University of Stuttgart
Germany

Optimizing Human Annotation of Word Usage Graphs in a Realistic Simulation Environment

December 6, 2021

Serge Kotchourko

Institute for Natural Language Processing, University of Stuttgart
Supervisor: apl. Prof. Dr. Sabine Schulte im Walde
Advisor: Dominik Schlechtweg

Introduction: What is a Word Usage Graph (WUG)?

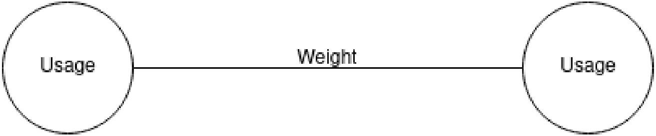

Figure 1: An example for a Word Usage Graph (**WUG**) of an unspecific word.

Introduction: Why and Where?

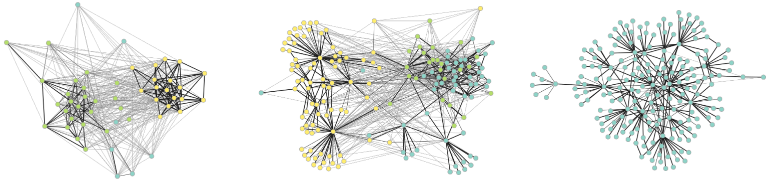

Figure 2: WUGs of *Zehner* from DiscoWUG (left), *abbauen* from DWUG DE (middle) and *bag* from DWUG EN (right).

Problems

- Annotation load grows with number of word usages
  - $|E| = \frac{|N|(|N|-1)}{2} = \frac{|N|^2-|N|}{2}$, where $|N|$ number of word usages and $|E|$ the number of possible pairs
  - Fully annotating even grows quadraticaly
  - Considering human annotator, not feasible for large sets
- Use of human annotators, thus error prone annotations due to (e.g.)
  - ambiguity
  - unknown context (Schlechtweg, Tahmasebi, Hengchen, Dubossarsky, & McGillivray, 2021)
  - non-expert annotators (Schlechtweg, Schulte im Walde, & Eckmann, 2018)

Motivation/Goal

- Building models, consisting of
  - sampling strategy
  - clustering strategy
  - stopping criterion
- and testing these exhaustively on capturing sense structures
  - efficiently, meaning reducing the annotation load
  - effectively, finding good edges and sense structures

Testing Models, but how?

Naive Approach by using models during annotation:

- Time consuming and costly due to human annotators
- Careful planing
- Measure of performance how?

Hence, simulating the full annotation process:

- Generating "ground truth" WUGs
- Simulation of annotation process
- Resulting WUGs evaluated against their "ground truth"

Simulation

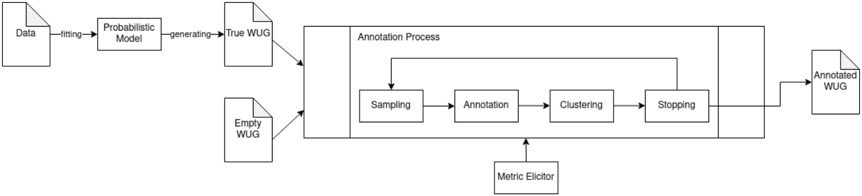

Figure 3: Overview of the complete simulation framework.

Simulation: Data

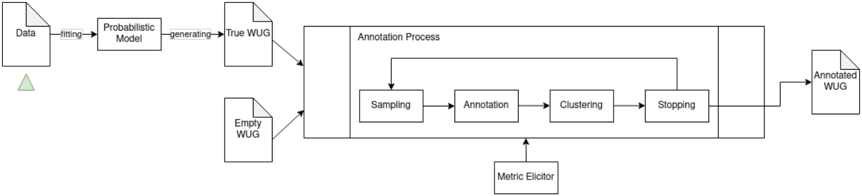

Figure 3: Overview of the complete simulation framework.

Data Used

- DWUG DE/EN[1] (Schlechtweg, Tahmasebi, et al., 2021)
  - Large WUGs
  - Usages sampled randomly from real corpus
  - Two different languages, same model used
  - High amount of annotations
- DiscoWUG[1] (Kurtyigit, Park, Schlechtweg, Kuhn, & im Walde, 2021)
  - Usages sampled randomly from real corpus
  - Exclusion of noisy words
  - Smaller WUGs
  - Randomly sampled word usage pairs

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

Simulation: Generating WUGs

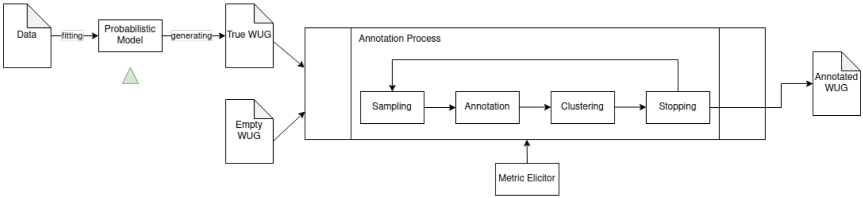

Figure 3: Overview of the complete simulation framework.

Simulation: Generating WUGs, Weighted Stochastic Block Model

- Extension to Stochastic Block Model (Holland, Laskey, & Leinhardt, 1983)
- WSBM is a Generative model for random graphs (Aicher, Jacobs, & Clauset, 2014; Peixoto, 2017)
- Takes three parameters into account:
  - Number and size of clusters
  - Symmetric probability matrix, defining the probability of an edge between clusters
  - Symmetric distribution matrix, defining the observed edge-weight between a pair
- Schlechtweg showed, that it is possible to generate reasonable graphs, modeling WUGs (Schlechtweg, Castaneda, Kuhn, & Schulte im Walde, 2021)

Simulation: True WUGs

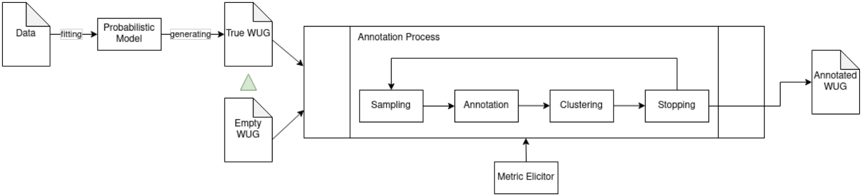

Figure 3: Overview of the complete simulation framework.

Simulation: Annotation Process

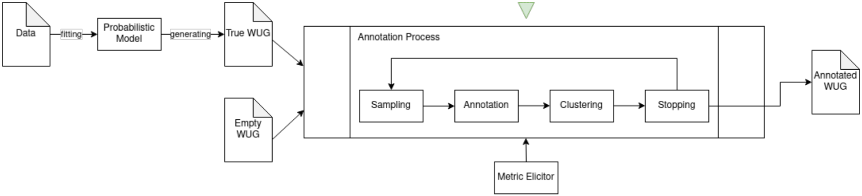

Figure 3: Overview of the complete simulation framework.

Simulation: Sampling

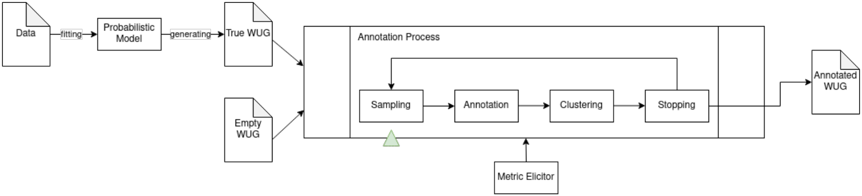

Figure 3: Overview of the complete simulation framework.

Simulation: Sampling, Random Sampling

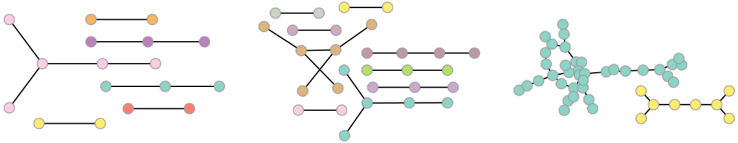

Figure 4: Random Sampling example, where the node-colors represent the connected component the node belongs to.

Simulation: Sampling, Modified Random Walk

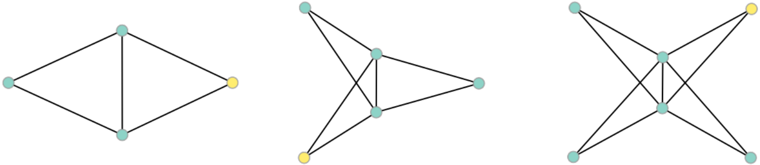

Figure 5: Example of Modified Random Walk sampling steps, illustrating the prioritization of new nodes (yellow) as well as building denser structures.

Simulation: Sampling, DWUG Sampling

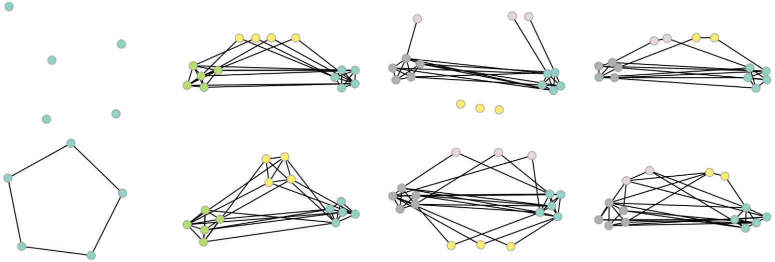

Figure 6: This is an illustration of the different phases of DWUG+RS Sampling strategy. **Left** is the initial seeding. **Middle left** shows the exploration phase performed on the yellow nodes, which do not belong to any cluster bigger than some threshold (green or blue). **Middle right** is an example of the combination phase performed on the purple nodes, which are currently not connected to all clusters bigger than some threshold (grey and blue) and the newly added nodes (yellow). **Right** highlights the intrinsic stopping criterion of DWUG and the random sampling thus performed by DWUG+RS. (Schlechtweg, Tahmasebi, et al., 2021)

Simulation: Annotation

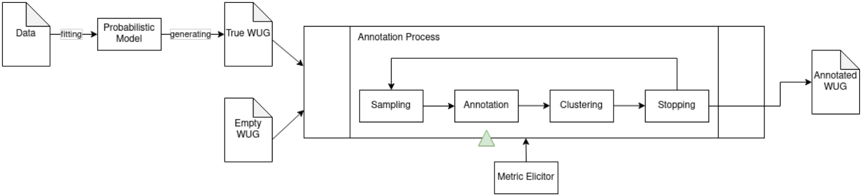

Figure 3: Overview of the complete simulation framework.

Simulation: Clustering

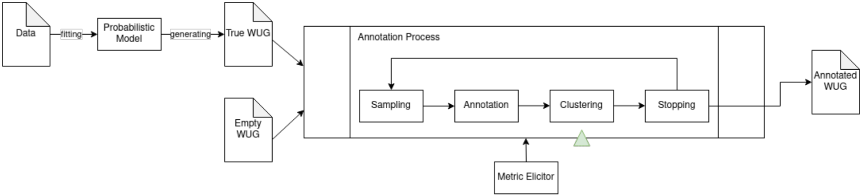

Figure 3: Overview of the complete simulation framework.

Simulation: Clustering, Connected Component Clustering

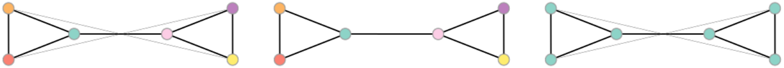

Figure 4: Example of the Connected Component Clustering, showing the individual steps performed. **Left:** Initial WUG. **Middle:** Edge removal step. **Right:** Connected component search. (Hopcroft & Tarjan, 1973)

Simulation: Clustering, Chinese Whispers

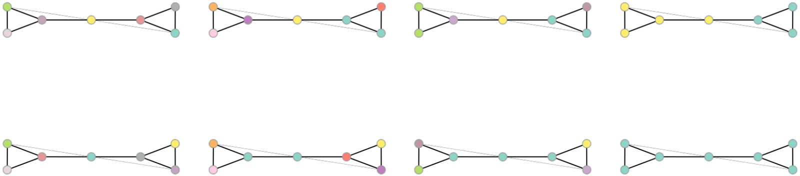

Figure 5: Illustrating two possible clustering (top and bottom) achieved by Chinese Whispers and illustrates how the initial ordering of nodes for the iteration step may impact the resulting clustering. We can also observe that the middle node (yellow) is trapped between two ideal clusters. (Biemann, 2006)

Simulation: Clustering, DWUG Correlation Clustering

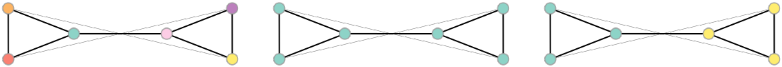

Figure 6: Example of how DWUG Correlation Clustering (**right**) finds a better clustering for a given WUG (**left**) compared to Connected Component Clustering (**middle**). (Bansal et al., 2004; Schlechtweg et al., 2020; Schlechtweg, Tahmasebi, et al., 2021)

Simulation: Stopping

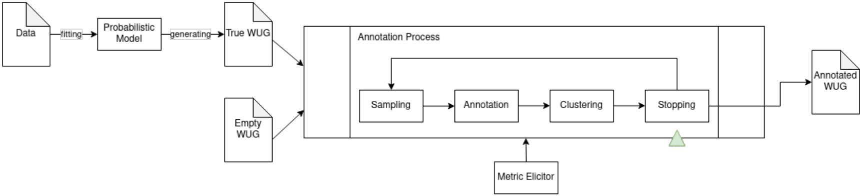

Figure 3: Overview of the complete simulation framework.

Simulation: Stopping, Gambette

Figure 4: An example round of Gambette, where **left** is the initial WUG, **middle** represents the perturbed WUG by the random annotator and **right** is the resulting new clustering for this modified WUG. Based on the left and right WUG the ARI score is calculated.(Gambette & Guénoche, 2011)

Simulation: Evaluation

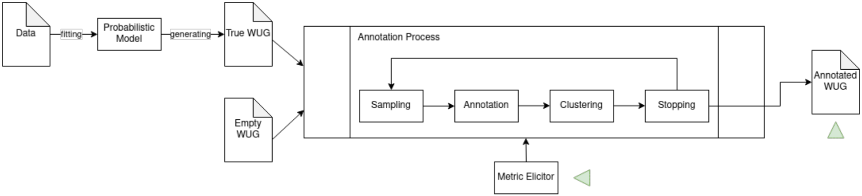

Figure 3: Overview of the complete simulation framework.

Approximation of Observed Data: Number of Clusters

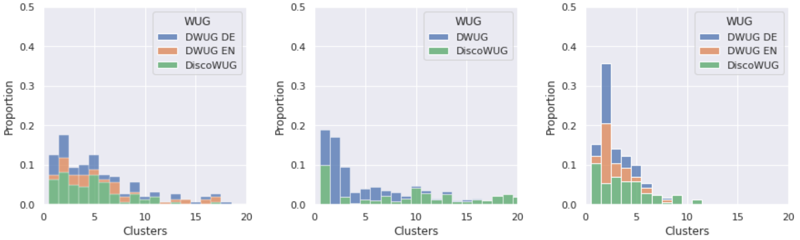

Figure 4: **Left:** Number of clusters (senses) in observed WUGs. **Middle:** Number of clusters for Coarse WUGs. **Right:** Number of clusters for Fitted WUGs, with corresponding models to the data-set.

Approximation of Observed Data: Sense Size distribution

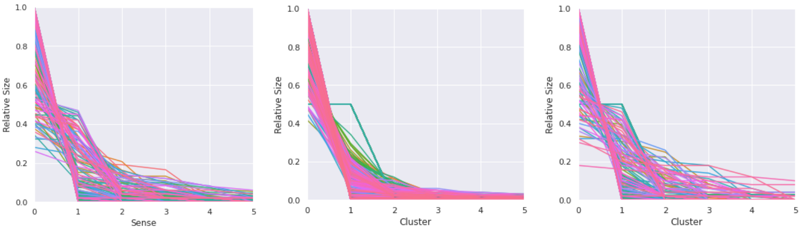

Figure 5: **Left:** Relative sense distribution for single WUGs for observed WUGs. **Middle:** Relative cluster (sense) distribution for single WUGs for Coarse WUGs. **Right:** Relative cluster (sense) distribution for single WUGs for Fitted WUGs.

Approximation of Observed Data: Variance

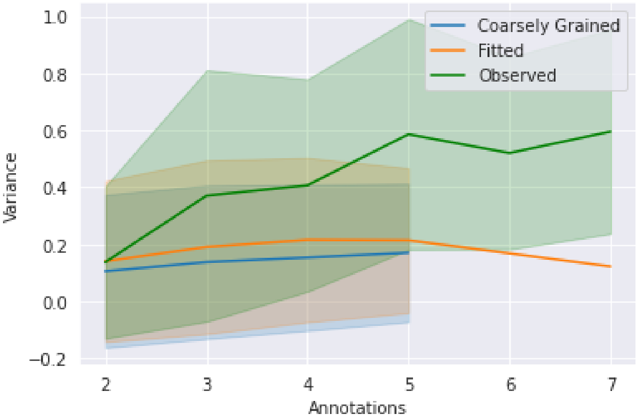

Figure 6: Comparison of median variance of edges per number of annotations of edges.

Approximation of Observed Data: Weight Distribution

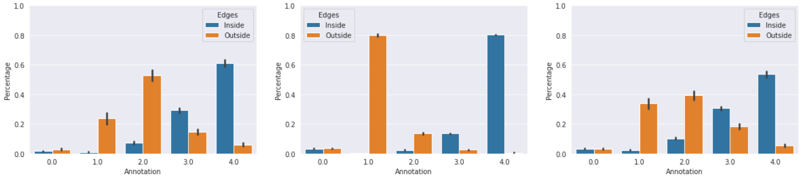

Figure 7: Relative annotation distribution between clusters (senses) for observed WUGs (**Left**), for Coarse WUGs (**Middle**) and for Fitted WUGs (**Right**)

Models on Coarse WUGs: Overview

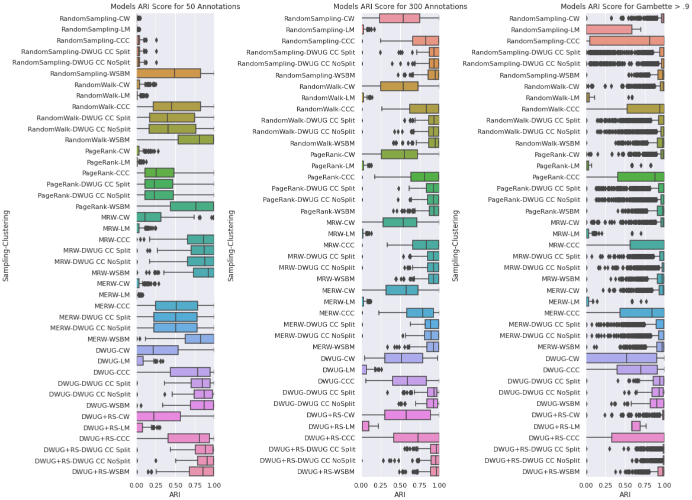

Figure 8: Models on Coarse WUGs in comparison for ARI Score based on the number of annotations with 50 annotations (**left**), with 100 annotations (**Middle**) and with Gambette > .9 and at least 100 annotations.

Models on Fitted WUGs: Overview

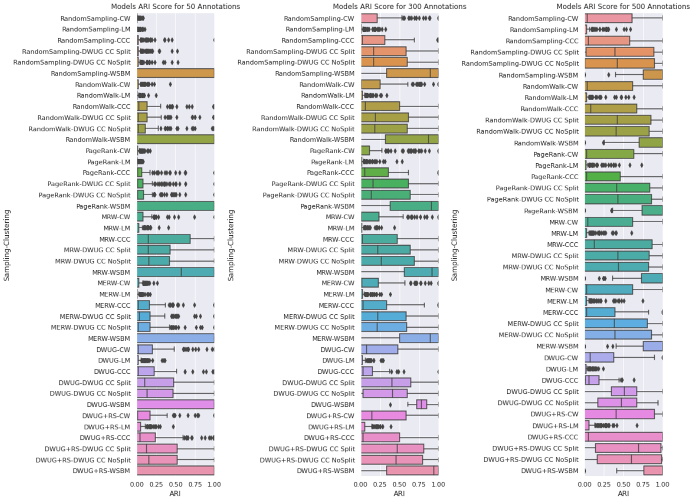

Figure 9: Models on Fitted WUGs in comparison for ARI score based on the number of annotations. **Left** 50, **middle** 300 and **right** 500 annotations.

Conclusion: Goal

The goal was to test different models on their ability to efficiently and effectively find the correct sense assignment.

- Hence: Simulation framework[2]
- Allowing: Extensive and automatic evaluation of any model on infinitely many WUGs
- Which showed: Possibility of generating closely resembling WUGs
- Which showed: Best Performing Models ...
  - Sampling: Modified Random Walk or DWUG
  - Clustering: DWUG Correlation Clustering or Weighted Stochastic Block Model
  - Stopping: Gambette
- Which showed: Performance heavily dependant on components behaviour & underlying data

[2] `https://github.com/confusedSerge/wug_sampling`

Conclusion: Drawbacks & Future

Drawbacks:

- Assumption: Data represents true state of a WUG
- Generative Process/Probabilistic model may not capture observed WUGs fully
- Lack of fully formalizing the annotator (only error & zero)

Future Work:

- Probabilistic model of the whole Annotation Process
- Possibility of using the Generative Process/Simulation as data-set creation
- Models and components as an optimization problem

References I

Aicher, C., Jacobs, A. Z., & Clauset, A. (2014, Jun). Learning latent block structure in weighted networks. *Journal of Complex Networks*, *3*(2), 221—248. Retrieved from http://dx.doi.org/10.1093/comnet/cnu026 doi: 10.1093/comnet/cnu026

Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, *56*(1-3), 89–113. doi: 10.1023/B:MACH.0000033116.57574.95

Biemann, C. (2006, June). Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the first workshop on graph based methods for natural language processing* (pp. 73–80). New York City: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W06-3812

Gambette, P., & Guénoche, A. (2011). Bootstrap clustering for graph partitioning. *RAIRO - Operations Research - Recherche Opérationnelle*, *45*(4), 339-352. doi: 10.1051/ro/2012001

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, *5*(2), 109 - 137.

Hopcroft, J., & Tarjan, R. (1973, June). Algorithm 447: Efficient algorithms for graph manipulation. *Commun. ACM*, *16*(6), 372–378. Retrieved from https://doi.org/10.1145/362248.362272 doi: 10.1145/362248.362272

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., & im Walde, S. S. (2021). *Lexical semantic change discovery.*

Peixoto, T. P. (2017, 08). Nonparametric weighted stochastic block models. *Physical Review E*, *97*. doi: 10.1103/PhysRevE.97.012306

Schlechtweg, D., Castaneda, E., Kuhn, J., & Schulte im Walde, S. (2021, August). Modeling sense structure in word usage graphs with the weighted stochastic block model. In *Proceedings of *sem 2021: The tenth joint conference on lexical and computational semantics* (pp. 241–251). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.starsem-1.23 doi: 10.18653/v1/2021.starsem-1.23

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (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. Retrieved from https://www.aclweb.org/anthology/2020.semeval-1.1/

References II

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

Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., & McGillivray, B. (2021). DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. *CoRR*, *abs/2104.08540*. Retrieved from `https://arxiv.org/abs/2104.08540`