiliary dataset in descending order (Strc. – Td.– Strg.– Rec.- Weak.– Oth.)

Strength	Weakness	Structure
0.936	0.892	1.0
Recap	**Todo**	**Other**
0.928	0.990	0.685

Code available at: https://github.com/lilys012/pragtag

Main results

	f1_mean	f1_case	f1_diso	f1_iscb	f1_rpkg	f1_scip	f1_secret
full	0.839	0.840	0.837	0.801	0.854	0.865	-
low	0.771	0.778	0.746	0.754	0.777	0.800	-
zero	0.516	0.502	0.518	0.551	0.492	0.516	-
final (full)	0.824	0.844	0.840	0.798	0.843	0.864	0.755
final (zero)	0.517	0.502	0.520	0.557	0.508	0.489	0.528

Full & Low

- Test data is labeled in a majority-vote manner using the best-performing models from Phase (3) Recall Labeling.
- We achieved average F1-score of **0.839** and **0.771** in ***Full*** and ***Low*** conditions, respectively, which rank **3rd** for each condition.

Zero

- In addition to auxiliary ARR dataset, we adopt a simple rule-based labeling approach for the "Structure" and "Other" label.
- We achieved an average F1-score of **0.517**, ranking **1st** rank for the zero condition.