# Uppsala University at SemEval-2022 Task 1: Can Foreign Entries Enhance an English Reverse Dictionary?

**Rafal Černiavski** and **Sara Stymne**
Department of Linguistics and Philology
Uppsala University
rafal.cerniavski.2286@student.uu.se, sara.stymne@lingfil.uu.se

## Abstract

We present the Uppsala University system for SemEval-2022 Task 1: Comparing Dictionaries and Word Embeddings (CODWOE). We explore the performance of multilingual reverse dictionaries as well as the possibility of utilizing annotated data in other languages to improve the quality of a reverse dictionary in the target language. We mainly focus on character-based embeddings. In our main experiment, we train multilingual models by combining the training data from multiple languages. In an additional experiment, using resources beyond the shared task, we use the training data in Russian and French to improve the English reverse dictionary using unsupervised embeddings alignment and machine translation. The results show that multilingual models occasionally but not consistently can outperform the monolingual baselines. In addition, we demonstrate an improvement of an English reverse dictionary using translated entries from the Russian training data set.

## 1 Introduction

In a reverse dictionary, one can look up a gloss, an explanation of a word's meaning, to find the most relevant word or word form. The applications of reverse dictionaries are numerous, as they can help language learners in expanding their vocabulary, authors and writers in looking for the most suitable word, and avid cruciverbalists in taking on some of the most challenging crosswords.

Reverse dictionary modelling has seen approaches ranging from traditional information retrieval using relevance scores (Zock and Bilac, 2004) to ones involving node-graph architectures (Zhang et al., 2020). As a general rule, the quality of a reverse dictionary appears to largely depend on the availability of annotated data. However, annotated data are scarcely available and expensive to produce for low-resource languages. We therefore explore the viability of multilingual approaches to improve the quality of a reverse dictionary.

This work is performed in the context of the reversed dictionary subtask of the SemEval 2022 task 1, COmparing Dictionaries and WOrd Embeddings (Mickus et al., 2022). Unlike standard reverse dictionaries, the target is to predict a word embedding vector for each gloss, rather than a word form. Three types of word embeddings are available: character-based embeddings (*char*), Skip-grams (*sgns*), and contextual embeddings (*electra*). No additional resources are allowed in the shared task. In this paper, we do present additional experiments, though, where we also used an external machine translation engine. While five languages were made available in the shared task, we mainly focus on English, but also give some results for Russian and French.

The main research question of this study is thus whether the performance of a monolingual reverse dictionary can be improved using data in other language(s) in a low supervision setup. We first explore what are the most suitable type of embeddings for a Transformer-based reverse dictionary. Having found the best-performing embeddings, we use them to train a joint model for multilingual reverse dictionary, which can map glosses to words in multiple languages. Finally, we use the training data in French and Russian to improve the quality of an English reverse dictionary by means of unsupervised embeddings alignment and machine translation.

We did not submit our results in the evaluation period since in one of the experiments we used a pre-trained neural machine translation model, which is prohibited in the shared task. Nevertheless, we report the performance of our jointly trained multilingual models on the test sets, as no additional data or pre-trained models were involved in training. For character-level embeddings, our best multilingual models, when tested on English,

*Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 88 - 93
July 14-15, 2022 ©2022 Association for Computational Linguistics

would rank 25th in terms of mean squared error (MSE), 20th in terms of cosine similarity (COS), and 9th in terms of cosine-based ranking (CRK); on French: 22nd (MSE), 10th (COS), 3rd (CRK); on Russian: 7th (MSE), 7th (COS), 13th (CRK).

## 2 Related Work

Recent research has explored bilingual and cross-lingual reverse dictionaries, the task of which is to map a gloss in a source language to a word in target language. An implementation by Qi et al. (2020) involved a machine translation API and bilingual dictionaries to re-direct a query in the source language through the target language pipeline. Yan et al. (2020) implemented the first cross-lingual reverse dictionary based on mBERT (Devlin et al., 2019), a Transformer-based language model trained on Wikipedia articles in 104 languages. Their study revealed that unaligned cross-lingual reverse dictionary achieves best performance when mBERT is tuned on unaligned multilingual data; its quality is substantially worse than that of a monolingual model. Yan et al. (2020) thus concluded that it remains unclear how multilingual data is to be utilized to improve the quality of unaligned reverse dictionary, which is to be explored in this project.

Joint multilingual models, which are trained on multiple languages at once, offer a solution for low-resource languages that often have little to none annotated data. This has for example been explored for dependency parsing, with positive results (Kondratyuk and Straka, 2019; Smith et al., 2018).

Cross-lingual embeddings are of central importance in word meaning similarity across languages (Jimenez et al., 2017), and are thus a crucial component of cross-lingual reverse dictionaries. As noted by Ruder et al. (2019), the applicability of cross-lingual embeddings relies on their quality, which, in turn, depends on the availability of bilingual corpora and dictionaries. Nevertheless, an unsupervised cross-lingual embeddings alignment method proposed by Lample et al. (2018) enables high quality cross-lingual embeddings with no or little supervision, further allowing for unsupervised machine translation. Unsupervised cross-lingual embeddings alignment thus offers a solution for both mapping the word embeddings and its glosses from one language to another.

## 3 System Description

We focus on the strategies of utilizing the data in foreign languages to improve reverse dictionary rather than the choosing of most suitable model. Therefore, we use the SemEval 2022 task 1 baseline system, a Transformer-based architecture with all parameters unchanged for all of our models.

### 3.1 Methodology

The methodology adopted can be divided into a preparatory step and two main experiments. The initial step sought to learn the most suitable type of embeddings for a Transformer-based English reverse dictionary. A baseline model was trained and tested three times on each type of embedding to learn whether there were notable deviations between the runs and the official baseline scores of the shared task. This was done to select the best performing type of embedding to be used in further experiments, thus avoiding spending the computational resources on numerous models with different embeddings.

The two main experiments build on the research of He et al. (2017), as they investigate joint training of multilingual models as well as cross-lingual embedding alignment. In the first experiment, the French and Russian training sets are concatenated to the English training set, one or both at a time. The joint models are then trained with a joint development set containing entries in all languages used in training. We choose the source languages, namely French and Russian, so as to investigate whether the similarities between the source and target language, such as shared words, similar script, and typological proximity can affect the performance of a multilingual reverse dictionary.

In the second experiment, the embeddings of source entries (in French and Russian) are firstly aligned to the target embedding space (English) with no supervision using the MUSE library (Lample et al., 2018). To ensure a fully unsupervised setup, the refinement and evaluation steps involving bilingual corpora are disabled. The alignment is conducted in five epochs using all standard parameters. In the process, the target embeddings are anchored. Their values are not updated in order to preserve the quality of the pre-trained embeddings. Secondly, the glosses of the first 4,500[1] entries from the now-aligned source training set are

[1] A relatively small number of glosses were translated due to the limited access to the tool used for machine translation.

| Embeddings | MSE | $\sigma$ | COS | $\sigma$ | CRK | $\sigma$ |
|---|---|---|---|---|---|---|
| *sgns* | 1.193 | 0.009 | 0.259 | 0.007 | **0.405** | 0.012 |
| *char* | **0.156** | 0.014 | 0.810 | 0.003 | 0.469 | 0.003 |
| *electra* | 1.846 | 0.172 | **0.840** | 0.001 | 0.483 | 0.002 |

Table 1: The baseline performance of a Transformer-based English reverse dictionary trained on different types of pre-trained embeddings, averaged over 3 runs. The standard deviation ($\sigma$) shows the fluctuation of the scores over the three runs.

translated and attached to the target (English) training set. Since the word forms are masked in the training data, we were unable to train an unsupervised machine translation model. The glosses are thus translated using a pre-trained neural machine translation model, namely Watson API[2]. Lastly, the translated glosses are tokenized using the spaCy tokenizer to mirror the tokenization in the original data sets provided by the organizers of the shared task.

### 3.2 Evaluation

All models are primarily evaluated on the trial data set. This is due to the fact that Experiment 2 used a pre-trained machine translation model, which goes against the rules of the contest. We, however, additionally evaluate our jointly trained multilingual models on the test set, as the model does not use any additional resources.

The models were evaluated based on the three official metrics of the shared task: mean squared error (MSE), cosine similarity (COS), and cosine-based ranking (CRK) (Mickus et al., 2022).

## 4 Results and Discussion

### 4.1 Choice of Embeddings

The performance of the baseline models trained on the English training data set with different embeddings can be seen in Table 1. The scores are highly similar to the baselines published by Mickus et al. (2022) and are primarily included to estimate the stability of the performance of a Transformer architecture on each type of embeddings.

Individually, each type of embedding achieves the highest score on one of the parameters, with *char* achieving lowest MSE, *electra* securing highest cosine similarity, and *sgns* having best cosine-based ranking. Overall, *char* embeddings demonstrate the most stable and good performance across all three parameters. The *char* embeddings also had a relatively low standard deviation between runs for all metrics, as opposed to *electra* on MSE.

The results seem to have several implications. Firstly, the three evaluation parameters favour divergent information encoded by the three types of embeddings. Most notably, character-level information stored in *char* embeddings substantially minimizes MSE of the predicted embeddings of a word. This might be because character-level embeddings are effective in addressing out-of-vocabulary words (Polatbilek, 2020). In other words, they seem to enable the Transformer model to learn the mapping between glosses and characters that add up to words denoting the glosses. However, such mapping suffers from a major limitation, as character-level embeddings do not differentiate between the senses of a word. Most effective in handling this task are the contextualized embeddings (*electra*), for they encode a word depending on the surrounding context. Depending on the context, the sense might differ, thus leading to completely different values in the embeddings space. It can thus be argued that both character-level and contextualized features are important for a reverse dictionary model; an ideal solution could perhaps utilize using both types of embeddings for fine-grained retrieval of words.

Seeing as *char* embeddings had a good and stable performance overall, we further explore them in the following experiments.

### 4.2 Multilingual Model

The performance of multilingual models jointly trained for two or three languages at a time is reported in Tables 2 and 3.

The multilingual models perform similarly to the monolingual baselines. As can be seen from comparing the models' performance across trial and test sets, some differences are likely due to chance and fall within the range of a standard deviation reported in 1. Nevertheless, it is rather surprising that the English reverse dictionary seems to bene-

[2] https://developer.ibm.com/components/watson-apis/

|  | English (E) |  |  |  | French (F) |  |  | Russian (R) |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | E (Base) | E+F | E+R | E+F+R | F (Base) | F+E | F+E+R | R (Base) | R+E | R+E+F |
| MSE | 0.17893 | 0.18897 | **0.14708** | 0.19417 | **0.39491** | 0.43406 | 0.51295 | **0.13858** | 0.15327 | 0.25199 |
| COS | 0.79591 | 0.78978 | **0.80659** | 0.79472 | **0.78361** | 0.77169 | 0.77499 | **0.84409** | 0.83503 | 0.83073 |
| CRK | **0.45771** | 0.46748 | 0.49775 | 0.48978 | 0.47125 | 0.45235 | **0.45225** | 0.42565 | 0.41385 | **0.40665** |

Table 2: The performance of a multilingual reverse dictionary jointly trained on *char* embeddings in the source and target language evaluated on the **trial** set. The performance of multilingual model (joint) is reported alongside its monolingual baseline.

|  | English (E) |  |  |  | French (F) |  |  | Russian (R) |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | E (Base) | E+F | E+R | E+F+R | F (Base) | F+E | F+E+R | R (Base) | R+E | R+E+F |
| MSE | 0.17893 | 0.22773 | **0.17821** | 0.21932 | **0.45808** | 0.50001 | 0.53045 | 0.16775 | **0.16075** | 0.24864 |
| COS | **0.79591** | 0.76452 | 0.78445 | 0.77336 | **0.77978** | 0.75831 | 0.76917 | **0.84044** | 0.83220 | 0.83349 |
| CRK | 0.45771 | **0.45639** | 0.46198 | 0.46480 | 0.45006 | **0.42284** | 0.43047 | 0.42073 | **0.40115** | 0.40776 |

Table 3: The performance of a multilingual reverse dictionary jointly trained on *char* embeddings in the source and target language evaluated on the **test** set. The performance of multilingual model (joint) is reported alongside its monolingual baseline.

fit from the Russian data more than it does from the French data. In addition, when trained on both English and Russian, the model performs better on Russian.

In the case of multilingual models, it might be productive to focus on the lack of losses rather than the lack of gains. The results indicate that the performance of Transformer-based English reverse dictionary remains unaffected by both a relatively close language (French), and a distant language (Russian). This might be due to the fact that the high-quality pre-trained embeddings exist in different vector spaces. Despite the fact that the data are concatenated, the Transformer architecture learns to differentiate between the two and only retrieve words from the relevant vector space.

The shared space of models like mBERT is arguably the main reason why the joint tuning of models on data in multiple languages at once leads to best performance of a cross-lingual reverse dictionary for Yan et al. (2020). Overall, it is debatable whether there is reason to train a multilingual reverse dictionary on several unaligned languages. Such a model takes longer to train and tune, occupies more space, and does not offer much apart from the convenience of not having to switch between multiple models.

### 4.3 Embeddings Alignment and Machine Translation

The last experiment involved unsupervised embeddings alignment and machine translation of the glosses from source language (French and/or Russian) to target language (English). During alignment, the target embeddings were anchored to retain the values of the pre-trained embeddings. However, due to system constraints, the target embedding values changed from ten decimal points to five. To address this and to see whether this could affect the results in a negative way, an additional model was trained with the restored original values (with ten decimal points) of the embeddings in English, while the source (French and Russian) embeddings were kept at five decimal points. The results are presented in Table 4 alongside the baseline results.

Alignment without translation of glosses in most cases affected the model in a negative way, as it only introduced noisy foreign data. However, the machine translated glosses attached to the aligned values from source language seemed to have a positive effect on the English reverse dictionary when the source language was Russian. In the case of French, the approach failed completely. The retention of the original embedding values as opposed to the last five digits being lost led to mixed results. Though in most cases the difference is small and might have occurred by chance, the results could also indicate that it is crucial for the source and target embeddings to be similar in terms of the quality

| | | English + French | | | English + Russian | | |
|---|---|---|---|---|---|---|---|
| Metric | Baseline | Al | Al+T | Al+TR | Al | Al+T | Al+TR |
| MSE | 0.156 | 0.184 | 0.171 | 0.184 | 0.162 | **0.135** | 0.171 |
| COS | 0.810 | 0.799 | 0.810 | 0.801 | 0.807 | **0.811** | 0.810 |
| CRK | 0.469 | 0.474 | 0.500 | 0.483 | 0.477 | 0.501 | **0.463** |

Table 4: The performance of a Transformer-based English reverse dictionary trained on aligned and joined data (Al), aligned with target embeddings cut off past five digits and machine translated glosses (Al+T), as well as aligned with recovered target embeddings and machine translated glosses (Al+TR).

in a cross-lingual space.

A rather surprising finding of the experiment was the improvement of an English reverse dictionary using the data in Russian. Contrary to the findings of Yan et al. (2020), a more substantial improvement for English was observed with a distant source language, which uses a completely different script. The Russian language has been previously proposed as a generally good source language across several tasks and target languages, though (Turc et al., 2021). As for this experiment, perhaps the alignment produced with no supervision was of higher quality with Russian, allowing to correctly project the foreign source entries in the target space. It is also possible, though unlikely, that the translations of glosses from Russian to English were of higher quality than those of French to English.

## 5 Conclusions

This project has investigated whether an English reverse dictionary can be improved using data in foreign languages. This research question was addressed by firstly determining the most suitable type of embeddings for a Transformer-based reverse dictionary. Secondly, multilingual joint models were trained to see the affects on the performance of English as target language and two source languages, namely French and Russian. Lastly, the embeddings from source language were aligned to the target embedding space, followed by machine translation of the respective glosses.

Three key findings emerged. Firstly, character-level features lead to best performance of an English Transformer-based reverse dictionary. Secondly, multilingual reverse dictionaries perform comparably with monolingual ones, as no substantial improvement or decline was observed. Thirdly, an English reverse dictionary can be improved using the available data in foreign languages, such as French and Russian, though the improvement is rather small. In the reported experimental setup, Russian was found to be a more suitable source language in enhancing an English reverse dictionary.

There are numerous possible extensions of the present study. One could, for instance, recreate the study in a fully supervised or fully unsupervised setup so as to see to what extent the lack of supervision affected the results. It would also be interesting to investigate whether combinations of embeddings, e.g. contextual and character-level, would lead to better performance of reverse dictionary models. Overall, the improvements recorded in this study were, arguably, hardly significant. It may therefore be productive to search for more successful ways of using data in foreign languages in creating or improving reverse dictionaries.

## Acknowledgements

Computations were enabled by resources in project UPPMAX 2020/2-2 at the Uppsala Multidisciplinary Center for Advanced Computational Science.

## References

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.

Junqing He, Long Wu, Xuemin Zhao, and Yonghong Yan. 2017. HCCL at SemEval-2017 task 2: Combining multilingual word embeddings and transliteration model for semantic similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 220–225, Vancouver, Canada. Association for Computational Linguistics.

Sergio Jimenez, George Dueñas, Lorena Gaitan, and

Jorge Segura. 2017. RUFINO at SemEval-2017 task 2: Cross-lingual lexical similarity by extending PMI and word embeddings systems with a Swadesh's-like list. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 239–244, Vancouver, Canada. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, Vancouver, Canada.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2022. SemEval-2022 Task 1: CODWOE – COmparing Dictionaries and WOrd Embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Ozan Polatbilek. 2020. *Enriching Contextual Word Embeddings with Character Information*. Ph.D. thesis, Izmir Institute of Technology (Turkey).

Fanchao Qi, Lei Zhang, Yanhui Yang, Zhiyuan Liu, and Maosong Sun. 2020. WantWords: An open-source online reverse dictionary system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–181, Online. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of English in zero-shot cross-lingual transfer. arXiv preprint, arXiv:2106.16171.

Hang Yan, Xiaonan Li, Xipeng Qiu, and Bocao Deng. 2020. BERT for monolingual and cross-lingual reverse dictionary. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4329–4338, Online. Association for Computational Linguistics.

Lei Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Multi-channel reverse dictionary model. arXiv preprint, arXiv:1912.08441.

Michael Zock and Slaven Bilac. 2004. Word lookup on the basis of associations : from an idea to a roadmap. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 29–35, Geneva, Switzerland. COLING.