JuniperLiu at CoMeDi Shared Task: Models as Annotators in Lexical Semantics Disagreements

Zhu Liu[1]*, **Zhen Hu**[2]*, **Ying Liu**[1]
[1]School of Humanities, Tsinghua University, Beijing, China
[2]College of Engineering, Beijing Forestry University, Beijing, China
liuzhu22@mails.tsinghua.edu.cn, huzhen@bjfu.edu.cn

Abstract

We present the results of our system for the CoMeDi Shared Task, which predicts majority votes (Subtask 1) and annotator disagreements (Subtask 2). Our approach combines model ensemble strategies with MLP-based and threshold-based methods trained on pretrained language models. Treating individual models as virtual annotators, we simulate the annotation process by designing aggregation measures that incorporate continuous relatedness scores and discrete classification labels to capture both majority and disagreement. Additionally, we employ anisotropy removal techniques to enhance performance. Experimental results demonstrate the effectiveness of our methods, particularly for Subtask 2. Notably, we find that standard deviation on continuous relatedness scores among different model manipulations correlates with human disagreement annotations compared to metrics on aggregated discrete labels. The code will be published at https://github.com/RyanLiut/CoMeDi_Solution.

1 Introduction

Lexical semantic similarity is a classical task that encompasses various forms, including multi-choice sense selection (Navigli, 2009), binary classification (Pilehvar and Camacho-Collados, 2019), and contextual word similarity (Islam and Inkpen, 2008), among others. However, the potential disagreements among annotators, arising from the inherent vagueness and continuous nature of meaning, have received comparatively less attention. To address these complexities, the CoMedi workshop (Context and Meaning - Navigating Disagreements in NLP Annotations[1]) introduced a Shared Task with two subtasks (Schlechtweg et al., 2025). Subtask 1 involves predicting the median judgment classification across four candidate labels, which represent the degree of similarity for a target word in context. Subtask 2 focuses on predicting annotator disagreement, which can be interpreted as a form of predictive uncertainty estimation (Gal, 2016).

In this paper, we first conceptualize the two subtasks as corresponding to two fundamental statistical properties of a Gaussian distribution: the mean and variance. Subsequently, we model each system, parameterized by specific variables, as an individual human annotator. These variables encompass both homogeneous factors, such as layers within the same model, and heterogeneous factors across different models. To address the tasks, we employ MLP-based and threshold-based approaches to generate continuous relatedness [2] scores and discrete classification labels, respectively. Additionally, we incorporate techniques for anisotropy removal to mitigate geometric biases inherent in embedding spaces. Finally, we propose diverse strategies for model ensembling to enhance performance. Our results demonstrate the effectiveness of threshold-based methods combined with anisotropy removal and MLP-based approaches. For Subtask 2, the findings further highlight the advantages of aggregating relatedness scores over discrete labels in capturing annotator disagreement.

2 Related Work

Probing for Contextual Word Meaning Tasks capturing word meaning in context include word sense disambiguation (WSD) (Navigli, 2009), which selects the most appropriate sense, and WiC (Pilehvar and Camacho-Collados, 2019), which determines semantic equivalence across contexts. Extending these, relatedness scoring provides a continuous measure of semantic relatedness.

*These authors contributed equally.
[1]https://comedinlp.github.io/
[2]We distinguish *similarity* from *relatedness*, with the task focusing on annotating relatedness scores.

Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation, pages 103–112
January 19, 2025.

The CoMeDi Shared Task reframes WiC as an ordinal classification task with four labels indicating relatedness degrees. Probing methods include MLP-based approaches (Tenney et al., 2019; Pilehvar and Camacho-Collados, 2019), which train dense networks, and threshold-based methods (Pilehvar and Camacho-Collados, 2019; Vulić et al., 2020; Liu et al., 2024), which optimize relatedness thresholds for pretrained representations. Since embeddings are often anisotropic (Ethayarajh, 2019), techniques like centering (Sahlgren et al., 2016) and standardization (Timkey and van Schijndel, 2021) are applied to improve representation quality.

Uncertainty Estimation Subtask 2 models annotator disagreement, aligning with the study of uncertainty estimation (UE), widely explored in computer vision (Gal, 2016) and robust AI (Stutz, 2022). UE arises from data uncertainty (aleatoric, linked to inherent data ambiguity like annotation disagreement) and model uncertainty (epistemic, due to biased learning on out-of-distribution data) (Gal, 2016). Researchers (Liu and Liu, 2023) combine these areas to model semantic uncertainty in sense selection. While Bayesian (Vazhentsev et al., 2022) and non-Bayesian (Szegedy et al., 2016) methods often use label probabilities, our threshold-based method lacks this feature. Instead, we treat the process as model ensemble (Lakshminarayanan et al., 2017) and propose aggregation measures.

Annotator Disagreement Annotator disagreement is common in lexical semantics tasks, such as word sense disambiguation (WSD) (Navigli, 2009; Chklovski and Mihalcea, 2003), due to the subjective and ambiguous nature of meaning (Navigli, 2008). While many studies resolve disagreement through majority voting, others exploit it by reframing tasks as multi-label classification (Conia and Navigli, 2021) or training on multiple judgments (Uma et al., 2021).

In this paper, we model annotator disagreement as uncertainty estimation, as both involve (1) output variability, (2) data noise [3], and (3) similar evaluation metrics.

3 System Overview

Most systems use MLP-based (Tenney et al., 2019; Pilehvar and Camacho-Collados, 2019) and threshold-based (Pilehvar and Camacho-Collados, 2019; Vulić et al., 2020; Liu et al., 2024) methods. They extract representations from pretrained language models, then MLP-based methods train a network to predict discrete labels (Subtask 1) or continuous values (Subtask 2). Threshold-based methods learn a threshold selector to map similarity scores to labels. However, naive baselines often fall short, as shown in Section 5. In our system, we applied anisotropy removal to the baseline code (Schlechtweg et al., 2025) and used a classifier-based method for comparison. For Subtask 1, we apply techniques to make data points more isotropic. For Subtask 2, we ensemble models, treating them as annotators, and use various strategies to model disagreement.

3.1 Formulation as Parameter Estimation

For a target word w appearing in a pair of contexts c_i and c_j, annotators from a hypothetical human space $\mathcal{H}$ provide a judgment score $s \in \mathcal{R}$, where higher values indicate greater similarity in meaning between c_i and c_j. These scores form a judgment distribution p on $\mathcal{R}$, which we assume follows a Gaussian distribution, $p \sim \mathcal{N}(\mu, \sigma^2)$, as it is a natural statistical choice (Jaynes, 2003). Here, μ represents the consensus similarity, while σ reflects disagreement among annotators.

In practice, the continuous Gaussian distribution is discretized due to the finite number of annotators and graded annotations. Nonetheless, we adopt the Gaussian framework to unify the two tasks: Subtask 1 estimates μ, while Subtask 2 estimates σ.

3.2 Subtask 1: Anisotropy Removal

Contextual representations are known to be anisotropic (Ethayarajh, 2019), clustering in a narrow region of the space. This inflates similarity scores, reducing their discriminative power in meaning-related tasks. For example, even unrelated words often exhibit high similarity. We adopt three techniques to reduce anisotropy: (1) centering by subtracting the mean vector (2) normal standardization (3) All-but-the-top (Mu and Viswanath, 2018): subtracting the projection on the component of the largest variance.

3.3 Subtask 2: Model Ensembling

To model annotator disagreement, we treat each model or its manipulation as an annotator and use three measures to reflect uncertainty. We explore

[3] Annotator disagreement can be viewed as label noise, contributing to data uncertainty—a key component of irreducible uncertainty.

three ensembling strategies: (1) homogeneous aggregation with model manipulations (e.g., layer and anisotropy removal), (2) heterogeneous ensembling across different models, and (3) a mixed approach combining both. After each model forward pass, we obtain a discrete label using the threshold-based model [4] and a continuous relatedness score. We apply three measures: standard deviation (STD) for continuous scores, mean pairwise absolute judgment differences (MPD) for discrete labels (as used in Subtask 2), and variation ratio (VR), the ratio of values not equal to the mode, commonly used in uncertainty estimation (Gal, 2016).

4 Experiment Setup

4.1 Task Description

The Shared Task in the workshop of CoMeDi (Schlechtweg et al., 2025) includes two subtasks. The first aims to predict a discrete label (from 1 to 4) to show the relatedness of the target word in two contexts while the second obtains a continuous value to indicate the disagreement. The task data was sampled from multilingual datasets, involving 7 languages, i.e., Chinese (Chen et al., 2023), English (Schlechtweg et al., 2021, 2024), German (Schlechtweg et al., 2018, 2021, 2024; Hätty et al., 2019; Kurtyigit et al., 2021; Schlechtweg, 2023), Norwegian (Kutuzov et al., 2022), Russian (Rodina and Kutuzov, 2020; Kutuzov and Pivovarova, 2021; Aksenova et al., 2022), Spanish (Zamora-Reina et al., 2022), Swedish (Schlechtweg et al., 2021, 2024).

4.2 Models

Our study focuses on threshold-based methods using pre-trained models: XLM-RoBERTa-base, XLM-RoBERTa-large (Conneau, 2019), BERT-multi-base (Pires, 2019), and Llama-7B (Touvron et al., 2023). For encoder-only models, we extract target word representations directly, while for the decoder-only Llama-7B, we use a prompt-based method (Liu and Liu, 2023) to extract the final colon representation. Inspired by in-context learning (Jiang et al., 2024), we apply layer-wise manipulations (centering, standardization, and all-but-the-top) to reduce anisotropy.

We discretize the continuous similarity scores into labels using a threshold selector based on the shared task baseline. The selector employs the Nelder-Mead method (Nelder and Mead, 1965) to optimize bin edges for Krippendorff's α, starting with evenly spaced bins and iteratively refining them. For Subtask 2, we explore model ensembling strategies (homo, hetero, mixed) and different measures (STD, MPD, VR), and also evaluate an MLP-based approach (details in Appendix 10.1).

[4] For Subtask 2, we use the majority of judgment scores as the GT label, avoiding the median to handle decimals.

4.3 Evaluation Phase Setting

During the evaluation phase, we selected models based on the development set.

For **Subtask 1**, we employ a threshold-based method using XLM-RoBERTa-base as the pre-trained model, except for Chinese and Russian (BERT-multi-base) and Norwegian (LERT-base-chinese). Representations are extracted from the 10th layer for XLM-RoBERTa-base and the final layer for other models. We apply normal standardization except for Norwegian to address anisotropy and utilized the threshold selection method from the official baseline code (Schlechtweg et al., 2025). Specifically,

For **Subtask 2**, we fine-tune an MLP regressor to predict disagreement scores, following the baseline methodology. It comprised of two linear layers and a ReLU activation function. For Swiss, we train for 50 epochs with a batch size of 32, while for other languages, we use 200 epochs with a batch size of 16. The learning rate is 1e-2 with a 0.1 dropout rate. We utilize AdamW for optimization with a warm-up ratio of 0.1.

4.4 Post Evaluation Phase Setting

For **Subtask 1**, we use the 25th layer of Llama and the 11th layer of XLM-RoBERTa-Base for all languages, with an MLP-based method fine-tuned using training data. All model representations are standardized to remove anisotropy. For the MLP-based model, we train for 50 epochs with a batch size of 128, an initial learning rate of 1e-2, and apply a dropout rate of 0.1 to prevent overfitting.

For **Subtask 2**, we employ ensembling strategies to significantly improve performance. We report two results from our ensembling methods. The first (ensembling) applies the same strategy across all languages: standardization with layer 24, no standardization with layer 16, centering with layer 24, and all-but-the-top with layer 16, all on Llama-7B. The second (ensembling*) presents language-specific ensembling strategies, as in Table 6.

Participator	Method	AVG	ZH	EN	DE	NO	RU	ES	SV
kuklinmike	-	**0.656**	**0.424**	**0.732**	0.723	**0.668**	**0.623**	**0.748**	**0.675**
comedy_baseline_2	-	0.583	0.379	0.654	**0.728**	0.515	0.550	0.656	0.601
daalft	-	0.555	0.317	0.555	0.656	0.589	0.487	0.636	0.648
ours	Thr* (XLM-R-B)	0.271	0.140	0.507	0.492	0.080	0.128	0.330	0.224
ours	Thr (LLM)	**0.451**	-0.090	0.474	**0.696**	**0.445**	**0.444**	**0.623**	**0.566**
ours	Thr (XLM-R-B)	0.339	**0.148**	**0.524**	0.485	0.240	0.301	0.348	0.325
ours	MLP	0.338	0.128	0.369	0.371	0.351	0.329	0.411	0.407

Table 1: Results for Subtask 1. The upper part shows the evaluation phase, and the lower part the post-evaluation phase. "Thr" denotes threshold-based methods, and "Thr*" indicates language-specific model selections. The same applies to other tables.

Participator	Method	AVG	ZH	EN	DE	NO	RU	ES	SV
kuklinmike	-	**0.226**	0.301	0.078	0.204	**0.286**	**0.175**	**0.187**	0.350
daalft	-	0.220	**0.539**	0.042	0.108	0.272	0.167	0.115	0.296
comedy_baseline_2	-	0.163	0.485	0.060	0.085	0.235	0.116	0.078	0.079
ours	MLP	0.082	0.358	0.038	0.022	-0.042	0.067	0.040	0.090
ours	ensembling	0.205	0.274	0.117	0.236	0.279	0.101	0.073	0.353
ours	ensembling*	0.220	0.347	**0.118**	**0.242**	0.283	0.108	0.078	**0.364**

Table 2: Evaluation results (upper part) and post-evaluation results (lower part) for Subtask 2. The method *ensembling** integrates language-specific ensembling strategies, while *ensembling* uses the strategy with the best average score across all languages.

5 Results

We present the results in Table 1 and Table 2 on the **test** set. The upper sections show evaluation phase scores submitted to the leaderboard, while the lower sections display post-evaluation results using public answers. We then conduct abalation studies on the **development** set in later sections.

In the evaluation phrase, for **Subtask 1**, our threshold-based method achieved moderate results, with LERT-base-chinese performing relatively better for Norwegian, though with limitations. For **Subtask 2**, we fine-tuned an MLP to predict disagreement scores but observed limited performance, prompting alternative methods in the post-evaluation phase.

In the post-evaluation phase, for **Subtask 1**, the threshold-based model performed comparably to the MLP-based model, while large language models (LLMs) showed superior results, highlighting their potential. For **Subtask 2**, our results matched the evaluation phase's top performances, confirming the effectiveness of the ensembling approach.

5.1 Abalation Study on Subtask 1

Figure 1 shows the average performance change with different anisotropy removal methods across layers. The large gap between removal and non-removal emphasizes the importance of this technique. Performance improves with higher layers, except for a drop in the last one or two layers. Standardization consistently performs best across all layers.

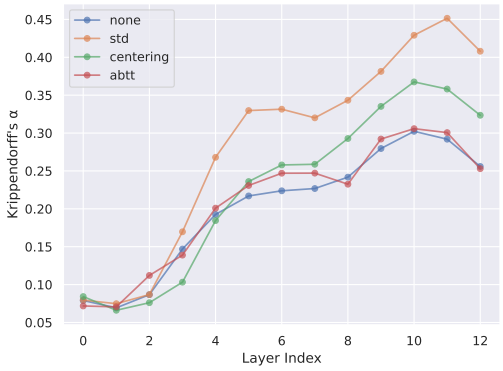

Figure 1: Performance of different types of anisotropy removal with the increase of layer index. 0 indicates the input embedding. "abtt" means all-but-the-top.

Figure 2 displays the performance of different models. Since Llama-7B is a decoder-only model with significantly more parameters and training data, its optimal result (Layer 25) serves as an upper bound [5]. The results show that XLM-RoBERTa-base outperforms all other models, including its larger counterpart.

[5]We attempt representations of different layers from Llama-7B, and the optimal layer index is 25.

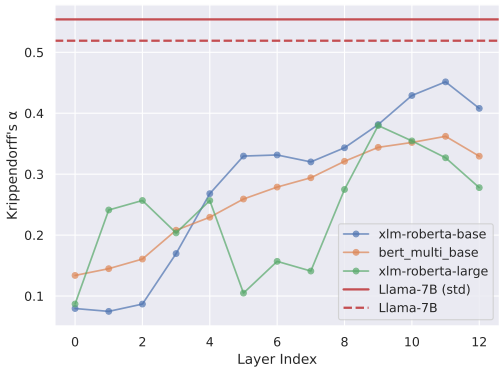

Figure 2: Performance of different models as the layer index increases. The optimal result (Layer 25) for Llama-7B and its standardized version are shown as the upper bound.

Figure 3: Performance of three types of measures across 500 random runs.

Type	1	2	3	4	5
homo	**0.237**	**0.235**	**0.235**	**0.234**	**0.233**
hete	0.217	0.216	0.215	0.209	0.201
mixed	0.228	0.222	0.219	0.213	0.203

Table 3: Top five groups for strategies of model selections

5.2 Abalation Study on Subtask 2

In this section, we analyze various factors influencing ensembling performance, including the choice of measure and model selection. We evaluate four candidate models i.e., XLM-RoBERTa-base, XLM-RoBERTa-large, BERT-multi-base, and Llama-7B, with four types of anisotropy removal and four layer levels. For layer levels, we extract layers 1, 4, 7, 10 for encoder-only models, and layers 8, 16, 24, 32 for the Llama model, yielding 64 possible model configurations. We use a threshold-based method for each model to obtain both a continuous similarity score and a discrete classification label, as we have done in Subtask 1. We randomly select a subset of 4 models from these possibilities, referred to as "mixed". Additionally, we experiment with homogeneous aggregation (using the same model) and heterogeneous aggregation (using different models). For homogeneous aggregation, we choose Llama-7B due to its superior performance. For each category, we sample 500 model subsets, obtaining both their classification labels using a threshold-based method and relatedness scores based on pre-trained embeddings. We first evaluate three measures (STD, MPD, and VR) in the mixed setting, selecting the best one to compare different category choices.

Measure Figure 3 presents the results for three measures. In most cases, STD on a continuous similarity score outperforms the others, while MDP slightly exceeds VR on the discrete classification labels . This suggests that similarity scores have an advantage over discrete labels due to the robustness of continuous values. Label prediction can be seen as a discretization of the continuous counterpart, leading to a loss of precision. Thus, we select STD as our final measure.

Model Selection Table 3 shows the top five results for three ensemble strategies. The specific model groups are listed in Tabel 7. Homogeneous model manipulations (homo) outperform mixed ensembles, while combining different models yields the worst performance. This suggests that model variance can still serve as an effective alternative, aligning with the use of dropout in uncertainty estimation (Gal, 2016).

6 Conclusion

We present our system for two subtasks released on CoMeDi Shared Task. We first formalize these tasks as parameter estimation where Subtask 1 estimates a mean and Subtask 2 the variance for a hypothetical Gaussian distribution. Then we mainly adopt threshold-based method with different techniques of anisotropy removal to classify the label for Subtask 1. Inspired by the area of uncertainty estimation, we utilize model ensembling with various strategies to select models and measures to reflect disagreement for Subtask 2. Experiments show the effectiveness of our method.

7 Limitations

We acknowledge several limitations in our system. First, the model training process utilizes data from all languages without considering their unique linguistic characteristics. For instance, Chinese exhibits rich formation rules (Zheng et al., 2021), yet lacks the morphological complexity found in Western languages, potentially leading to distinct patterns of disagreement. Second, our parameter estimation for the Gaussian distribution does not account for the estimation of the mean, which could be incorporated into Subtask 1 for a more comprehensive approach. Furthermore, in Subtask 2, we employ the median of all annotations as an independent label for the model instead of using individual annotations. This approach may introduce inconsistencies with our formulation of *models as annotators*. Lastly, while our experiments highlight the potential of large language models (LLMs) compared to pretrained language models, future work will focus on exploring more effective strategies for extracting lexical representations from LLMs.

8 Ethics Statement

We do not foresee any immediate negative ethical consequences arising from our research.

9 Acknowledgements

The authors thank the anonymous reviewers for their valuable comments and constructive feedback on the manuscript. This work is supported by the 2018 National Major Program of Philosophy and Social Science Fund "Analyses and Researches of Classic Texts of Classical Literature Based on Big Data Technology" (18ZDA238) and Research on the Long-Term Goals and Development Plan for National Language and Script Work by 2035 (ZDA145-6).

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.

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.

Timothy Chklovski and Rada Mihalcea. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Recent Advances in Natural Language Processing*.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Yarin Gal. 2016. *Uncertainty in Deep Learning*. Phd thesis, University of Cambridge.

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 Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 1–8, Minneapolis, Minnesota. Association for Computational Linguistics.

Aminul Islam and Diana Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, 2(2):10:1–10:25.

Edwin T Jaynes. 2003. *Probability theory: The logic of science*. Cambridge university press.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.

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. 2021. Rushifteval: a shared task on semantic shift detection for russian. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*.

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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024. Fantastic semantics and where to find them: Investigating which layers of generative LLMs reflect lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14551–14558, Bangkok, Thailand. Association for Computational Linguistics.

Zhu Liu and Ying Liu. 2023. Ambiguity meets uncertainty: Investigating uncertainty estimation for word sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3963–3977.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Roberto Navigli. 2008. A structural approach to the automatic adjudication of word sense disagreements. *Natural Language Engineering*, 14(4):547–573.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

John A. Nelder and Roger Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, pages 1267–1273.

T Pires. 2019. How multilingual is multilingual bert. *arXiv preprint arXiv:1906.01502*.

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.

Magnus Sahlgren, Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Akshay Viswanathan, and Anders Holst. 2016. The gavagai living lexicon. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Dominik Schlechtweg. 2023. *Human and computational measurement of lexical semantic change*. Ph.D. thesis, University of 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*, Miami, Florida. Association for Computational Linguistics.

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth. 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.

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, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

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.

David Stutz. 2022. Understanding and improving robustness and uncertainty estimation in deep learning. *Saarländische Universitäts-und Landesbibliothek*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, et al. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

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.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021. Leveraging word-formation knowledge for Chinese word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 918–923, Punta Cana, Dominican Republic. Association for Computational Linguistics.

10 Appendix

10.1 MLP-based Methods

We attempt the MLP-based method in two subtasks, freezing XLM-RoBERTa-base model parameters to obtain the vector representation of the target word, and training a classifier or a regression model downstream. MLP1 is a linear layer, MLP2 represents two linear layers, and uses the ReLU activation function.

In Subtask 1, we use the cross-entropy loss function to train a classifier. The results on the development dataset are shown in Table 4. We find that two linear layers achieve better results. We attempt to use a weighted cross-entropy loss function to alleviate the problem of sample imbalance, but shows slight improvement. We compare the results of different layers of the model and find that the vector representation of the shallower layers(11th) achieves better results. We attempt layer fusion and average pooling of the vectors in the last 4 layers, which results in more stable improvements.

Training settings for the MLP-based method in Subtask 1: 50 epochs, batch size of 128, 1e-2 learning rate, AdamW optimizer, and a dropout rate of 0.1 to improve generalization.

In Subtask 2, we use the mean square error loss function to train a regression model that directly predicts continuous values of inconsistent labeling of target words, similar to the baseline provided by the official source. The results on the development dataset are shown in Table 5. We find that two linear layers are worse than a single linear layer.

We try multiple different hyperparameter settings on Task 2. In MLP1, we ultimately chose 200 epochs, batch size of 16, while in MLP2, we chose 50 epochs, batch size of 32. Other training settings: 1e-2 learning rate, AdamW optimizer, and a dropout rate of 0.1.

10.2 Model Groups

We use letters to denote different models: A, B, C, and D represent Llama-7B, XLM-RoBERTa-base, BERT-multi-base, and XLM-RoBERTa-large, respectively.

For encoder-only models, h, i, j, and k indicate layers 1, 4, 7, and 10, respectively; whereas in large language models (LLMs), these symbols correspond to layers 8, 16, 24, and 32.

X, Y, Z, and W correspond to four standardization methods: non-standard, std, centering, and all-but-the-top.

Model groups for specific languages. We experiment with various model groups, and different groups achieve the best results in different languages. Table 2 shows the best results for the test dataset in Subtask 2, and the specific model groups are shown in Table 6.

Top 5 model groups. We employ three ensemble strategies, and the top five results of each strategy on the development dataset of Subtask 2 are presented in Table 3, with corresponding model groups shown in Table 7.

Method	AVG	ZH	EN	DE	NO	RU	ES	SV
MLP1	0.191	0.105	-0.140	0.192	0.337	0.276	0.418	0.151
weighted loss	0.240	0.361	0.110	0.166	0.156	0.255	0.354	0.277
layer11	0.265	0.267	0.009	0.261	0.357	0.298	0.341	0.321
MLP2	0.407	0.519	0.268	**0.609**	0.360	0.265	0.565	0.262
layer11	0.418	0.530	**0.384**	0.511	**0.416**	0.311	0.576	0.198
last4layer	**0.429**	**0.509**	0.229	0.570	0.306	**0.416**	**0.584**	**0.386**

Table 4: Evaluation results for Subtask 1 in MLP-based methods. The upper part presents the outcomes of using a single linear layer as a classifier, where "weight loss" indicates the employment of a weighted cross-entropy loss function, and "layer11" denotes utilizing the vector representations from the 11th layer of the language model. The lower part illustrates the results obtained by employing two linear layers as classifiers, showing the performance of the 11th layer of the model as well as the outcome after applying average pooling to the last four layers of the model.

Method	AVG	ZH	EN	DE	NO	RU	ES	SV
MLP1	**0.128**	**0.323**	**0.088**	**0.179**	**0.132**	**0.061**	0.026	0.083
MLP2	0.098	0.232	-0.061	0.131	0.119	0.020	**0.061**	**0.187**

Table 5: Evaluation results for Subtask 2 in MLP-based methods, demonstrating the results of Multi-Layer Perceptrons (MLPs) with different numbers of layers.

Language	Model Groups
Chinese	AiX-AkX-AhX-AkW
English	AjZ-AiX-AjX-AjW
German	AhW-AjX-AjW-AjZ
Norwegian	AjZ-AiX-AjX-AjW
Russian	AiX-AiW-AkW-AkZ
Spanish	AhY-AiZ-AhX-AhW
Swedish	AiX-AkY-AjZ-AjY

Table 6: The optimal model groups for each specific language for the development set in Subtask 2.

Type	1	2	3	4	5
homo	**AjY-AjZ-AiX-AiW**	**AjY-AiW-AjZ-AjX**	**AjX-AiX-AiW-AjY**	**AjZ-AiX-AjX-AjW**	**AiX-AjZ-AjX-AhX**
hete	AjY-BkW-ChZ-DkX	AjY-BiX-ChW-DjX	AjY-BkW-CiX-DiW	AjZ-BkW-ChY-DhW	AjY-BhW-ChY-DiX
mixed	AjX-ChX-AiX-AjZ	AjY-AiX-AjW-ChY	AkW-ChX-AjY-AjW	AjZ-ChY-DhX-DkX	ChX-AkY-AiX-AiW

Table 7: Top five model groups when ensembling models for Subtask 2.