# TartuNLP @ AXOLOTL-24: Leveraging Classifier Output for New Sense Detection in Lexical Semantics

**Aleksei Dorkin** and **Kairit Sirts**
Institute of Computer Science
University of Tartu
{aleksei.dorkin, kairit.sirts}@ut.ee

## Abstract

We present our submission to the AXOLOTL-24 shared task. The shared task comprises two subtasks: identifying new senses that words gain with time (when comparing newer and older time periods) and producing the definitions for the identified new senses. We implemented a conceptually simple and computationally inexpensive solution to both subtasks. We trained adapter-based binary classification models to match glosses with usage examples and leveraged the probability output of the models to identify novel senses. The same models were used to match examples of novel sense usages with Wiktionary definitions. Our submission attained third place on the first subtask and the first place on the second subtask.

| Team | ARI | F1 |
|---|---|---|
| deep-change | **0.413** | **0.750** |
| Holotniekat | 0.312 | 0.641 |
| *TartuNLP (ours)* | 0.310 | 0.590 |
| IMS_Stuttgart | 0.287 | 0.487 |
| ABDN-NLP | 0.221 | 0.431 |
| WooperNLP | 0.187 | 0.316 |
| Baseline | 0.041 | 0.207 |

Table 1: Overall results on the Subtask 1.

| Team | Overall | BLEU | BERTScore |
|---|---|---|---|
| *TartuNLP (ours)* | **0.467** | **0.208** | **0.726** |
| WooperNLP | 0.340 | 0.020 | 0.660 |
| ABDN-NLP | 0.253 | 0.045 | 0.461 |
| baseline | 0.218 | 0.013 | 0.423 |

Table 2: Overall results on the Subtask 2.

## 1 Introduction

The subject of the AXOLOTL-24 shared task (Fedorova et al., 2024) is diachronic semantic change detection and explanation. Diachronic semantic change is understood as the change in word meanings (i.e., words losing old senses and obtaining new ones) over shorter or longer periods. Accordingly, given a dataset containing usage examples from different periods (old and new), the task is to identify and define the new senses that words gain in the new time period compared to the old one.

The goal of the shared task is to implement a semantic change modeling system for two tasks:

1) Correctly assigning existing senses to target word usages and identifying novel, previously unseen senses;
2) Describing the identified novel senses.

The data in the shared task is provided in three languages: Finnish, Russian, and German (the surprise language for which only the test split is available). For each language, examples from old and new periods are given. Each data point consists of a target word and its usage example, gloss (target word definition), the period the example comes from, and usage and sense IDs. The data includes glosses for both time periods in the training and validation splits, while glosses for the new time period are not provided in the test splits. The “old” and “new” periods differ for each language. For Finnish, old texts are dated before 1700, and new ones are dated after 1700. For Russian, the old target word usages are from the 19th century, and the new data represents modern usages of words. For German, the old period is from 1800 to 1899, and the new period is from 1946 to 1990 (Schlechtweg, 2023).

Although we participated in both subtasks, we were primarily interested in the second subtask of producing definitions for new senses. We implemented a solution that matches identified novel sense usages with definitions from an external resource (Wiktionary). Our approach is based on a binary classification task to predict whether a proposed definition matches the sense under consideration. We reused this binary classification model for the second subtask of describing the identified novel senses.

Our system attained the first place on the second subtask (Table 2) and obtained competitive results on the first subtask (Table 1).

## 2 Methodology

We propose a simple classification-based solution for both subtasks. We adopt the GlossBERT approach (Huang et al., 2019) that treats word sense disambiguation as a sentence pair classification task, where each pair comprises a usage example and a sense definition. In turn, we frame the problem of new sense identification as the problem of matching between usage examples and sense definitions. Accordingly, the matching problem can be solved with a binary classification model that, given a usage example and a sense definition, outputs the probability of the sense definition correctly describing the usage example.

We adopt the cross-encoder model that simultaneously processes the usage examples and the sense definitions with the same model. Given a usage example and a sense inventory, we apply the classification model to predict binary probabilities for each example/sense definition combination. If the highest probability over all candidate pairs exceeds a predefined threshold, the system assigns the highest probability sense to the usage example. Otherwise, the sense used in the example is deemed to be new.

### 2.1 Subtask 1: Bridging Diachronic Word Uses and a Synchronic Dictionary

This subtask aims to assign a sense ID to every usage example from the new period; the sense ID may come either from the senses in the old period or, if the system identifies a novel sense, a unique new sense ID is created.

The data for the first subtask contains sense definitions and correct usages, which can be used to construct positive examples for our task formulation. However, having only positive examples for a classification model is generally insufficient. To produce negative examples, we employ a simple algorithm. We only consider words associated with at least two distinct sense IDs. For a given sense ID and its associated gloss, we create all possible combinations of the gloss with the usage examples associated with the other sense IDs of the same word (consult Appendix A for additional details). We expect that the negatives obtained with this algorithm are hard and, as such, are more useful for training that could be obtained, for instance, via random sampling.

We transform every split of every language by extending it with negative examples created with the procedure described above. For each language, we train a separate classification model on the train split and evaluate on the development split. When training and evaluating the classifier, we do not consider the period (old or new) from which the examples come. The best checkpoint for each model is selected based on the development F1 score.

Having trained the classification models, we perform inference on the test set and transform the output into the expected format. During inference, the usage examples from the old period are ignored and the classification is performed only on pairs of usage examples from the new period and sense definitions from the old period. If the highest predicted probability for a usage example is above a threshold, we assign the sense ID of the most probable sense definition to the usage example. Otherwise, a new sense ID is created. The final result submitted for evaluation contains both the predicted senses for the examples from the new period as well as the positive examples in the test split from the old period.[1]

For the surprise language—German—the training process is slightly different. No training or validation data is provided, so we train and validate the classification model on the positive and negative examples obtained from the old period in the test data. Inference, however, is exactly the same.

### 2.2 Subtask 2: Definition Generation for Novel Word Senses

Subtask 2 aims to define each novel sense identified in the first subtask. Despite the name of the subtask, our approach does not generate any new definitions. We also do not train any additional models. As previously mentioned, we consider this task a matching problem, except that the definitions for the novel senses are not present in the data provided in the shared task. To solve this problem, we scrape the definitions of the surface forms, for which we identified at least one example as the usage of a novel sense, from the Wiktionary. More specifically, we scrape the definitions from

[1] Since the test examples for the old period were already annotated, we simply copied their sense definitions to the submitted result file.

the language-specific Wiktionary versions for each language (i.e. Finnish,[2] Russian,[3] and German[4] Wiktionaries).

Having scraped the necessary definitions, we head straight to inference on the test set. We reuse the models and predictions from the first subtask. We collect the examples identified as the usages of the novel senses from the predictions and match them with the Wiktionary definitions using the classifier models trained in the first subtask. After that, we add the matched definitions to the predicted new senses.

### 2.3 Implementation Details

Different from GlossBERT (Huang et al., 2019), which is based on the BERT (Devlin et al., 2019) model, we instead use XLM-RoBERTa (Conneau et al., 2020) as the base model for our classifiers. XLM-RoBERTa is a multilingual model that includes Finnish, Russian, and German in its training data. We expect our system to benefit from the multilinguality. Instead of full fine-tuning, we opt for parameter-efficient fine-tuning. More specifically, we train bottleneck adapter (Houlsby et al., 2019) classifiers for each language. We adopted this approach because it makes our solution computationally lightweight and easily reproducible.

In GlossBERT, Huang et al. (2019) differentiate between training setups with and without weak supervision, with the former including the defined word itself in the gloss, as well as highlighting it in the usage example. According to the experimental results reported by Huang et al. (2019), weak supervision appears to bring minor improvements in sense prediction. However, we do not use weak supervision in our submission. The reason is that Finnish and Russian are substantially more morphologically rich languages than English; thus, the target words rarely appear in their dictionary forms in the usage examples. Moreover, in some cases, the orthography also differs between old and new periods.

To delimit context and gloss, Huang et al. (2019) use the special [SEP] token that is pre-trained into the BERT model via the next sentence prediction task. However, RoBERTa (Zhuang et al., 2021), and by extension XLM-RoBERTa, omitted the next sentence prediction task in the pre-training. As a result of that, the </s> token that is used by RoBERTa in place of [SEP] does not have the same classification-oriented meaning. For this reason, we employed the tabulation symbol as the delimiter instead.

For each language, we employed different variations of the base model and varying training setups. For Finnish, we used the large version of XLM-RoBERTa and trained for ten epochs in half-precision with a batch size of 128 and 3 steps of gradient accumulation. We observed that the training did not converge with a smaller effective batch size. For Russian, we trained the classifier adapter with the base version of XLM-RoBERTa for 50 epochs with a batch size of 144. We also experimented with the large version of the model for the Russian language; however, it showed no improvements compared to the base version. For German, we did not train the classifier from scratch. Instead, we continued training from the best checkpoint trained on the Finnish data. The motivation is that there is considerably more data in Finnish than in Russian in the shared task, so we assume the Finnish model to be stronger. We continued training the Finnish classifier for 20 epochs in half-precision with a batch size of 48 and 6 steps of gradient accumulation. All models were trained with a 5e-4 learning rate.

The threshold value for the classifier's probability to identify novel senses was selected as the highest scoring option in the first subtask using the evaluation script provided by the organizers. We tested a small number of values in the range of of 0.2 to 0.5 on Russian and determined the best value to be **0.35**. The same value was used for all languages without additional testing due to time limitations.

The models were trained on the University High-Performance Cluster (University of Tartu, 2018). We used a single Tesla V100 GPU for Russian and German, while for Finnish, we used a single A100 80GB GPU. The time elapsed on training is 9 hours for Finnish, 3 hours for Russian, and 9 minutes for German. We implemented our solution using the transformers[5] and the adapters[6] libraries. The source code and the data are available on GitHub[7] and HuggingFace Hub,[8] respectively.

[2] https://fi.wiktionary.org/

[3] https://ru.wiktionary.org/

[4] https://de.wiktionary.org/

[5] https://github.com/huggingface/transformers

[6] https://github.com/adapter-hub/adapters

[7] https://github.com/slowwavesleep/ancient-lang-adapters/tree/axolotl

[8] https://huggingface.co/datasets/adorkin/axolotl-wiktionary-definitions

| Team | Fi-BLEU | Ru-BLEU | De-BLEU | Fi-BERTScore | Ru-BERTScore | De-BERTScore |
|---|---|---|---|---|---|---|
| *TartuNLP (ours)* | 0.028 | **0.587** | **0.01** | 0.679 | **0.869** | 0.63 |
| WooperNLP | 0.023 | 0.027 | **0.01** | 0.675 | 0.656 | **0.65** |
| ABDN-NLP | **0.107** | 0.027 | 0.0 | **0.706** | 0.677 | 0.0 |
| baseline | 0.033 | 0.005 | 0.0 | 0.403 | 0.377 | 0.49 |

Table 3: Language specific results for the Subtask 2.

## 3 Results

The overall results of both subtasks are presented in Tables 1 and 2. For subtask 1, the metrics reported are the average macro-F1 score and the average Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) across target words per language. The overall F1 and ARI scores are computed as the mean across all languages. For subtask 2, the evaluation metrics are the BERTScore (Zhang et al., 2020) and BLEU (Papineni et al., 2002) averaged across target words per language. BLEU and BERTScore values for the entire subtask are the respective averages across all languages. The overall score is the mean of BLEU and BERTScore.

Our submission attained the third place out of eight participants in the first subtask and the first place out of four participants in the second subtask (Table 2). This aligns with our expectations since we focused on the second subtask from the beginning and applied the system developed for the second subtask to the first subtask. When looking at the language-specific measures of subtask 2 (Table 3), one can see considerable differences between languages. Our system works the best in Russian while also performing well in German in terms of BERTScore (although the BLEU score is close to 0 for all systems). In Finnish, our system is competitive in terms of BERTScore but underperforms compared to the baseline in terms of BLEU.

## 4 Discussion

Our submission to the second subtask is well ahead of the other participants in the overall leaderboard (Table 2) despite the simplicity of our approach. However, the language-specific results show that it is not so clear-cut (Table 3). Some of the success can be attributed to accidentally matching the source of definitions for the Russian language, which is the Russian Wiktionary. We believe so because the value of the BLEU metric of our submission in the Russian language is higher than that of the other teams and in the other languages by an order of magnitude. However, we do not consider this a critical issue because the BERTScore metric is reasonably high and well above the baseline for all languages, suggesting that the matched definitions capture the expected senses well. However, the corresponding low BLEU scores highlight the inadequacy of the BLEU metric for this task.

| **Wiktionary language** | **Number of unique pages** |
|---|---|
| Finnish | 586,439 |
| German | 1,314,597 |
| Russian | 2,877,010 |

Table 4: The number of unique Wiktionary pages per language.

Secondly, our approach to the second subtask has limitations. More specifically, matching usage examples only against the definitions of the target word, while efficient, considerably limits the system's ability to describe completely new senses. Intuitively, a definition associated with a different word may be a more suitable description of a new sense. A more robust solution would involve matching usage examples against all available definitions. However, that would likely require using a bi-encoder architecture (as proposed by Blevins and Zettlemoyer (2020), for instance) instead of a cross-encoder due to the computational complexity of matching every example with every definition.

Accessing the definitions for all the words in a given language-specific Wiktionary is time-consuming because the layout, article structure, and templates used are completely different for each Wiktionary version. While there is a resource providing Wiktionary dumps in a much more convenient format,[9] it is mostly limited in its support to the English language, with the support for some languages, such as Russian and German, being work in progress, and for others, such as Finnish, completely missing at the time of writing. Moreover, the Finnish, German, and Russian Wiktionaries differ in size and the fullness of their coverage. A rough estimate can be made by accessing the **Special:Statistics** page of each Wiktionary and examining the total number of unique pages (Table 4).

[9] https://kaikki.org/

We note the correlation between the smaller sizes of the Finnish and German Wiktionaries and the lower performance of our system on these languages.

Lastly, although we did not focus on the first subtask, we believe the results of the sense prediction task obtained with our systems could also be improved. For instance, the choice of the threshold value for determining a new sense could be done in a more systematic manner or made learnable. Similarly, adjusting the training data or the hyperparameters might bring further improvements.

## 5 Conclusion

This paper described our solution to both subtasks of the AXOLOTL-24 shared task based on leveraging classifier probabilities for usage example/sense definition pairs. The developed system is conceptually simple, adopting a binary classification approach to predict the probability of a sense definition matching the usage example and employing the adapters framework to reduce computation resource requirements. Our submission attained the third place in the first subtask and the first place in the second subtask, showing the feasibility of our approach.

## Acknowledgements

This research was supported by the Estonian Research Council Grant PSG721.

## References

Terra Blevins and Luke Zettlemoyer. 2020. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient Transfer Learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. 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 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Lawrence Hubert and Phipps Arabie. 1985. Comparing Partitions. *Journal of Classification*, 2:193–218.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dominik Schlechtweg. 2023. *Human and computational measurement of lexical semantic change*. Ph.D. thesis, Universität Stuttgart.

University of Tartu. 2018. UT Rocket.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with Bert.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

## A Training Examples

Table 5 presents two training examples for the Russian word “Перо” (*Feather*). In the first row, we have a gloss and a matching usage example for the figurative meaning of the word (*a symbol of the writer’s art*), which is denoted by the label **1**. Each usage example of the word in its other senses is paired with this gloss and used as a negative example labeled **0**. For instance, in the second row, the same gloss is paired with a mismatching usage example in the literal sense of the word. We omit the rest of the negative examples and the other senses for brevity.

| **Gloss** | **Usage example** | **Label** |
|---|---|---|
| “Символ искусства писателя, писательского труда, его ремесла.” | “У него бойкое, острое перо.” | 1 |
| “Символ искусства писателя, писательского труда, его ремесла.” | “Перья зверя.” | 0 |

Table 5: A subset of training examples for a single word.