University of Stuttgart
Germany

Supervised Semantic Proximity, Noise and Disagreement Detection

January 17, 2025

Tejaswi Choppa
Institute for Natural Language Processing, University of Stuttgart

Introduction

- Gold data is required for training and testing of models
- **Common approach:** Adjudicate multiple annotations into single gold label
- **Problem:** This discards valuable information
- **Aim:** Predict and analyze noise and semantic proximity disagreement

Example of Disagreement

(1) ...and taking a knife from her pocket, she opened a vein in her little **arm**.

(2) ...It stood behind a high brick wall, its back windows overlooking an **arm** of the sea ....

- Sample judgments: [2,3,2]; median: 2; mean disagreement: 0.67; noise label: 0

DURel Annotation Scale

↑	4: Identical
	3: Closely Related
	2: Distantly Related
	1: Unrelated
	- - - - - - - - -
	0: Cannot Decide

Table 1: The DURel relatedness scale (Schlechtweg et al., 2018).

Example of Noise

(1) ...and taking a knife from her pocket, she opened a vein in her little **arm**.

(3) ...the com pany create a new **arm**

- Sample judgments: [1,0,0]; noise label: 1

Table of Contents

Task

- Given the pair of word usages:
  - **OGWiC**: predict median semantic proximity label

$$M(J) = median(J)$$

  - **DisWiC**: predict the mean disagreement

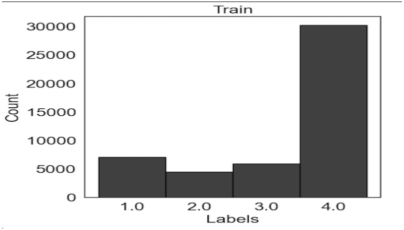
$$D(J) = \frac{1}{|J|} \sum_{(j_1, j_2) \in J} |j_1 - j_2|,$$

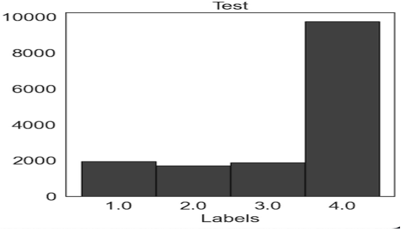
  - **NoiseWiC**: predict noise

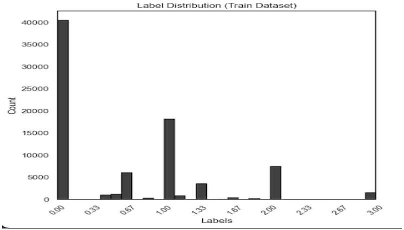
$$N(J) = \begin{cases} 1, & \text{if } (\# \text{ non-zero } < \# \text{ zero}) \\ \text{NaN}, & \text{if } (\# \text{ non-zero} \geq \# \text{ zero})\text{and}(\# \text{ zero} > 0) \\ 0, & \text{otherwise} \end{cases}$$

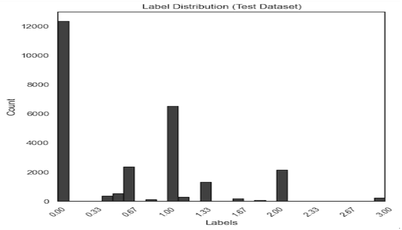
Data

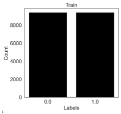
- For all our tasks, we make use of publicly available ordinal WiC datasets from the CoMeDi shared task (Schlechtweg, Choppa, Zhao, & Roth, 2025)
- Datasets are highly skewed having class imbalance
- It is a multi-lingual dataset
- For NoiseWic, we employ a sampling strategy to downsample the majority class to match the size of the minority class

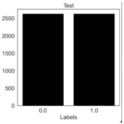

Train
Count
Labels
1.0
2.0
3.0
4.0
0
5000
10000
15000
20000
25000
30000


Test
Labels
1.0
2.0
3.0
4.0
0
2000
4000
6000
8000
10000

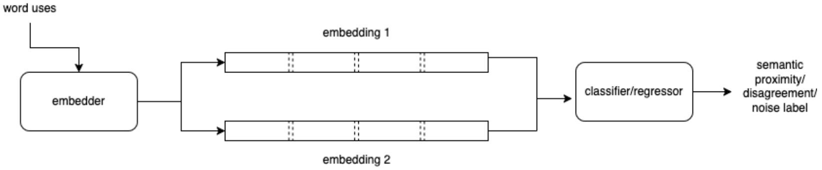

Label Distribution (Train Dataset)
Count
Labels
0.00
0.33
0.67
1.00
1.33
1.67
2.00
2.33
2.67
3.00


Label Distribution (Test Dataset)
Count
Labels
0.00
0.33
0.67
1.00
1.33
1.67
2.00
2.33
2.67
3.00


Train
Count
Labels
0.0
1.0


Test
Labels
0.0
1.0

Model Architecture



Models

- Contextual embedders:
  - **XL-Lexeme**: An extension of S-BERT model pre-trained on WiC datasets
  - **XLM-R (Baseline)**: An extension of RoBERTa using self-supervised training techniques to achieve state-of-the-art performance in cross-lingual understanding

Models

- Heads:
  - **OGWiC**: Cosine+Threshold (CosTH), MLP, Linear Regression
  - **DisWiC**: MLP, Linear Regression
  - **NoiseWiC**: Logistic Regression

Baselines

- Majority Baseline:
  - Used for the NoiseWic task
  - Provides a minimum performance threshold that a model should exceed
- Feature Baseline:
  - For DisWic, feature vectors with character length and non-alpha character ratio
  - Uses MLP to predict disagreement labels

Evaluation

- **OGWiC**: Krippendorff's α
- **DisWic**: Spearman's ρ
- **NoiseWic**: Accuracy and Krippendorff's α

Upperbound Metric

- Represents the maximum potential performance of a model on a specific task
- OGWiC:
  - Compute α iteratively across annotators by excluding one, weighted by their annotation contribution
- DisWiC:
  - Compute ρ iteratively comparing excluded annotator pairs with remaining annotators
  - Requires a minimum of four annotators for analysis

Experiments

- For each of the subtasks, the models are fit on the training data in two ways:
  - **Per language** i.e, hyperparameters or thresholds are learned per language
  - **All Data** i.e, on the entire training data available

Experiments

- For each of the models in **OGWiC** and **DisWic**:
  - We fit models with best parameters by searching over a defined parameter grid

Results-OGWiC

Model	Setting	AVG	ZH	EN	DE	NO	RU	ES	SV
Upperbound	All	**.95**	**1.**	**.97**	**.88**	**.94**	**.96**	**.96**	**.96**
XL-Lexeme + CosTH	Lang	**.58**	**.38**	**.65**	**.72**	**.51**	**.55**	**.65**	**.60**
XL-Lexeme + LR	All	.16	.04	.26	.15	.06	.15	.26	.18
	Lang	.09	.06	.04	.15	.03	.22	.22	-.07
XL-Lexeme + MLP	All	.42	.35	.49	.39	.37	.44	.51	.40
	Lang	.28	.20	.36	.36	.23	.32	.34	.13
XLM-R + CosTH	Lang	.12	.06	.10	.27	.12	.11	.17	.02

Results-DisWiC

Model	Setting	AVG	ZH	EN	DE	NO	RU	ES	SV
Upperbound	All	**.18**		**.07**	.04		**.22**	.08	**.48**
XL-Lexeme+ LR	All	.10	.30	.02	.03	.06	.07	.05	.18
	Lang	.09	.06	.04	**.15**	.03	**.22**	**.22**	-.07
XL-Lexeme+ MLP	All	.15	.45	**.07**	.07	.10	.13	.08	.16
	Lang	**.16**	**.48**	.04	.11	**.25**	.04	.06	.16
XLM-R + LR	All	.11	.38	.06	.09	.07	.04	.07	.08
	Lang	.05	.10	.01	.13	.04	.11	.05	-.11
Feature Baseline	All	-.00	-.00	-.00	.00	-.03	-.01	-.01	.02

Results-NoiseWiC

Metric	Model	AVG	EN	DE	NO	ES	SV
Accuracy	XL-Lex. +Logistic Reg	.58	**.59**	.63	**.58**	.48	.63
Accuracy	XLM-R +Logistic Reg	**.59**	**.59**	**.65**	.47	**.60**	**.63**
Krippendorff	XL-Lex.+Logistic Reg	.15	.19	.27	.15	-.08	.26
Krippendorff	XLM-R+Logistic Reg	.14	.17	**.30**	-.21	.20	.25
Accuracy	Majority Baseline	.50	.50	.50	.50	.50	.50

Exemplary Disagreement Pattern

(1) Willoughby's as the family possess and will submit for examination, carefully searched, in the hope that some **record** may be found in his hand-writing.

(2) For the **record**, your information is inaccurate on Governor Rockefeller's visit on Sept. 21.

- `Judgments`: [3, 4, 2]
- `Mean Disagreement Label`: 1.333

Exemplary Noise Pattern

(3) The public, gene- /z/ **rally**, remained indifferent, notwithstanding the marvellous things which were related of the terri tory which had been ceded to the company.

(4) Once or twice I have known him touch nerves that go close to the heart; but gene **rally**, he is no master of the feelings.

- `Judgments:` [1, 0, 0, 0, 4]
- `Noise Label:` 1

Factors Influencing Annotator Disagreement

- Grammatical errors and misspelled words
- Lack of contextual information
- Complex language misinterpretation
- Annotator uncertainty raising reliability concerns
- Historical contexts, scientifically specific concepts

Conclusion I

- Task Formulation and Model Performance:
  - Introduced OGWiC, DisWiC, and NoiseWiC tasks for semantic proximity and disagreement analysis
  - XL-Lexeme achieved highest Krippendorff's α scores of 0.67 (dev) and 0.58 (test)
  - Consistently outperformed baseline XLM-R, especially in language-specific configurations
  - In DisWiC, ZH and NO perform better

Conclusion II

- Addressed class imbalance in NoiseWiC through a downsampling strategy
- Demonstrated the importance of per-language hyperparameter tuning
- Research can be expanded by looking into main factors affecting the disagreement

References I

Schlechtweg, D., Choppa, T., Zhao, W., & Roth, M. (2025). The CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of the 1st workshop on context and meaning–navigating disagreements in nlp annotations.* Abu Dhabi, UAE.

Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 169–174). New Orleans, Louisiana. Retrieved from `https://www.aclweb.org/anthology/N18-2027/`