Similarity-Based Cluster Merging for Semantic Change Modeling

Christopher Brückner[1] and **Leixin Zhang**[2] and **Pavel Pecina**[1]
[1]Charles University, Faculty of Mathematics and Physics
{bruckner,pecina}@ufal.mff.cuni.cz
[2]University of Twente
l.zhang-5@utwente.nl

Abstract

This paper describes our contribution to Subtask 1 of the AXOLOTL-24 Shared Task on unsupervised lexical semantic change modeling. In a joint task of word sense disambiguation and word sense induction on diachronic corpora, we significantly outperform the baseline by merging clusters of modern usage examples based on their similarities with the same historical word sense as well as their mutual similarities. We observe that multilingual sentence embeddings outperform language-specific ones in this task.

1 Introduction

Semantic change modeling is the task of computationally determining how the meanings of words change over time. This semantic shift can be observed in the change of contexts in which the words appear (Kutuzov et al., 2018).

Given a diachronic corpus of old and new word usage examples and an inventory of old word senses with their dictionary definitions, the modeling task can be split further into the disambiguation and the induction of word senses: New usage examples are aligned with old usage examples and sense definitions. If an appropriate old sense does not exist in the sense inventory and an alignment is thus impossible, a novel word sense is induced instead, indicating that the word gained a new meaning.

This joint task has been defined by Subtask 1 of the AXOLOTL-24 Shared Task (Fedorova et al., 2024), our contribution to which we describe in the following sections. Like the baseline proposed by the shared task organizers, we approach the challenge by measuring the similarity of modern usage example clusters and old sense definitions. We further explore impacts on the performance by merging clusters based on a similarity criterion and by ensembling different embedding models and different clusterings.

Our implementation is available on GitHub.[1]

2 Related Work

The idea of unsupervised clustering to discriminate word senses goes back at least to using Gaussian Mixture models on a synchronic corpus (Schütze, 1998). More recently, neural approaches have been applied to diachronic corpora to detect and quantify semantic change (Kutuzov et al., 2018).

SemEval-2020 Task 1 produced several approaches for lexical semantic change detection between two time-specific corpora (Schlechtweg et al., 2020). The task was split into the binary classification of whether words lost or gained senses, and the ranking of words according to their degree of change. These sub-tasks were solved, e.g., by clustering contextual word embeddings and comparing their cluster assignments (Karnysheva and Schwarz, 2020), or by measuring the average cosine distances between contextual embeddings of the same word (Kutuzov and Giulianelli, 2020). In contrast with AXOLOTL-24, this task considers whether an old word sense is still present in the new time period.

Another task more similar to AXOLOTL-24 was defined by the *Reverse Dictonary* track of SemEval-2022 Task 1 (Mickus et al., 2022): Given a dictionary consisting of words, their definitions, and definition embeddings, user-written definitions are to be mapped to the correct word by reconstructing the reference embedding. As these embeddings were pre-computed, submitted systems were limited to three specific models. While participating teams achieved reasonable average cosine similarities using token-level transformers, this was not evaluated as a classification task.

The Sentence-BERT architecture promises better performance than token-level transformers on sentence-level downstream tasks such as paraphras-

[1]https://github.com/chbridges/axolotl24

ing and the measurement of sentence similarities (Reimers and Gurevych, 2019). While the original publication proposes a model for paraphrasing in over 50 languages based on MPNet (Song et al., 2020), the general-purpose LaBSE doubles the size of the language inventory and suggests current state-of-the-art performance in cross-lingual settings (Feng et al., 2022).

3 Datasets and Task Definition

The AXOLOTL-24 Shared Task provides training corpora in Finnish and Russian. The Finnish corpus covers the years 1543 to 1650 in its old time period and the years 1700 to 1750 in its new period, whereas the Russian corpus covers approximately the 19th century and the years after 1950. Both datasets consist of different target words with multiple word senses, and each sense comprises a sense ID, a definition, and a usage example. In the case of Russian, usage examples of old words are often noisy or missing.

The goal of Subtask 1 is to determine the correct sense IDs of word usage examples in the new period. Thus, the corresponding test datasets only contain sense IDs and definitions in the old period. Subtask 2, the generation of novel sense definitions, is out of the scope of this paper. In addition to the Finnish and Russian test datasets, a third, German test set based on the DWUG dataset (Schlechtweg, 2023) is provided to quantify the developed systems' multilingual performance.

Systems are evaluated with respect to word sense disambiguation and the joint task including the induction of novel word senses. System performance on the disambiguation task is measured with the macro F1 score of sense classifications only of sense IDs present in the old sense inventory. Additionally, the overall performance is measured with the adjusted Rand index, thus ignoring specific sense assignments but validating whether modern usage examples of old and novel word senses are correctly grouped together.

4 Methodology

In this section, we briefly summarize the baseline algorithm before describing our improvements.

4.1 Baseline

The general approach can be divided into two steps: the embedding of old word sense definitions and modern usage examples, and the alignment of the respective embeddings to assign word senses to the examples. These steps are computed target word by target word, i.e., no defined sense of a different word can leak into the assignment.

In the first step, the sense definitions and usage examples from the old time period are concatenated. These concatenations and all usage examples from the new time period are then embedded in a shared vector space by a transformer model.

In the second step, the new usage examples are clustered. For each cluster C, an old word sense s_{old} is assigned to all corresponding usage examples if the cosine similarity $\cos(\mathbf{s}_{old}, \mathbf{c})$ between the old sense embedding $\mathbf{s}_{old}$ and a cluster embedding $\mathbf{c}$ is greater than a threshold $\tau \in [0, 1]$. If no old sense satisfies this condition, a new sense s_{novel} is assigned to all usage examples in the cluster.

This final step is solved in a greedy manner: Once an old sense is assigned to a cluster C_i, it is removed from a list of candidate senses S and cannot be assigned to any other cluster C_j, even if $\cos(\mathbf{s}_i, \mathbf{c}_j) > \cos(\mathbf{s}_i, \mathbf{c}_i)$. This poses a problem if word senses are split into multiple clusters.

The following subsections propose methods to alleviate this weakness of the baseline approach. The impact of each described method on the system performance is summarized in Section 5.

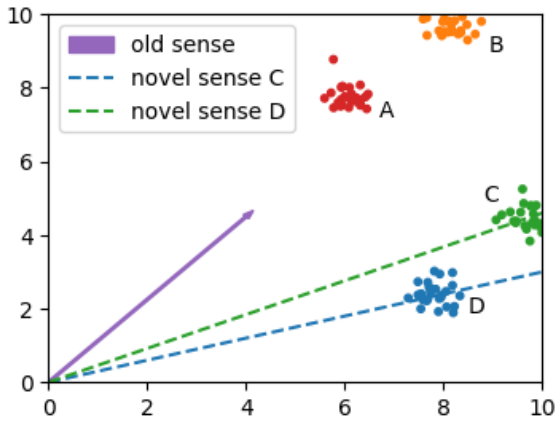

Figure 1: A conceptual cluster merging. Clusters A and B get merged, as the angles of their centers with the old sense vector are small. Novel senses are generated for clusters C and D. In the second pass, novel sense vectors are fitted through C and D, merging them if the angle between these vectors is sufficiently small.

4.2 Cluster Merging

A straightforward technique to improve the alignment of old senses and new usage examples is to keep the set of candidate senses S fixed and assign

Model	Finnish ARI	Finnish F_1	Russian ARI	Russian F_1
Baseline	*0.022*	*0.222*	**0.098**	*0.274*
Merge-1	0.420	0.557	*0.052*	0.428
Merge-5	0.420	0.557	0.058	0.428
Merge-1c	**0.437**	**0.570**	0.071	0.447
Merge-5c	**0.437**	**0.570**	0.077	**0.449**

Table 1: Development scores at a fixed similarity threshold $\tau = 0.3$. Where Affinity Propagation is used, the model name indicates the number of ensembled clusterings and the usage of cosine similarity affinity. Highest scores are indicated in bold, lowest scores in italics.

Embedding	Finnish ARI	Finnish F_1	Russian ARI	Russian F_1
LEALLA-large	0.437	0.570	0.077	*0.449*
LaBSE	*0.277*	*0.462*	0.081	0.572
Finnish-Paraphrase	0.561	**0.676**	—	—
Sentence RuBERT	—	—	*0.056*	0.608
Multi-Paraphrase	0.554	0.661	0.118	**0.612**
Multi $\oplus$ LaBSE	**0.572**	0.669	**0.120**	0.603

Table 2: Development scores at a fixed similarity threshold $\tau = 0.3$ for different sentence embeddings, based on the best models in Table 1. Highest scores are indicated in bold, lowest scores in italics.

each cluster C to the sense with the greatest similarity, provided that the similarity is greater than the previously chosen threshold τ. The similarity is computed between the sense embedding $\mathbf{s}$ and the cluster mean $\bar{\mathbf{c}}$ to capture the overall semantics of C. Thus, the sense alignment step is defined as:

$$s_C = \begin{cases} \underset{s \in S}{\operatorname{argmax}} \cos(\mathbf{s}, \bar{\mathbf{c}}), & \cos(\cdot) \geq \tau \\ s_{novel}, & otherwise \end{cases} \quad (1)$$

As each old sense can now be mapped to multiple clusters, this alignment is equivalent to the merging of clusters when their similarities with the same old sense are sufficiently large. This reduces the granularity of old sense clusters. In a second pass, each novel sense cluster center is considered a novel sense embedding and novel sense clusters are merged by the same criterion based on pairwise cosine similarities. A conceptual such merging in two dimensions is depicted in Figure 1.

4.3 Two-Stage Ensembling

In addition to merging clusters with respect to the similarity with the same old sense, we propose two methods to ensemble results at different stages of the algorithm: The ensembling of embedding models and the ensembling of clusterings.

The ensembling of n models is straightforward and can be solved via the concatenation

$$\mathbf{e} = \mathbf{e}_1 \oplus \ldots \oplus \mathbf{e}_n \quad (2)$$

of each model output $\mathbf{e}_i$ which is then used as the input to the alignment step of the algorithm.

A crucial part of the alignment step is the clustering of modern usage examples. Some clustering algorithms such as K-means (Lloyd, 1982) and Affinity Propagation (Frey and Dueck, 2007) are initialized using a random seed r based on which they can converge to different local minima. We mitigate resulting errors by clustering the input embeddings multiple times using different random seeds r_i and selecting the final cluster assignments via a majority vote. Reproducibility is ensured by fixing the initial random seed r_0 and incrementing it for the subsequent clusterings, i.e., $r_i = r_0 + i$.

5 Results and Discussion

The AXOLOTL-24 baseline uses LEALLA-large (Mao and Nakagawa, 2023) in the embedding step, a lightweight language-agnostic sentence transformer distilled from LaBSE (Feng et al., 2022), and clusters these embeddings with Affinity Propagation (Frey and Dueck, 2007) using the negative Euclidean distance as the cluster affinity. We begin our study by comparing the baseline with our approach based on LaBSE embeddings on the Finnish and Russian development sets in Table 1. We generally prioritize the ARI since the F_1 score only quantifies the classification of old word senses. While the cluster merging significantly improves the ARI and F_1 score for Finnish, there is a slight trade-off between them in the Russian dataset where a greatly increased F_1 score comes at the cost of a decreased ARI. The ensembling of clusterings does not affect Finnish but leads to better results for Russian. The scores further increase when using the cosine similarity as the cluster affinity.

We further evaluate additional language-specific and language-agnostic sentence embeddings from the Hugging Face Hub in Table 2: a Finnish paraphrasing model[2] (Kanerva et al., 2021), Sentence RuBERT[3] (Kuratov and Arkhipov, 2019), and a multilingual paraphrasing model[4] (Reimers and Gurevych, 2019). Interestingly, we observe that multilingual models can outperform language-

[2] TurkuNLP/sbert-cased-finnish-paraphrase

[3] DeepPavlov/rubert-base-cased-sentence

[4] sentence-transformers/paraphrase-multilingual-mpnet-base-v2

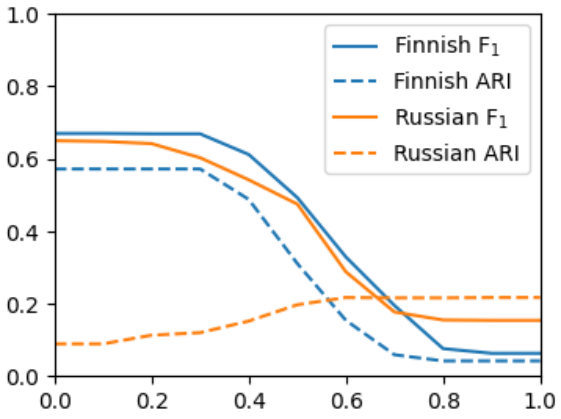

Figure 2: Threshold analysis on the development sets of both languages. ARI and F1 on the y-axis are mapped against different similarity thresholds τ on the x-axis.

specific ones, in particular, a concatenation of the multilingual paraphrasing model and LaBSE. Thus, we consider this embedding to generalize best and choose it for further experiments. We do not observe any improvement when concatenating a third model.

Next, we analyze how well the system performs for different similarity thresholds τ in Figure 2. It shows different behavior for the two datasets: Increasing τ leads to a decreasing F_1 in both languages but to a decreasing ARI for Finnish and an increasing ARI for Russian. This indicates that initial old sense clusters are too granular in both languages, whereas old and novel sense clusters are less discriminative in Russian as novel senses tend to get merged into old ones, reducing the overall clustering quality when the similarity threshold is set too small. Thus, increasing τ increases the proportion of granular novel sense clusters to coarse-grained old sense clusters. We find that the preset threshold $\tau = 0.3$ used by the baseline and our previous experiments is a reasonable choice, as the Finnish scores are stable up to this value. For Russian, a slightly higher threshold of $\tau = 0.4$ or $\tau = 0.5$ might be preferred to account for a greater ARI without sacrificing too much F_1. We choose τ for Finnish and Russian based on this graph but suggest cross-validation as a more robust method to choose the parameter for unseen data. For German, which has no development set, we select the same parameter as for Finnish since the different performance on Russian can possibly be attributed to its often noisy or missing word usage examples in the old period.

Finally, for further analysis, we skip the clustering step and assign a single most probable old sense to each target word. The result is surprising: While we achieve an ARI of merely 0.015 on Russian, we outperform our method on Finnish with an ARI of 0.614 and an F_1 score of 0.680. We attribute this anomaly to the quality of the dataset, as the numbers of senses per target word and usage examples per word sense are imbalanced, including several words with only one sense. However, this characteristic also reveals a weakness in our algorithm: Clusters are often aligned with an old sense in the first pass even though the word has no documented old sense. Possible improvements are a combination of both passes into one or the usage of two different similarity thresholds for old and novel senses.

ARI			
System	Finnish	Russian	German
Baseline	0.023	0.079	0.022
deep-change	0.638	0.059	0.543
Ours (old)	0.596	0.043	0.298
Ours (new)	0.578	0.130	0.298
F_1			
System	Finnish	Russian	German
Baseline	0.230	0.260	0.130
deep-change	0.756	0.750	0.745
Ours (old)	0.655	0.661	0.608
Ours (new)	0.655	0.563	0.608

Table 3: ARI and F_1 scores of the baseline, the winning team deep-change, our submission to the shared task, and our updated system on the three test datasets.

Our final results are summarized in Table 3. In our submission to the shared task, we tuned the similarity thresholds less carefully, used $\tau = 0.1$ for all three test sets, and did not cluster the Finnish dataset. The new system uses $\tau = 0.2$ for Finnish and German, and $\tau = 0.45$ for Russian. It does not affect our ranking on the leaderboards.

6 Conclusion

We presented a simple method to discriminate word senses on diachronic corpora by clustering usage examples and merging the resulting clusters if either their similarity with a known word sense or their mutual similarities are sufficiently large. It depends on a similarity threshold τ that can be tuned on annotated data. The resulting system performs best when embedding usage examples and word sense definitions with two different multilingual models and thus adapts well to different languages.

However, there is room for improvement. For the proposed algorithm, we suggest the usage of

two different similarity thresholds for old and novel sense cluster merging. We further see a weakness in prioritizing the disambiguation of old word senses while solving the induction of novel word senses as a subsequent step.

We support the publication of a similar, better-normalized dataset for improved comparability between languages.

Limitations

The AXOLOTL-24 Shared Task takes a step from the pure quantification of semantic change to more interpretable results by assigning concrete word senses to groups of word usage examples and simultaneously identifying word usages with no recorded definition. The presented results do not go beyond the scope of this shared task. There may be limitations in the comparability between languages due to significant amounts of noise and imbalance in the provided dataset. Furthermore, the evaluation does not take the absence of recorded word senses in the new period into account and thus does not consider the full spectrum of semantic change observable in the data. These aspects should be investigated further in future research.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was partially supported by SVV project number 260 698 and Horizon Europe grant agreement number 101061016.

References

Mariia Fedorova, Timothee Mickus, Niko Tapio Partanen, Janine Siewert, Elena Spaziani, and Andrey Kutuzov. 2024. AXOLOTL'24 shared task on multilingual explainable semantic change modeling. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, Bangkok. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972–976.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Jenna Saarni, Maija Sevón, and Otto Tarkka. 2021. Finnish paraphrase corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 288–298, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Anna Karnysheva and Pia Schwarz. 2020. TUE at SemEval-2020 task 1: Detecting semantic change by clustering contextual word embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 232–238, Barcelona (online). International Committee for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *Preprint*, arXiv:1905.07213.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

S. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1886–1894, Dubrovnik, Croatia. Association for Computational Linguistics.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. 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.

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

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 Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

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

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.