CoMeDi Shared Task: Median Judgment Classification & Mean Disagreement Ranking with Ordinal Word-in-Context Judgments

Dominik Schlechtweg[1] **Tejaswi Choppa**[1] **Wei Zhao**[2] **Michael Roth**[3]
[1]University of Stuttgart [2]University of Aberdeen [3]University of Technology Nuremberg
schlecdk@ims.uni-stuttgart.de st180670@stud.uni-stuttgart.de
wei.zhao@abdn.ac.uk michael.roth@utn.de

Abstract

We asked task participants to solve two subtasks given a pair of word usages: Ordinal Graded Word-in-Context Classification (OGWiC) and Disagreement in Word-in-Context Ranking (DisWiC). The tasks take a different view on modeling of word meaning by (i) treating WiC as an ordinal classification task, and (ii) making disagreement the explicit detection aim (instead of removing it). OGWiC is solved with relatively high performance while DisWiC proves to be a challenging task. In both tasks, the dominating model architecture uses independently optimized binary Word-in-Context models.

1 Introduction

Recent developments in language modeling and word embeddings have made it possible to achieve near-human performance in several semantic NLP tasks (Wang et al., 2019). One of these is the Word-in-Context task (WiC, Pilehvar and Camacho-Collados, 2019), asking if the same word in two contexts has the same meaning. WiC treats the problem of meaning distinctions as a **binary classification task**. The state-of-art model has obtained near-human performance (77.9% vs. 80%, Wang et al., 2021). On the one hand, WiC is an elegant simplification of the classical Word Sense Disambiguation task (Navigli, 2009) avoiding the need for sense glosses and opening new avenues for auxiliary tasks such as Word Sense Induction (WSI Schütze, 1998) or Lexical Semantic Change Detection (LSCD, Schlechtweg et al., 2020). On the other hand, the binary nature of the task is a strong and inadequate simplification of the problem of word meaning distinction (Tuggy, 1993; Cruse, 1995; Kilgarriff, 1997; Erk et al., 2013; McCarthy et al., 2016). A more theory-adequate formulation is the Graded Word Similarity in Context task (GWiC, Armendariz et al., 2020). It asks to provide graded WiC predictions. However, the GWiC shared task did not require models to **reproduce** human annotations as the evaluation metric (harmonic mean of Pearson and Spearman correlations) does not restrict the label set in the predictions, effectively treating the problem as a **ranking task**. Such a task can be fulfilled by predictions on an arbitrary scale (e.g. real numbers). However, exactly reproducing human annotations can have certain advantages such as providing linguistic interpretations. These can be exploited for modeling auxiliary tasks such as WSI or LSCD where linguistic interpretations such as *context variance* or *polysemy* can be crucial to decide whether a new sense was found. Hence, we introduce Ordinal Graded Word-in-Context Classification (OGWiC), asking participants to exactly reproduce instance labels instead of just inferring their relative order.

WiC Datasets annotated on ordinal scales often show considerable disagreement. Consequently, we lose information when discarding instances during aggregation or summarizing them by majority judgment. Recent research has started to incorporate this information by using alternative label aggregation methods (Uma et al., 2022; Leonardelli et al., 2023). Modeling this disagreement is important because in a real world scenario we most often do not have clean data. We need to predict on samples where high disagreement is expected and which are inherently difficult to categorize. Predicting disagreement can help to detect or filter highly complicated samples. Therefore, we introduce the task of Disagreement in Word-in-Context Ranking (DisWiC). It differs from previous tasks (Leonardelli et al., 2023) by aggregating "gold" labels purely over judgment differences, thus making disagreement the explicit ranking aim.

Both tasks, OGWiC and DisWiC, were introduced in a shared task organized as part of the 2025 CoMeDi workshop.[1] This paper describes

[1] https://comedinlp.github.io/

Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation, pages 33–47
January 19, 2025.

4: Identical	Identity
3: Closely Related	Context Variance
2: Distantly Related	Polysemy
1: Unrelated	Homonymy

Table 1: The DURel relatedness scale (Schlechtweg et al., 2018) on the left and its interpretation from Schlechtweg (2023, p. 33) on the right.

the setup, participating systems and results of the shared task.

2 Related work

2.1 Word-in-Context task

The Word-in-Context task (WiC, Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020; Martelli et al., 2021) is a relatively new task reframing Word Sense Desambiguation in a context-only setting. It asks if the same word in two contexts/usages has the same meaning. WiC treats the problem of meaning distinctions as a binary classification task. The state-of-the-art model has obtained near-to-human performance on English data (78% vs. 80% accuracy, Wang et al., 2021). A more theory-adequate formulation is the Graded Word Similarity in Context task (GWiC, Armendariz et al., 2020). It asks to provide graded WiC predictions on an arbitrary scale, treating the problem of meaning distinctions as a ranking task. The state-of-the-art model reaches near-to-human performance on English data (73% vs. 77% harmonic mean of the Spearman and Pearson, Al-khdour et al., 2020).

Recently, a number of WiC-like datasets have been annotated with semantic proximity labels on an ordinal scale from 1 (the two uses of the word have completely unrelated meanings) to 4 (the two uses of the word have identical meanings) following the four-point scale in Table 1 (e.g. Schlechtweg et al., 2021; Kurtyigit et al., 2021; Kutuzov and Pivovarova, 2021b; Chen et al., 2023).[2] This scale was developed within the DURel annotation framework (Schlechtweg et al., 2018), which is based on Blank's concept of semantic proximity (Blank, 1997, pp. 413–418)). This ordinal scale is similar to the one used for the original annotations in GWiC (before data transformation).

Each level of the DURel scale has an exact linguistic interpretation as depicted in Table 1, where **polysemy** is located between **identity**, **context variance**, and **homonymy** (Schlechtweg, 2023, pp. 22–23). The pair (1,2) is classified as **identical** as the referents of two uses of the word *arm* are both prototypical representatives of the same extensional category corresponding to the concept 'a human body part':

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

(2) [. . .] and though he saw her within reach of his **arm**, [. . .]

The usage pair (1,3) is classified as **context variance** as both referents still belong to the same extensional category, but one is a non-prototypical representative. Hence, there is some variation in meaning, e.g. the arm of a statue loses the function of the physical arm to be lifted:

(3) [. . .] when the disembodied **arm** of the Statue of Liberty jets spectacularly out of the sandy beach.

The usage pair (1,4) would be classified as **polysemy** as the two referents of *arm* belong to different extensional categories, but the corresponding concepts still hold a semantic relation (in this case a similarity relation regarding physical form).

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

In contrast, the referents of *arm* in the **homonymic** pair (1,5) belong to different extensional categories and the corresponding concepts do *not* hold a semantic relation:

(5) And those who remained at home had been heavily taxed to pay for the **arms**, ammunition; fortifications, [. . .]

2.2 Disagreement detection

Most data for NLP tasks is created by discarding disagreement. However, some approaches try to incorporate disagreement into the task through alternative label aggregation methods. One of the oldest approaches, as suggested by Dawid and Skene (1979), is the probabilistic label aggregation method. This method calculates the posterior probability of a label for a particular instance conditioned on predicted label, true label and reliability of the annotator, i.e., the annotator's past annotations. The final label is chosen based on the

[2] There are further ordinal datasets annotated on different scales (e.g. Trott and Bergen, 2021).

posterior probability. While this method incorporates disagreement for choosing gold labels, it still reduces the data down to a single dominant view. Sheng et al. (2008) modify this approach proposing an uncertainty-preserving labeling scheme that retains information about annotator disagreement instead of resolving it. They represent labels as a probability distribution over classes based on annotator ratings (“soft labels”). This preserves ambiguity and uncertainty when multiple plausible labels exist. Aligning with these approaches, Uma et al. (2021) develop machine learning models that can effectively learn from and capture the disagreement of annotations, rather than just relying solely on a single aggregated label. To learn from the full distribution of annotations, the annotator distributions are converted into soft labels and the model is optimized to predict these soft label distributions (Uma et al., 2021). They employ techniques like standard normalization of annotator distributions, softmax function over annotator distributions and use of probabilistic label aggregation models like MACE to generate soft labels.

Although these approaches capture the distribution of disagreeing annotations, there is no significant research on directly predicting the **amount** of disagreement in a supervised way.

3 Tasks

Participants are asked to solve two subtasks. Both rely on data from human WiC judgments on the ordinal DURel scale, as described in Section 2.1. Each instance has a target word w, for which two word usages, u_1 and u_2, are provided (usage pair). Each of these usages expresses a specific meaning of w. As an example, consider the two annotation instances below. Pair (1,2) would likely receive label 4 (identical) while pair (1,3) would rather receive a lower label such as 2 (distantly related).

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

(2) ...and though he saw her within reach of his **arm**, yet the light of her eyes seemed as far off.

- Sample judgments: [4,4]; median: 4; mean pairwise difference: 0.0

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

(3) It stood behind a high brick wall, its back windows overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat.

- Sample judgments: [2,3,2]; median: 2; mean pairwise difference: 0.667

3.1 Subtask 1: Median Judgment Classification with Ordinal Word-in-Context Judgments (OGWiC)

For each usage pair (u_1, u_2), participants are asked to predict the median of annotator judgments.[3] This task is similar to the previous WiC and GWiC tasks. However, we limit the label set in predictions and penalize stronger deviations from the true label (see Section 6). This makes OGWiC an **ordinal classification task** (Sakai, 2021), in contrast to binary classification (WiC) or ranking (GWiC). Predictions are evaluated against the median labels with the ordinal version of Krippendorff’s α (Krippendorff, 2018).

Treating graded WiC as an ordinal classification task instead of a ranking task constrains model predictions to exactly reproduce instance labels instead of just inferring their relative order. This is advantageous if ordinal labels have an interpretation because predictions then inherit this interpretation. Such an interpretation can be assigned to the DURel scale as explained in Section 2.1: Judgment 1-4 can be interpreted as "homonymy" (1), "polysemy" (2), "context variance" (3) and "identity" (4), respectively.

3.2 Subtask 2: Mean Disagreement Ranking with Ordinal Word-in-Context Judgments (DisWiC)

For each usage pair (u_1, u_2), participants are asked to predict the mean of pairwise absolute judgment differences between annotators:

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

where J is the set of unique pairwise combinations of judgments. For pair (1,2) from above,

$$D(J) = \frac{1}{2}(|(4-4)| + |(4-4)|) = 0.0$$

while for (1,3) it amounts to

$$\frac{1}{3}(|(2-3)| + |(2-2)| + |(3-2)|) = 0.667.$$

[3] We choose the median instead of other summary statistics because it is robust to outliers and frequently used in studies using ordinal WiC data (e.g. Schlechtweg et al., 2020; Zamora-Reina et al., 2022).

Dataset	LG	Reference	JUD	VER	KRI	SPR
ChiWUG	ZH	Chen et al. (2023)	61k	1.0.0	.60	.69
DWUG	EN	Schlechtweg et al. (2021)	69K	3.0.0	.63	.55
DWUG Res.	EN	Schlechtweg et al. (2024)	7K	1.0.0	.56	.59
DWUG	DE	Schlechtweg et al. (2021)	63K	3.0.0	.67	.61
DWUG Res.	DE	Schlechtweg et al. (2024)	10K	1.0.0	.59	.7
DiscoWUG	DE	Kurtyigit et al. (2021)	28K	2.0.0	.59	.57
RefWUG	DE	Schlechtweg (2023)	4k	1.1.0	.67	.7
DURel	DE	Schlechtweg et al. (2018)	6k	3.0.0	.54	.59
SURel	DE	Hätty et al. (2019)	5k	3.0.0	.83	.84
NorDiaChange	NO	Kutuzov et al. (2022)	19k	1.0.0	.71	.74
RuSemShift	RU	Rodina and Kutuzov (2020)	8k	1.0.0	.52	.53
RuShiftEval	RU	Kutuzov and Pivovarova (2021a)	30k	1.0.0	.56	.55
RuDSI	RU	Aksenova et al. (2022)	6k	1.0.0	.41	.56
DWUG	ES	Zamora-Reina et al. (2022)	62k	4.0.1	.53	.57
DWUG	SV	Schlechtweg et al. (2021)	55K	3.0.0	.67	.62
DWUG Res.	SV	Schlechtweg et al. (2024)	16K	1.0.0	.56	.65

Table 2: Datasets used for our task. All are annotated on the DURel scale. Spearman and Krippendorff values for RuShiftEval are calculated as average across all time bins. 'LG' = Language; 'JUD' = Number of judgments; 'VER' = Dataset version; 'KRI' = Krippendorff's α; 'SPR' = Weighted mean of pairwise Spearman correlations; 'Res.' = Resampled.

DisWiC can be seen as a **ranking task**. Participants are asked to rank instances according to the magnitude of disagreement observed between annotators. It differs from previous tasks (Leonardelli et al., 2023) by aggregating "gold" labels purely over judgment differences, thus making disagreement the explicit ranking aim. Predictions will be evaluated against the mean disagreement labels with Spearman's ρ (Spearman, 1904).

4 Data

For both subtasks, we make use of publicly available ordinal WiC datasets from multiple languages, as summarized in Table 2.[4] These provide a large number of judgments for usage pairs on the DURel scale and have so far not been used primarily for WiC-like tasks, but only for semantic change detection purposes. We additionally augment DWUG DE/EN/SV and DiscoWUG with roughly 33k unpublished instances which we have recently annotated guaranteeing evaluation on hidden data (DWUG Resampled). For all datasets, overall agreement as well as cleaning procedures ensure data quality.

Language	Mean	Std
Chinese	2.00	0.00
English	2.32	0.62
German	2.82	1.06
Norwegian	2.31	0.46
Russian	3.78	1.03
Spanish	2.23	0.48
Swedish	2.36	0.63

Table 3: Mean and standard deviation for number of annotators per instance after cleaning and aggregation per language.

4.1 Dataset pre-cleaning

The data setswe rely on show various problems such as erroneous target word indices or duplicate contexts and judgments. This holds in particular for the Norwegian, Russian and Spanish datasets. Hence, we apply multiple cleaning measures. We describe them in the order they were applied: First, we load all uses from all datasets into one Pandas DataFrame, similarly for judgments, resulting in 82,286 uses and 492,796 judgments to process. Usage pairs with the same use identifiers are considered to be the same pair irrespective of the identifier order in the pair. We start by removing all judgments by annotator 'gecsa' from the Spanish judgments as the annotators was also excluded by the creators of the dataset. Then we drop missing judgments (empty fields). Spanish usages have non-consistent CSV quoting characters. Hence, we drop enclosing quotes and double quotes from contexts while updating target word indices accordingly. Next, we drop duplicate uses if they have the same identifier, context and target word indices.

Then, we reconstruct erroneous target word indices. We start out by excluding punctuation at the beginning or end of the target word; we then check a number of properties on the target word indices and the selected substring to find erroneous indices:

- the start and end index should be in the range of the context length,
- the target word should have at least one character,
- the preceding and following character should not be alphabetic (except in Chinese) and

[4] https://www.ims.uni-stuttgart.de/data/wugs

- the string similarity between target lemma and selected target word string should be above or equal to 0.5.[5]

All usages not meeting any of these conditions are further considered for index reconstruction. Usages with modified punctuation (see above) also enter the reconstruction. The index reconstruction proceeds as follows: We tokenize the context by splitting at white spaces. We then compare the lowercased version of each token with punctuation removed to the lower-cased version of the lemma. For each candidate token with the maximum string similarity, we first remove punctuation from beginning and end and then search for the candidate string with start index nearest to the original index. This candidate is chosen as the new target substring. For cases with multiple candidates with the same distance between new and original start index, we choose the first candidate. For Russian, we additionally split tokens at hyphens as the data contains many compounds connected by hyphens.[6] The finally chosen candidate is regarded as the new target word substring and we extract its start and end index. In order to make sure that the new target substring choice is reasonable, we check its string similarity with the target lemma as described above. Substrings with a string similarity below or equal 0.5 are filtered out and later removed.[7] We manually inspect filtered-out usages below different thresholds of string similarity to make sure not to filter out valid usages not meeting our conditions. This frequently happens where target lemma and substring were very different because of strong inflection, or plural forms with different lemma than the singular forms, such as лю́ди as plural of человек. This leads to a number of additional special conditions making sure to keep certain particular usages or usages meeting certain conditions.

Next, we find usages having the same context, lemma and target word indices (but not identifier, as checked above). For each such equivalence set, one identifier is chosen to represent all of them and used to replace the other identifiers in the judgments. The rest is dropped from the uses. We further aggregate duplicate judgments (same pair judged multiple times by the same annotator) with the median of judgments or as 0 (special judgment for “Cannot decide”) if the number of 0-judgments was larger than judgments between 1–4. Finally, judgments are removed if they contain an identifier that is not present in the uses. After applying this preprocessing, we are left with 80,514 uses and 490,696 judgments.

4.2 Data aggregation and cleaning

For cleaning and aggregation, we initially exclude annotation instances with less than two annotations. For OGWiC, then instances with any 0-judgments (“Cannot decide”) and instances with any pair of annotators disagreeing more than one point on the annotation scale are discarded. We then calculate the median of all judgments, for each instance. Instances with a non-integer median (e.g. 3.5) are discarded. For all remaining instances, gold labels are given by the median of judgments as explained in Section 3.1. For DisWiC, we derive instance labels by aggregating over judgments with the average of pairwise absolute annotator deviations as explained in Section 3.2. 0-judgments are ignored in this process.

For each subtask, we then randomly split the target words per language into train/test/dev with sizes of 70/20/10%. Instances are then assigned to each split according to their lemma. Note that there is no overlap in uses between splits and no overlap in target lemmas. In contrast to previous tasks, we intentionally do not balance out the label distribution in order to keep more realistic data conditions. Find an overview of the final splits per language in Table 4.[8] Find plots of the aggregated label distributions for both subtasks in Appendix A. Table 3 gives additional statistics regarding the number of annotators per language after cleaning and aggregation.

5 Models

Five teams participated in at least one of the shared task’s official evaluation phases. Additionally, there were three unofficial submissions (Choppa et al., 2025; Loke et al., 2025; Sarumi et al., 2025).[9] In the description below, for each team we mark those modeling approaches with the label “top” which produce the team’s top result in the evaluation phase, as reported in Table 5.

[5] We use the ratio measure from the difflib library, ranging between 0.0 and 1.0.

[6] Compounds are only split for index reconstruction. The original context is left untouched.

[7] For compounds, we choose the maximum string similarity of any subtoken after splitting at hyphens.

[8] Data is available through our website: https://comedinlp.github.io/.

[9] User “sunfz1” did not submit a system description paper.

Task	# Instances	# Uses	# Lemmas	Split
OGWiC	48K	55K	520	Train
	8K	8K	77	Dev
	15K	16K	152	Test
DisWiC	82K	55K	521	Train
	13K	8K	77	Dev
	26K	16K	152	Test

Table 4: Data statistics after cleaning and aggregation per split and and over all languages combined.

5.1 Participating teams

Deep-change (Kuklin and Arefyev, 2025) The employed model is based on a previous Word-in-Context model for binary prediction (same sense vs. different sense), which has already shown high performance in lexical semantic change detection (DeepMistake, Arefyev et al., 2021; Homskiy and Arefyev, 2022). The model uses XLM-R, a multilingual BERT variation (Devlin et al., 2019; Conneau et al., 2019), as base embeddings, which were fine-tuned on binary multilingual WiC data (Martelli et al., 2021) and/or binary or binarized Spanish data (Pasini et al., 2021; Zamora-Reina et al., 2022). For OGWiC, the model is further fine-tuned on the shared task data or a binarized version of it thresholding the binary predictions to map them to four classes (top). The team also experiments with different models per language (top). For DisWiC, multiple disagreement measures are tested including linear regression directly predicting the disagreement labels, binary classification of perfect agreement and the class probability standard deviation of a four-class model trained on individual annotations (top).

GRASP (Alfter and Appelgren, 2025) For OGWiC, multiple models are tested including a probabilistic sequential model predicting annotation sequences from annotation co-occurrence frequencies, a simple XLM-R-based Word-in-Context model fine-tuned on the task data and an XLM-R-based Word-in-Context model (XL-Lexeme, Cassotti et al., 2023) previously fine-tuned on binary multilingual WiC data (Martelli et al., 2021) with thresholds on cosine similarity (top). For DisWiC, the team tests regression models using cosine similarities from XL-Lexeme and XLM-R, as well as linguistic features. Further, Word-in-Context models are optimized on different dataset splits representing individual annotators and models are optimized specifically for subsets of languages (top).

MMLabUIT (Le and Van, 2025) Predictions were only submitted for OGWiC. One set of models uses variations of BERT including XLM-R as base embeddings, applies stacking and averaging modifications and measures the final labels by thresholds on cosine similarity. Another set relies on BERT variations (top) and BART (Lewis et al., 2019) as base embeddings, fine-tuning these on Natural Language Inference data, based on a postulated similarity of the shared task objective with Natural Language Inference.

JuniperLiu (Liu et al., 2025) The OGWiC models build on BERT variations including XLM-R (top) and Llama (Touvron et al., 2023) to extract embeddings, apply matrix transformations to remove vector anisotropy, then calculate cosine similarity, and map these to discrete labels using thresholds on the similarity values. For DisWiC, a multi-layer perceptron regressor (Hinton, 1990) is learned on embedding features predicting the disagreement label (top).

FuocChu_VIP123 (Chu, 2025) Only DisWiC predictions are submitted. The model uses Sentence Transformers (Reimers and Gurevych, 2019) based on XLM-R to generate embeddings and a multi-layer perceptron regressor to predict disagreement labels (top).

5.2 Baselines

We employ a number of baseline models to put participants' results into context. Code for Baseline 1 and 3 was published at the beginning of the respective development phases of the shared task.

XLM-R + CosTH (Baseline 1) For each usage pair, we use XLM-R to generate contextualized embeddings for the two target words in context and calculate the cosine similarity (Salton and McGill, 1983) between the two embeddings. Similarity is mapped to discrete OGWiC labels using three thresholds θ. These are optimized on the training set by minimizing the following loss function (cf. Choppa, 2024):

$$L(\mathbf{y}, \hat{\mathbf{y}}|\theta) = 1 - \alpha(\mathbf{y}, \hat{\mathbf{y}}_{\theta})$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are gold labels and predicted cosine similarities respectively, α is Krippendorff's

α and $\hat{\mathbf{y}}_\theta$ is a mapping of cosine similarities to discrete labels according to thresholds θ. We optimize thresholds per language.

XL-Lexeme + CosTH (Baseline 2) This is the same model as XLM-R + CosTH with the exception of using XL-Lexeme (Cassotti et al., 2023) as contextual embedder. XL-Lexeme is a bi-encoder model utilizing a Siamese Network that extends the Sentence Transformers (Reimers and Gurevych, 2019) architecture to focus on the target word within input sentences. The model is trained using a contrastive loss function, which minimizes the cosine distance between the encoded representations when the target word has the same meaning and maximizes the distance when the meanings differ. It is pre-trained on a large multilingual binary WiC dataset (Martelli et al., 2021). We learn one mapping from similarities to thresholds per language.

XLM-R + LR (Baseline 3) For each usage pair, we use XLM-R to generate contextualized embeddings for the two target words in context and concatenate these to create a single representation. We then use linear regression to learn a mapping from embedding representations to continuous disagreement labels for DisWiC. This mapping is optimized on the training set. We learn one mapping per language, and one on the full dataset. Then, we choose the condition which yields highest performance on the development set to apply to the test set. The optimized condition is given by the full dataset model.

XL-Lexeme + MLP (Baseline 4) For each usage pair, we use XL-Lexeme to generate contextualized embeddings for the two target words in context and concatenate these to create a single representation. We then use a multi-layer preceptron regressor (Hinton, 1990) to learn a mapping from embedding representations to continuous disagreement labels for DisWiC. This mapping is optimized on the training set. We learn one mapping per language, and one on the full dataset. We further vary the batch size, activation function, hidden layer size and alpha parameters, and apply feature scaling. Refer to Table 6 in Appendix B for an overview of the hyperparameter grid used. Then, we choose the combination which yields highest performance on the development set to apply to the test set. The optimized condition is given by the per language model with hyperparameters as shown in Table 7 in Appendix B.

Upper bound (OGWiC) For each language, we iterate over annotators and calculate Krippendorff's α between the current annotator's judgments and the remaining ones aggregating them by their median per instance. This number reflects how well each annotator can predict the median of the other annotators' judgments. We then take the average α over annotators weighted by their number of judgments as the final upper bound.[10]

6 Evaluation

WiC is a binary classification task suggesting accuracy as evaluation measure. In contrast, the GWiC shared task used the harmonic mean of Pearson and Spearman correlations (Spearman, 1904). For our OGWiC task, we want to produce ordinal classifications corresponding to the nature of our gold labels. This requirement makes the evaluation measure employed in GWiC unsuitable because it does not limit the label set. Using accuracy is also not ideal in that it does not capture the ordinal nature of the classes. For example, suppose that an instance has a gold label of 4. A model prediction of 1 should be penalized more heavily than a prediction of 3.

With the above considerations in mind, we will use Krippendorff's α (Krippendorff, 2018), which, in its ordinal formulation, penalizes stronger deviations from the gold label more heavily. It has the additional advantage of controlling for expected disagreement and has been demonstrated to be superior to other measures such as Mean Absolute Error for ordinal classification (Sakai, 2021).

For DisWiC, we do not ask participants to reproduce the exact disagreement label as it has no direct interpretation. We are more interested in the relative amount of disagreement observed between usages. Hence, it is formulated as a ranking task and accordingly evaluated with Spearman's rank order correlation coefficient (Spearman, 1904).

Participants were provided with a starting kit implementing our XLM-R-based baseline models

[10] Surprisingly, this upper bound is 1.0 for Chinese. This is a consequence of our cleaning process combined with the specific properties of this dataset: All instances in the dataset have exactly two annotations. As described in Section 4, we remove those with a disagreement of more than one point on the scale. This means that remaining instances with disagreement all have exactly one point disagreement, such as [3, 4]. These instances all have a non-integer median, which is also removed by our cleaning process. Hence, all instances in the cleaned Chinese dataset have perfect agreement.

Task	Team	AV	-ES	ZH	EN	DE	NO	RU	ES	SV
OGWiC	Upper bound	.95	.95	1.	.97	.88	.94	.96	.96	.95
	deep-change	**.66**	**.64**	**.42**	**.73**	.72	**.67**	**.62**	**.75**	**.68**
	Baseline 2	.58	.57	.38	.65	**.73**	.52	.55	.66	.60
	GRASP	.56	.54	.32	.56	.66	.59	.49	.64	.65
	MMLabUIT	.52	.51	.36	.57	.67	.44	.42	.60	.61
	JuniperLiu	.27	.26	.14	.51	.49	.08	.13	.33	.22
	Baseline 1	.12	.12	.06	.10	.27	.12	.11	.18	.02
DisWiC	deep-change	**.23**	.23	.30	**.08**	**.20**	**.29**	**.18**	**.19**	**.35**
	GRASP	.22	.23	**.54**	.04	.11	.27	.17	.12	.30
	Baseline 4	.16	.17	.49	.06	.09	.24	.12	.08	.08
	FuocChu.	.12	.14	.36	.02	.10	.16	.05	.01	.17
	Baseline 3	.12	.12	.39	.06	.09	.08	.05	.08	.08
	JuniperLiu	.08	.09	.36	.04	.02	-.04	.07	.04	.09
	sunfz1	.07	.07	.30	.05	-.00	-.07	.07	.04	.09

Table 5: Top results of evaluation phases. ‘AV’ = Average over languages; ‘-ES’ = Average over languages excluding Spanish; ‘FuocChu.’ = FuocChu_VIP123.

(see Section 5) as well as training and development data (see Section 4) during the development phases for both subtasks lasting from August 23 to September 14 and September 15 to October 13, respectively.[11] During the evaluation phases, which lasted October 14–21 and October 21–27 respectively, participants were allowed to make three submissions, which were evaluated on the hidden test data, where the leaderboard on Codalab was kept hidden at all times.[12] Public test instances were only published at the start of the evaluation phases. Task results were released on October 28. The hidden gold labels of test instances were published during the respective post-evaluation phases.

7 Results

Find an overview of participants' top results in both evaluation phases in Table 5 and results for all submitted predictions in Table 8 in Appendix C. OGWiC is solved with rather high performances across the board. The winning team **deep-change** has an average performance of .66 with minimum of .42 on Chinese and a maximum of .75 on Spanish. The team has top performance on all languages except for German where our Baseline 2 excels. Second and third winners are **GRASP** and **MMLabUIT** with average performances of .56 and .52. The overall maximum performance reached in any language is .75 on the Spanish dataset while the lowest maximum performance for any language is Chinese where no team reached a higher performance than .42. Baseline 1 is outperformed by all participants while Baseline 2 is only outperformed by the winner. The nearest any performance gets to the upper bound is for German with a .15 difference for Baseline 2.

In contrast, DisWiC is solved with rather low performance, turning out to be a very challenging task. The winning team **deep-change** has an average performance of .23 with a minimum of .08 on English and a maximum of .30 on Chinese. The team has top performance on all languages except for Chinese, where **GRASP** excels with .54. Second and third winners are **GRASP** and **FuocChu_VIP123** with average performances of .22 and .12. The overall maximum performance reached in any language is .54 on the Chinese dataset, which is generally solved with rather high performances, while the lowest maximum performance for any language is English where no team reached a higher performance than .08. We hypothesize that maximum performance differences between languages may be related to different numbers of annotators on annotation instances per language, and the effect this has on our disagreement measure defined in Section 3.2, see the discussion in Section 9. Baseline 3 is outperformed by the top three participants while Baseline 4 is only outperformed by the top two participants.

In the post-evaluation phase we noticed that the winning team **deep-change** had (unknowingly) used some of the previously publicly available Spanish test data for training some of their models. This data leakage may have contributed to the exceptionally high result of the team on Spanish. Hence, we also report the average performance across languages excluding Spanish in Table 5 (column ‘-ES’). As we see, this does not change the average performances significantly, whereas **GRASP** now performs slightly better than **deep-change** in DisWiC (.235 vs. .231). However, this is mainly due to the exceptional performance on Chinese.

In both tasks, those teams excel that use independently optimized binary Word-in-Context models, i.e., **deep-change** and **GRASP**. This fits well with the strong performance of our Baselines 2 and 4 building on the same type of model. This could be explained by the similarity of the binary WiC task to OGWiC and the derivation of DisWiC labels from absolute differences between ordinal WiC an-

[11] Starting kits are available through our website: https://comedinlp.github.io/.

[12] Evaluation phase 1 was extended by one day because of technical problems.

notations. Further, across top-scoring submissions, OGWiC is solved by thresholding graded similarity predictions, as in our Baseline 2.

8 Conclusion

We introduced two new tasks based on ordinal Word-in-Context annotations between word usages, and described the results of a shared task based on these: OGWiC asks to predict the median semantic proximity judgment label for each annotated instance. This is a more traditional task definition where data is cleaned and summarized beforehand. DisWiC instead asks to predict the mean of pairwise absolute deviations between annotators. This takes a new and more perspectivist view on data, yet differing from previous tasks in making disagreement the explicit prediction aim. The traditional task was solved with rather high performances while the new approach proves to be challenging. However, on some languages performance is exceptionally high suggesting future modeling possibilities. Both tasks were dominated by the same teams employing a Word-in-Context model optimized on independent binary Word-in-Context data. The dominant approach to solve OGWiC was thresholding of graded similarity predictions.

In the future, it would be interesting to solve the two tasks we introduced with different data splitting conditions, such as sharing target words across splits. Models presumably better generalize from training data with the same target words as in the test data. It would also be interesting to tie the published task data to individual annotators enabling participants to build models for individual annotators accounting for individual judgments and corresponding disagreements.

9 Limitations

As a result of different numbers of annotators per instance, mean absolute disagreement values may not be completely comparable across instances. Consider this example: If an instance has two annotations, the maximum possible mean pairwise disagreement is 3.0, e.g. for

$$D(\{1,4\}) = \frac{1}{1}(|(4-1)|) = 3.0.$$

If one adds one more annotation, the maximum possible disagreement reduces to 2.0, e.g. for

$$D(\{1,1,4\}) = \frac{1}{3}(|(1-1)| + |(1-4)| + |(1-4)|) = 2.0.$$

This means that our measure is influenced by the number of annotators, which was not available to participants at test time. There is considerable variation across languages in the annotator number per instance: Table 3 gives the mean and standard deviation for the number of annotators per instance for each language. Chinese is exceptional with a mean of 2.0 and a standard deviation of 0.0, which means that each instance is annotated by exactly two annotators. As the number of annotators is constant across instances in Chinese, the mean disagreement values are not influenced by annotator number, facilitating prediction for participants, as opposed to the other languages. This may have supported exceptionally high DisWiC results for Chinese, see Table 5. In the future, the number of annotations per instance should be controlled or provided a test time, or the measure should be normalized. Also, other disagreement measures should be explored.

One of the major shortcomings of our setup is the narrowness of training, development and test data in terms of target words. While the task used data for seven languages with tens of thousands of usage pair instances per language, these instances were only sampled from a few hundred target words. The data was additionally split at target words (lexical split), asking participants to generalize from a huge number of instances of few target words to instances of unseen target words. It is questionable whether the training data provides enough information to generalize to unseen target words, and overfitting on the narrow training data is likely. Some task results indicate that models not using the training data at all perform competitively (Kuklin and Arefyev, 2025). In the future, one could run the tasks with alternative data splits where training, development and test data would not be split at target words, but at usages, asking models to generalize to usages from the same target words in the test data as seen in the training data. This would be a relevant task setup as in many annotation studies the budget allows to annotate a limited number of instances per word. If a model can be learned from these instances to reasonably predict the labels for unseen instances, this would be of enormous practical usefulness to analyze greater samples.

Another limitation is related to our choice of Krippendorff's α as evaluation measure for OGWiC. Despite its advantages and being recommended by Sakai (2021) for ordinal classification, the measure estimates the expected label distribution from both model and gold labels, which seems a reasonable assumption when measuring annotator agreement where none of the annotators should be given prevalence, but seems less reasonable in a model evaluation setup where the expected label distribution is given by the gold labels. In the future, one could explore modifications of Krippendorff's α estimating the expected label distribution solely from the gold data.

The performance upper bound we calculated for OGWiC may be influenced positively by our data cleaning process: While all left-over instances after the cleaning have high agreement, it may have occurred randomly for some of them, i.e., even two random annotators would agree on some instances, but this would not make their annotations for those instances reliable. Such instances will push the upper bound, but will be impossible to model.

Almost all of the datasets we used have a diachronic component, i.e., usages sampled from historical time periods often containing historical spelling variants and outdated meanings. While we completely ignored this component in this task, it puts additional difficulties on models and may be responsible for some of the performance differences observed between languages. In the future, one could include this information into the task setup by reporting results for historical and modern language instances separately.

Acknowledgments

Dominik Schlechtweg has been funded by the research program 'Change is Key!' supported by Riksbankens Jubileumsfond (under reference number M21-0021). Work by Michael Roth was funded by the DFG Emmy Noether program (RO 4848/2-1).

References

Anna Aksenova, Ekaterina Gavrishina, Elisei Rykov, and Andrey Kutuzov. 2022. RuDSI: Graph-based word sense induction dataset for Russian. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 77–88, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Nour Al-khdour, Mutaz Bni Younes, Malak Abdullah, and Mohammad AL-Smadi. 2020. JUSTMasters at SemEval-2020 task 3: Multilingual deep learning model to predict the effect of context in word similarity. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 292–300, Barcelona (online). International Committee for Computational Linguistics.

David Alfter and Mattias Appelgren. 2025. GRASP at CoMeDi Shared Task: Multi-strategy modeling of annotator behavior in multi-lingual semantic judgments. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Nikolay Arefyev, Maksim Fedoseev, Vitaly Protasov, Daniil Homskiy, Adis Davletov, and Alexander Panchenko. 2021. DeepMistake: Which senses are hard to distinguish for a word-in-context model. volume 2021-June, pages 16–30.

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49, Barcelona (online). International Committee for Computational Linguistics.

Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Niemeyer, Tübingen.

Pierluigi Cassotti, Lucia Siciliani, Marco de Gemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th International Workshop on Computational Approaches to Historical Language Change*, Singapore. Association for Computational Linguistics.

Tejaswi Choppa. 2024. Supervised semantic proximity noise and disagreement detection. Master thesis, University of Stuttgart.

Tejaswi Choppa, Michael Roth, and Dominik Schlechtweg. 2025. Predicting median, disagreement and noise label in ordinal word-in-context data. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Phuoc Duong Huy Chu. 2025. FuocChu_VIP123 at CoMeDi Shared Task: Disagreement ranking with xlm-roberta sentence embeddings and deep neural regression. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

D. A. Cruse. 1995. *Polysemy and related phenomena from a cognitive linguistic viewpoint*, chapter 2. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Anna Hätty, Dominik Schlechtweg, and Sabine Schulte im Walde. 2019. SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*, pages 1–8, Minneapolis, MN, USA.

Geoffrey I. Hinton. 1990. *Connectionist learning procedures*, pages 555—-610. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Daniil Homskiy and Nikolay Arefyev. 2022. DeepMistake at LSCDiscovery: Can a multilingual word-in-context model replace human annotators? In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 173–179, Dublin, Ireland. Association for Computational Linguistics.

Adam Kilgarriff. 1997. “I don’t believe in word senses”. *Computers and the Humanities*, 31(2).

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Mikhail Kuklin and Nikolay Arefyev. 2025. Deepchange at CoMeDi: the cross-entropy loss is not all you need. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical Semantic Change Discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021a. Rushifteval: a shared task on semantic shift detection for russian. *Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Dialog conference*.

Andrey Kutuzov and Lidia Pivovarova. 2021b. Three-part diachronic semantic change dataset for russian. *Preprint*, arXiv:2106.08294.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Tai Duc Le and Thin Dang Van. 2025. MMLabUIT at CoMeDi Shared Task: Text embedding techniques versus generation-based nli for median judgment classification. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Zhu Liu, Zhen Hu, and Ying Liu. 2025. JuniperLiu at CoMeDi Shared Task: Models as annotators in lexical semantics disagreements. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Ying Xuan Loke, Dominik Schlechtweg, and Wei Zhao. 2025. ABDN-NLP at CoMeDi Shared Task: Predicting the aggregated human judgment via weighted few-shot prompting. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics*, 42(2):245–275.

Roberto Navigli. 2009. Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: an extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13648–13656. AAAI Press.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tetsuya Sakai. 2021. Evaluating evaluation measures for ordinal classification and ordinal quantification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2759–2769.

Gerard Salton and Michael J McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw - Hill Book Company, New York.

Olufunke O. Sarumi, Charles Welch, Christin Seifert, Lucie Flek, and Jörg Schlötterer. 2025. Funzac at CoMeDi Shared Task: Modeling annotator disagreement from word-in-context perspectives. In *Proceedings of the 1st Workshop on Context and Meaning–Navigating Disagreements in NLP Annotations*, Abu Dhabi, UAE.

Dominik Schlechtweg. 2023. *Human and Computational Measurement of Lexical Semantic Change*. Ph.D. thesis, University of Stuttgart, Stuttgart, Germany.

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte Im Walde, and Nina Tahmasebi. 2024. More DWUGs: Extending and evaluating word usage graph datasets in multiple languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14379–14393, Miami, Florida, USA. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 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*, pages 169–174, New Orleans, Louisiana.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA. Association for Computing Machinery.

Charles Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Sean Trott and Benjamin Bergen. 2021. RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7077–7087, Online. Association for Computational Linguistics.

David Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4(3):273–290.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385—–1470.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

A Label distributions

Find aggregated label distributions for all languages combined in Figures 1 and 2.

B Hyperparameter grid and optimized parameters

Find the hyperparameter grid used for tuning Baseline 4 in Table 6 and the final optimized hyperparameter combinations in Table 7.

C Submission overview

Find results for all submitted predictions in Table 8.

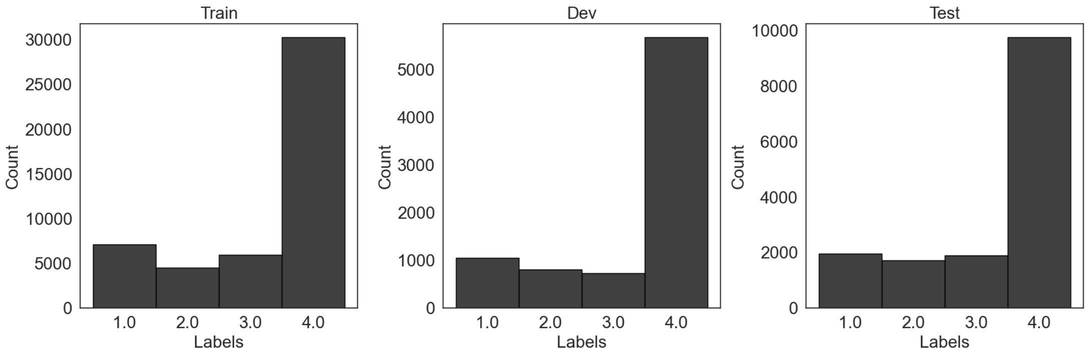

Figure 1: Label distribution for OGWiC task for all languages combined.

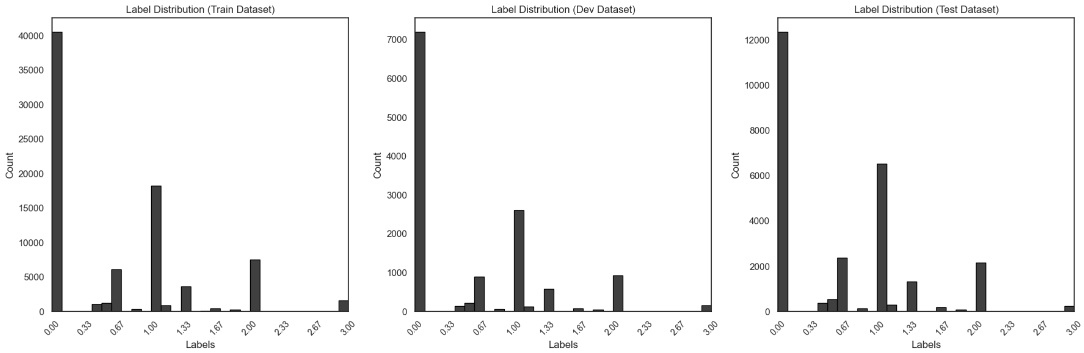

Figure 2: Label distribution for DisWiC task for all languages combined.

Hyperparameter	Values
activation	relu, tanh
solver	Adam
hidden layer sizes	10, 50, 100
alpha	.0001, .001, .01, .1
batch size	32, auto, 50, 100
scaler	StandardScaler(), None

Table 6: Hyperparameter grid used for tuning Baseline 4.

Hyperparameter	ZH	EN	DE	NO	RU	ES	SV
Activation	tanh	tanh	relu	relu	tanh	tanh	tanh
Alpha	.001	.1	.0001	.001	.1	.1	.0001
Batch Size	auto	auto	auto	100	100	auto	100
Hidden Layer Sizes	(50,)	(50,)	(50,)	(50,)	(100,)	(100,)	(100,)
Scaler	None	yes	yes	yes	yes	yes	yes

Table 7: Final set of hyperparameters for Baseline 4 per language.

Task	Team	AV	-ES	ZH	EN	DE	NO	RU	ES	SV
OGWiC	deep-change	**.66**	.64	**.42**	**.73**	.72	**.67**	.62	**.75**	**.68**
	deep-change	.65	.64	.42	.73	.72	.63	**.63**	.75	.68
	Baseline 2	.58	.57	.38	.65	**.73**	.52	.55	.66	.60
	GRASP	.56	.54	.32	.56	.66	.59	.49	.64	.65
	MMLabUIT	.52	.51	.36	.57	.67	.44	.42	.60	.61
	MMLabUIT	.52	.51	.32	.52	.65	.46	.42	.57	.66
	MMLabUIT	.52	.51	.35	.53	.66	.45	.43	.58	.63
	GRASP	.51	.50	.33	.57	.62	.47	.46	.59	.56
	GRASP	.43	.41	.18	.61	.51	.29	.34	.58	.48
	JuniperLiu	.27	.26	.14	.51	.49	.08	.13	.33	.22
	Baseline 1	.12	.12	.06	.10	.27	.12	.11	.18	.02
DisWiC	deep-change	**.23**	.23	.30	.08	**.20**	.29	**.18**	**.19**	**.35**
	GRASP	.22	.23	**.54**	.04	.11	.27	.17	.12	.30
	GRASP	.22	.23	.50	**.10**	.12	**.32**	.16	.10	.23
	GRASP	.16	.17	.26	.06	.13	.27	.11	.10	.20
	Baseline 4	.16	.17	.49	.06	.09	.24	.12	.08	.08
	Baseline 3	.12	.12	.39	.06	.09	.08	.05	.08	.08
	FuocChu_VIP123	.12	.14	.36	.02	.10	.16	.05	.01	.17
	FuocChu_VIP123	.11	.13	.35	.01	.10	.13	.04	.01	.15
	JuniperLiu	.08	.09	.36	.04	.02	-.04	.07	.04	.09
	JuniperLiu	.08	.09	.36	.04	.02	-.04	.07	.04	.08
	sunfz1	.07	.07	.30	.05	-.00	-.07	.07	.04	.09

Table 8: All results for both evaluation phases. ‘AV’ = Average over languages; ‘-ES’ = Average over languages excluding Spanish.