Final report: Probabilistic linear regression of pairwise proximity judgments

Introduction to Bayesian statistics
March 31, 2025

Long Le Hoang
IMS, University of Stuttgart
Immatrikulation: 3796175
st192583@stud.uni-stuttgart.de

Abstract

Following the recent shared task on Disagreement in Word-in-Context Ranking (DisWiC), we show the application of Bayesian methods in predicting the annotator disagreement on semantic proximity judgments between two usages of a word. Instead of directly modeling the disagreement like previous works, we use linear regression to estimate the annotated pairwise proximity from embedding features, then sample from the proximity posterior distribution to get a prediction on the disagreement. As the linear model is very simple, we cannot draw much conclusion from the results, but this is only the first step toward a more rigorous Bayesian approach to annotator disagreement modeling in NLP.

1 Introduction

Contextual embeddings (Devlin et al., 2019; Conneau et al., 2019) are representations of word tokens in their context, learned by language models from a large amount of text data. They have been used to achieve great success for many downstream NLP tasks, especially ones that try to understand words in different contexts, such as the Word-in-Context task (WiC, Pilehvar and Camacho-Collados 2019). With the same word in two different contexts, this task originally asks the model to classify if the pair has the same meaning or not, but other later datasets have been used to ask the model to detect a more nuanced distinction between two tokens (Armendariz et al., 2020; Schlechtweg et al., 2018). Along with this development of the task is the realization that the disagreement in labeling, once discarded as noise, has now become increasingly important as a source of information for difficult semantic tasks (Leonardelli et al., 2023).

Following these developments, the 2025 Workshop on Context and Meaning - Navigating Disagreements in NLP Annotations (CoMeDi, Schlechtweg et al., 2025 aims to investigate the WiC task with a newer lens. They criticize old tasks for their ranking nature, stating that an ordinal numerical representation of the semantic closeness between two usages can provide researchers with a beneficial linguistic interpretation. They also claim that annotators for WiC datasets very often disagreed on the ordinal labels, and propose a new shared task called Disagreement in Word-in-Context Ranking (DisWiC) to utilize the disagreement among annotators. Unlike previous methods that aggregate labels (Leonardelli et al., 2023), Schlechtweg et al. (2025) asked participants at the shared task to rank the disagreement, thus detecting highly complicated context pairs. We also agree that knowing the level of disagreement is highly important in real world scenarios, but the relationship between i) features of two tokens and ii) whether it is difficult to classify the relationship between them is clearly complex and directly modeling the disagreement is a demonstrably hard task (Schlechtweg et al., 2025; Liu et al., 2024).

To better predict the annotator disagreement, we first devise a theoretical model of how people use the semantic distance between two contexts of the same word to judge their semantic proximity. We then propose a probabilistic linear model to represent the theory with careful consideration of the observed variables. The model is fitted with a training dataset comprising pairwise annotated judgments from different languages and time periods (Schlechtweg et al., 2018). Only after having a judgment model that we use posterior predictive sampling to simulate the process of annotation and evaluate the disagreement between those samples. Following this procedure is our best guess on how to replicate the mechanism that produces the observed disagreement, and the results can therefore be used to validate our theory. Unfortunately, even after experimenting with different methods to represent the details in our model, we have not got a good result to draw any worthwhile conclusion.

To investigate more thoroughly, we inspect the posterior distribution we got from fitting the data to find any possible interpretation. We also discuss the implications for the theory and for future works, with suggestions on limitations and possible innovations. The overall outlook is still optimistic, as there are still a lot of spaces for improvement within this probabilistic approach to measure disagreement.

2 Related Work

WiC (Pilehvar and Camacho-Collados, 2019) is a binary classification task which a model predicts whether two tokens of the same word have the same meaning. It has been used in many domains such as Lexical Semantic Change Detection (LSCD, Schlechtweg et al., 2018. However, the problem of semantic proximity in different contexts is clearly more complex than a simple binary decision. Therefore, it requires a more fine-grained task and set of labels, such as the Graded Word Similarity in Context task (GWiC, (Armendariz et al., 2020)), where the semantic distinction is graded instead.

While this new task better represents the human intuition on token meanings, Schlechtweg et al. (2025) consider it not enough and instead propose a new ordinary regression based on the DURel annotation framework (Schlechtweg et al., 2018). This framework has a clear linguistic interpretation along each grade on the scale: highest is **4: identical meaning**, then **3: closely related** meaning - corresponds with context variance, then **2: distantly related** - often as polysemy, and farthest in semantic is **1: two unrelated meanings** of a homonymic pair.

This ordinal scale clearly has more advantages for tasks like LSCD than the labels of previous datasets, so many teams have worked with it at the CoMeDi workshop. For the ordinal regression task (OGWiC) there, Kuklin and Arefyev (2025) use a binarized version of XLM-RoBERTa (Conneau et al., 2019) as the based model and further fine-tuned with multilingual WiC data from different sources. Alfter and Appelgren (2025) experiment with 3 approaches: a probabilistic chain model, an ensemble of XLM-RoBERTa, and a model based on XL-lexeme. Le and Van (2025) try various BERT-based models (Devlin et al., 2019) further trained with nature language inference dataset. Liu et al. (2024) use cosine similarity over normalized XL-lexeme embeddings. All of them use a set of tuned adaptive thresholds to map the output to a suitable proximity category, as common in ordinal regression (Choppa et al., 2025)

WiC, like most other NLP tasks, used to simply filter out disagreed samples, but recently researchers have started to create methods that can utilize information coming with disagreement. Previous works either aggregate all the labels (Leonardelli et al., 2023) or try to learn and evaluate with the entire label distribution (Uma et al., 2021). Schlechtweg et al. (2025) instead create a ranking task that models the disagreement for WiC pairs (DisWiC). This is a difficult and very ambitious task, but most of the participating teams still use a variation of XLM and BERT to directly predict the disagreement. Their results are mostly not bad, explained by "the derivation of DisWiC labels from absolute differences between ordinal WiC annotation" (Schlechtweg et al., 2025), but there is still a huge gap to improve.

3 Data

The dataset used in this report, following (Schlechtweg et al., 2021), is created from a variety of Word Usage Graphs (WUGs) datasets. These datasets depict the different usages of a word as nodes within a graph, where the nodes are linked by weighted edges that represent the annotated semantic proximity. The overall dataset combines data from 7 languages (German, English, Swedish, Chinese, Russian, Spanish, Norwegian). The dataset is cleaned (by ensuring the identifiers are unique and removing NaN values) and aggregated into one pair of tokens for each column.

Similarly to (Schlechtweg et al., 2025), the labels annotated by different annotators are used to compute the median judgment and the mean absolute disagreement between labels:

$$D = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} |x_i - x_j|$$

A difference in our data compared to Schlechtweg et al. (2025) is that we keep the pairs that only has one annotator and consider the disagreement $D = 0$. This may be bad for the evaluation of disagreement ranking, but as we try to model the judgment first, we need as much data as possible. Then we will inspect the posterior distribution for better evaluation.

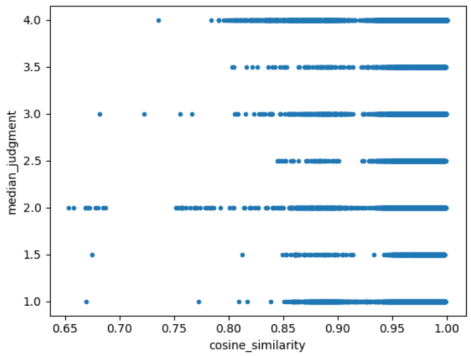

Figure 1: Cosine similarity plotted against the median semantic proximity judgment. As demonstrated, the data is not very suitable for a linear regression, but most pairs with low similarity have lower proximity (<=2), while a lot of pairs with high similarity have proximity judgment = 4

We split the dataset into a train set and a dev set with a proportion of 80/20, as we only fit the model on the train set while sample from both training and dev set. Afterward, we first filter out tokens without context, and then compute the contextual embeddings for each token. We choose to use XLM-RoBERTa (Conneau et al., 2019) to get the embeddings, since it is easier to use and faster to run even though XL-lexeme gives better results with WiC data (Schlechtweg et al., 2025). We then experiment with two types of representation for the token pair: cosine similarity (Fig. 3) and a concatenation version of two vectors, down-rank with PCA. Overall, the clean training set has 200912 token pairs and the clean dev set has 50083 token pairs, either each pair is represented by a scalar (cosine similarity) or a vector with $d = 8$) (PCA of concatenation).

4 Model

The ultimate goal of the DisWiC task is to rank the difficulty of categorization for each token pair, which Schlechtweg et al. (2025) use the mean absolute disagreement as the observed variable. To clarify the assumptions in this choice and motivate further statistical inquiry, we demonstrate a simple scientific model: For each annotator and for each token pair, they will read each of the two contexts that a word appears in, understand each meaning and intuitively compare these meanings. They then choose one of the DURel categories to judge the semantic distinction. As there can be disagreement in all the steps from understanding, comparison to grading choice, the level of difficulty is influenced by both hidden semantic variables and observed judgment decision (Fig. 2)

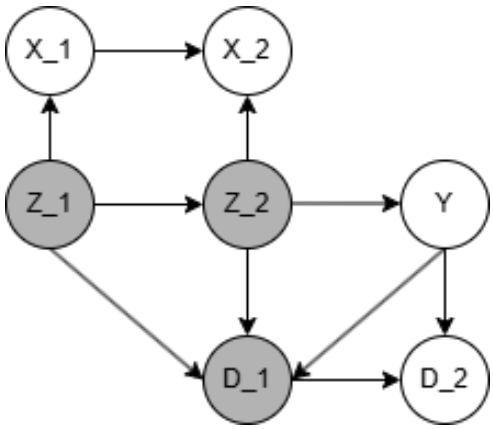

Figure 2: The Directed Acyclic Graph (DAG) representing all the variables in our theory. Z_1: hidden meanings of words in the semantic space, Z_2: hidden distance of meanings in the semantic space. X_1: contextual embeddings used to approximate the hidden meanings, X_2: features extracted from embeddings to simulate the comparison. Y: observed proximity rating. D_1: hidden level of complexity of token pairs, D_2: mean absolute disagreement rank

As we cannot access the hidden understanding and proximity intuition of an annotator in the semantic space, we will only approximate them with contextual embeddings X_1 (learned by XLM-RoBERTa) and a variable X_2 representing the distance (either cosine similarity or a feature vector). We then model the observed median judgment for each token pair indirectly from embeddings as:

$$Y \sim Normal(\beta X\_2 + \alpha, \sigma)$$

$$X\_2 = f(X\_1)$$

$$\beta \sim Normal(1, 1)$$

$$\alpha \sim LogNormal(2, 1)$$

$$\sigma \sim Uniform(0, 1)$$

The function f can be either a cosine or a PCA version of concatenation, but both of them do not really have a good interpretation (even cosine or its log version is not very nice when projected) so we have to choose less informative priors. As the proximity judgment is on the scale from 1 to 4 and we normalize X_2 to have a mean of 0, we choose β to be the Gaussian prior with a small variance to ensure the rating is positive. The α prior is set to be a LogNormal distribution so that most values are positive and their mode equals $e^{2-1} = e$ (most samples are around e).

As in Fig. 2, there is no direct relation between X and Y, the model only represents an indirect relation that cannot justify the hidden variables Z through Y alone, but if we use posterior predictive sampling from Y to get D_2, we will partially have evidence for Z, D_1 and their relation. For that purpose, we use a Markov Chain Monte Carlo (MCMC) method called No U-Turn Sampler (NUTS) to approximate the posterior distribution and then sample from it to simulate the procedure of different annotators giving different judgments.

5 Experiments

After processing and standardizing the variables, we fit the formulated model to the whole training set and approximate the posterior distribution with 4 different chains and 200 samples for each input. To evaluate the disagreement ranking, we sample 10 ratings for each input from the posterior predictive distribution, using input variables from both training set and dev set. Therefore, with each token pair we can calculate the mean absolute disagreement of those ratings and compare them to the gold disagreement measure of the annotated labels using Spearman ρ (Spearman, 1961).

Unfortunately, the results are not good enough to justify the aforementioned theory. We use two types of distance features: cosine similarity get 0.003 for train set ρ score and 0.002 for dev set ρ score; a PCA with $d = 8$ for the concatenation of two embeddings get 0.005 for both train and dev set ρ. This is worse than the baseline in the CoMEDi workshop (Schlechtweg et al., 2025) and requires more thorough investigation. We therefore follow (McElreath, 2020) and project the posterior distribution with 89% samples with respect to cosine similarity (Fig 3)

We may see that the interval of the prediction is mostly in the area from 1 to 4 of the proximity scale, so the priors have been chosen correctly. However we can also see that the interval is still too big, showing that the cumulative uncertainty in the linear model is too large and we have not been able to fit the model to the data. This is understandable as a linear model is too simple, while the judgment data is designed for ordinary regression instead. One clear way to improve the simulation therefore is using an ordinal model with the thresholds as prior variables. Another way is to try to incorporate other confounding variables that can lead to judgment, like change in style or time period (Schlechtweg et al., 2018).

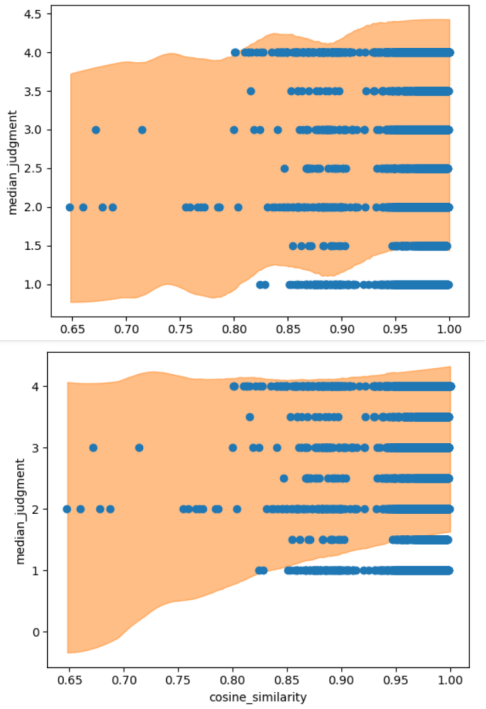

Figure 3: 89% prediction interval of proximity from the Normal distribution. The upper figure is using PCA concatenation vector as predictors, the lower one is using cosine similarity as predictors. Both are projected against cosine similarity

While this does not look too bad for modeling the median judgment, the results measured by Pearson ρ are not as good. There are many ways to interpret this: 1) The disagreement in the DAG (Fig. 2) are not directly caused by the uncertainty of all the step, as there may be further hidden variables that a simple linear model have not captured; 2) The data and metric we used to estimate the disagreement is not a good approximation of the true annotation disagreement, so the results therefore can be improved with better data processing and other formulation of the task; 3) The MCMC model is not a good fit for estimating the disagreement indirectly, and thus we should try other methods that link the posterior distribution to a measurement for disagreement. All of these interpretations can lead us to further directions in the future, but we still need to carefully inspect each of them to draw a better conclusion.

6 Conclusion

In conclusion, this paper presents an initial exploration of applying Bayesian methods to model annotator disagreement in semantic proximity judg-

ments. While the simplicity of the linear model limits the ability to draw definitive conclusions, it lays the groundwork for future research into more sophisticated Bayesian models to understand annotator disagreement in NLP tasks. Moving forward, more complex and nuanced models will be necessary to fully capture the intricacies of annotator judgment variation, offering deeper insights into this challenging aspect of NLP

Contribution statement

As the sole author and contributor to this project, I declare that I have written this work independently and have not used any sources other than those indicated. I have fully identified all parts of this work that are taken from other works in terms of wording, meaning or argumentation (including the World Wide Web and other electronic text and data collections), stating the sources. I have documented every use of AI tools that was related to my work, which only includes a part of the conclusion I get for prompting the abstract to ChatGPT.

References

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 Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 78–89, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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.

Tejaswi Choppa, Michael Roth, and Dominik Schlechtweg. 2025. Predicting median, disagreement and noise label in ordinal word-in-context data. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 65–77, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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*.

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.

Mikhail Kuklin and Nikolay Arefyev. 2025. Deepchange at CoMeDi: the cross-entropy loss is not all you need. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 48–64, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Tai Duc Le and Thin Dang Van. 2025. MMLabUIT at CoMeDiShared task: Text embedding techniques versus generation-based NLI for median judgment classification. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 113–121, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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.

Zhu Liu, Zhen Hu, and Ying Liu. 2024. Comedi shared task: Models as annotators in lexical semantics disagreements. *arXiv preprint arXiv:2411.12147*.

Richard McElreath. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, second edition. Texts in Statistical Science. CRC, Boca Raton, Florida.

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.

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth. 2025. Comedi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 33–47.

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.

Charles Spearman. 1961. The proof and measurement of association between two things. *The American Journal of Psychology*.

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.