FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

universität wien

REPORT

"Enhancing Cognate Detection Through Multilingual Model Fine-Tuning"

VU Introduction to Computational Linguistics
University of Vienna
(WiSe 2024)

Written by

Fabian Schambeck BA

PIN, Mat. No.: 2410843026, 12033754
Degree Programme Codes: 0843, UA 066 587
Joint-Master's Degree Multilingual Technologies

Submitted on February 7, 2025
Assoz. Prof. Mag. Dr. Dagmar Gromann BSc

1 Introduction

Cognate detection is a task in computational and historical linguistics that involves identifying words in different languages that share a common etymological origin. Two words are considered cognates if they derive from the same ancestral word in a proto-language, often preserving similarities in spelling, pronunciation, and meaning (List, 2014). Accurate cognate identification plays a crucial role in various multilingual NLP applications. By recognizing cognate relationships, computational models can improve word alignment, enhance multilingual embeddings, and support tasks requiring cross-linguistic transfer learning. While traditional approaches to cognate detection rely primarily on phonetic and orthographic similarity metrics, recent advances in deep learning have enabled more robust and scalable solutions using transformer-based language models.

The project presented in this work focuses on fine-tuning a pretrained language model to improve its ability to predict the French cognate of a given English word. The initial approach involved fine-tuning `xlm-roberta-base` on a dataset of aligned English-French sentences from `Helsinki-NLP/europarl`. However, this dataset consists of translations, meaning that word alignments are based on semantic equivalence rather than etymological relatedness, making it inaccurate for true cognate detection. To address this issue, I switched to a dataset by Frossard et al. (2020), which was specifically designed for English-French cognate detection and contains true cognate word pairs.

2 Methodology

The main objective of this project was to fine-tune `xlm-roberta-base` to improve its ability to predict French cognates using a masked language modeling (MLM) approach. The task was designed such that the model is given an English word and must predict the correct masked French cognate. The predicted French word was then compared to the actual French word from the dataset to assess accuracy. A bilingual masked input sentence pair was used to provide context for the model, ensuring that it takes advantage of cross-lingual knowledge when predicting the masked token. The input followed the format:

(1) *"In English, the word is* ***word_en****. En Français, le mot est <****mask****>."*

The project's pipeline consisted of three main steps. First, the pretrained model's performance was evaluated without fine-tuning to establish a baseline. Next, `xlm-roberta-base` was fine-tuned on a curated dataset specifically designed for this task, with the goal of improving the model's ability to predict cognates. Finally, the fine-tuned model was re-evaluated to measure improvements in cognate prediction accuracy. Performance was assessed by checking whether the correct French cognate appeared in the top-5 retrieved predictions. The masked prediction accuracy assessment was conducted on the test split of the dataset, where the model was used purely for inference without fine-tuning.

The dataset, derived from Frossard et al. (2020), consists of 492 English-French word pairs, which were split into 70% training (n=344), 20% validation (n=98), and 10% test data (n=50). The dataset includes the following key features: the English and French word pairs (`word_en`, `word_fr`), tokenized representations (`input_ids`), attention masks

(attention_mask), and masked language modeling labels (labels). Table 1 provides an overview of these features.

Feature	Description
word_en	English word in the pair
word_fr	Corresponding French word
input_ids	Tokenized sentence representation
attention_mask	Binary mask for valid tokens
labels	Target tokens for masked positions

Table 1: Dataset features and descriptions.

The input_ids represent the tokenized form of the input sentence, including special tokens and the <mask> token at the masked position. The attention_mask is a binary mask that distinguishes valid tokens (1) from padding tokens (0). The labels feature contains the expected token(s) for the masked position, while all other positions are set to -100 to ensure they are ignored in the loss calculation.

Hyperparameter tuning was performed manually over approximately 20 trials to determine an optimal configuration. The final training setup used a batch size of 32, a learning rate of 3e-5, and a linear learning rate scheduler with 10% warmup steps. Training was conducted over 6 epochs with gradient accumulation (steps=2) to stabilize updates. To mitigate overfitting, weight decay (0.01) and max gradient norm clipping (1.0) were applied. Training was performed on a CPU in Google Colab.

Fine-tuning and evaluation were conducted using the Hugging Face Trainer API, automating training and validation while ensuring efficient performance tracking. The model was configured to save checkpoints at each epoch, and the best-performing model was selected based on validation loss. The final evaluation on the test set quantifies the improvement achieved through fine-tuning.

3 Results

The fine-tuned xlm-roberta-base model was evaluated based on training loss, validation loss, accuracy, and masked token prediction performance. The training process spanned six epochs, with loss values recorded for both training and validation sets to track the model's optimization. Training loss decreased from 8.1771 in the first epoch to 2.4609 in the final epoch, while validation loss followed a similar trend, decreasing from 5.8079 to 2.8598. Accuracy improved steadily from 12.28% at the beginning of training to 55.56% by the final epoch, indicating that the model successfully learned to identify cognates more effectively. The training and validation losses suggest that the model converged without severe overfitting, as validation loss continued to decrease alongside training loss. In Table 2, the training statistics are presented.

Epoch	Training Loss	Validation Loss	Accuracy
1	8.1771	5.8079	12.28%
2	4.4220	3.8068	42.69%
3	3.4863	3.3821	46.78%
4	3.6457	2.9702	52.05%
5	2.9118	2.9007	54.39%
6	2.4609	2.8598	55.56%

Table 2: Training and validation loss along with accuracy per epoch.

While overall accuracy improved, the masked prediction performance on the test set revealed a key limitation of the fine-tuning process. The top-5 accuracy, which measures whether the correct French cognate appears among the model's five most likely predictions, remained unchanged at 46% before and after fine-tuning. This indicates that while the fine-tuned model learned to predict cognates more confidently, it did not improve in ranking the correct cognate among its top predictions. The lack of improvement in top-5 accuracy suggests that the model's initial ability to capture cognate relationships was already strong and that fine-tuning primarily influenced token ranking probabilities rather than retrieval performance.

To assess top-5 accuracy, I implemented a custom evaluation function that retrieves the model's five most likely predictions for each test instance and checks whether the correct cognate appears among them. Table 3 summarizes the results.

Model	Top-5 Accuracy
Pretrained `xlm-roberta-base`	46%
Fine-tuned `xlm-roberta-base`	46%

Table 3: Comparison of pretrained and fine-tuned model accuracy on the test set.

The qualitative analysis compares the pretrained and fine-tuned `xlm-roberta-base` model outputs in the cognate prediction task. Examining individual predictions reveals patterns in model behavior and highlights the effects of fine-tuning. Firstly, the results show that both models correctly predicted the cognate in multiple cases. These cases demonstrate that the model has an inherent ability to identify cognates, likely due to multilingual pretraining on large corpora that include frequent cross-lingual word correspondences.

(2) **English:** abdominal, **Expected French:** abdominal
Pretrained: [abdominal, muscle, bouche, :, stomach]
Fine-tuned: [abdominal, e, né, eur, bé]

(3) **English:** festival, **Expected French:** festival
Pretrained: [festival, fête, Festival, concert, spectacle]
Fine-tuned: [festival, festival, e, é, Festival]

(4) **English:** negative, **Expected French:** négatif
Pretrained: [positive, positif, négatif, negative, negativ]
Fine-tuned: [négatif, negative, positif, negativ, positive]

Secondly, the fine-tuned model displays a shift in prediction strategy, often generating subword components rather than complete words. This behavior suggests that the fine-tuned model has learned word formation patterns rather than memorizing complete words, possibly due to the BytePair tokenization method.

(5) **English:** affirmative, **Expected French:** affirmatif
Pretrained: [:, :, de, expression, souvent]
Fine-tuned: [firma, firm, i, tive, tif]

(6) **English:** rehabilitation, **Expected French:** réhabilitation
Pretrained: [recovery, :, :, de, reparation]
Fine-tuned: [tion, té, ité, adaptation, é]

Another noticeable downside of the fine-tuned model is semantic drift: its predictions become less meaningful compared to the pretrained version. In some cases, fine-tuning led to outputs favoring high-frequency syllabic patterns instead of coherent words.

(7) **English:** syndicate, **Expected French:** syndicat
Pretrained: [syndicat, collective, :, production, de]
Fine-tuned: [e, ne, é, né, ine]

(8) **English:** civilization, **Expected French:** civilisation
Pretrained: [nation, :, :, population, société]
Fine-tuned: [ité, tion, cité, té, isation]

All in all, the fine-tuned model shows a shift in prediction behavior, losing the ability to generate fully formed words and instead favoring syllabic units. This suggests an over-reliance on tokenized subwords rather than holistic word representations. One noticeable change is the model's improved recognition of morphological components, as it successfully identifies subword units. However, it often fails to assemble them into complete words, resulting in fragmented or partial outputs. While the pretrained model occasionally produced semantically related but incorrect words, the fine-tuned version often generates word fragments that lack coherence.

Despite these structural differences, the fine-tuned model does not achieve a measurable improvement or decrease in overall performance, as the top-5 accuracy remains unchanged at 46%. This suggests that fine-tuning primarily affects token selection and ranking rather than enhancing the model's ability to detect cognates.

The visualizations presented in Figures 1 and 2 compare the embedding distributions of the test dataset at layers 4 and 12, respectively. Figure 1 depicts the embeddings at layer 4, where a pronounced clustering pattern can be observed. Distinct groupings are visible across the plot, suggesting that at this stage, the model's embeddings maintain clear separations between certain data points. This indicates that layer 4 captures higher-level syntactic or semantic similarities within the data. The relatively tight and discrete nature of the clusters at this stage suggests that the model is in an early phase of refining its representations, focusing on localized patterns or shallow features.

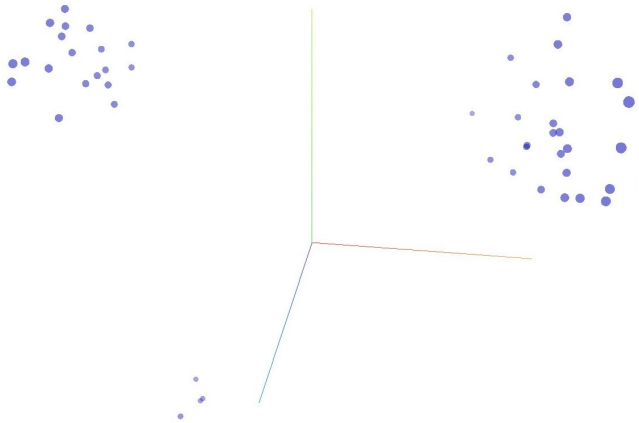

Figure 1: Embedding distribution of the test dataset at layer 4.

In contrast, Figure 2 illustrates the embeddings at layer 12, where a notable dispersion of data points is evident. While some areas of density indicate residual grouping tendencies, the overall structure appears less defined compared to layer 4. This increased spread and reduced clustering reflect the model's deeper processing of the input, as representations are transformed into more generalized and abstract encodings.

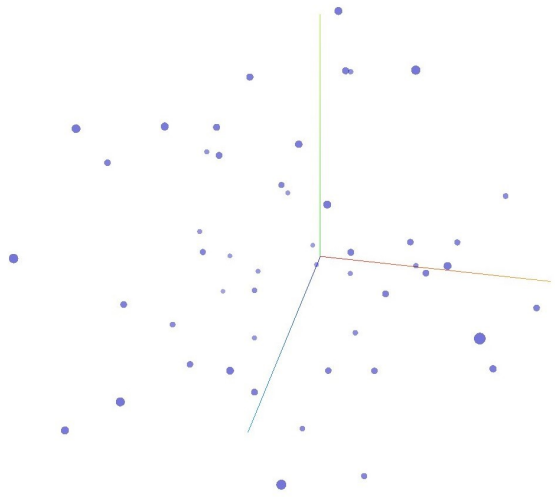

Figure 2: Embedding distribution of the test dataset at layer 12.

The transition from structured clusters in layer 4 to a more diffused representation in layer 12 suggests that the model evolves from explicit, interpretable feature extraction to a phase of more abstract and nuanced encodings. While this progression highlights deeper learning processes, the reduced cohesion in layer 12 may pose challenges for tasks requiring well-defined separability, such as cognate detection.

4 Discussion

The observed results highlight the benefits and limitations of fine-tuning `xlm-roberta-base` for cognate detection. While the model demonstrated improved confidence in selecting correct cognates, the overall unchanged top-5 accuracy at 46% underscores the model's inability to consistently predict full cognate words.

One notable challenge arises when the model encounters words it does not know. Due to the use of byte-pair encoding, the model can still process unseen words by breaking them down into subword units. However, this ability often leads to fragmented or incomplete predictions, as evidenced by the fine-tuned model's tendency to favor subword components over complete words. This behavior can undermine performance, particularly when full-word predictions are required for accurate cognate detection. Incorporating character-level supervision could help address this issue by ensuring that the model generates whole-word outputs rather than partial morphemes.

A promising direction for improvement might lie in the use of contrastive learning objectives. By providing the model with explicit negative examples—pairs of words that are not cognates, it could learn to better differentiate between true cognates and false friends. This approach could be further enhanced by introducing a binary feature to the dataset, labeling each pair as either a true cognate (1) or not a cognate (0). Such a feature could encourage the model to develop a more robust understanding of etymological relationships, potentially improving its generalization capabilities.

Another possible enhancement involves training the model on multiple example sentences for each word pair. Providing a variety of contextual sentences for the masked language modeling task could give the model more diverse linguistic contexts to consider, helping it better capture the subtleties of word relationships. This additional contextual information may also reduce the model's reliance on local token patterns and encourage it to focus on broader semantic cues.

Expanding the dataset to include a wider range of cognates across more language pairs could also yield significant benefits. A larger and more diverse dataset would not only improve the model's exposure to various etymological patterns but also make it less prone to overfitting on a limited set of examples. Furthermore, experimenting with alternative model architectures, such as `xlm-roberta-large` or multilingual BERT, could provide insights into whether increased model capacity leads to better generalization and improved performance on cognate detection tasks.

Additional improvements might also focus on enhancing the training objective. For instance, combining masked language modeling with auxiliary tasks, such as predicting phonetic similarity scores or identifying orthographic patterns, could encourage the model to learn features more directly relevant to cognate detection. Fine-tuning the model with custom pretraining objectives tailored to cross-lingual word alignment could also help align representations more effectively across languages.

Overall, while the fine-tuning approach demonstrated potential, it is clear that cognate detection remains a challenging task requiring a combination of data-centric and model-centric improvements. Addressing these challenges through the strategies discussed could significantly enhance the model's ability to identify cognates, ultimately contributing to broader advancements in multilingual NLP.

5 Conclusion and Reflections

This project explored fine-tuning `xlm-roberta-base` for the task of cognate detection, focusing on improving the model's ability to predict cognates in a multilingual context. While the model demonstrated an increased confidence in selecting correct cognates, the unchanged top-5 accuracy of 46% highlighted the complexity of the task and the need for further improvements in both model design and data preparation. Several potential enhancements were identified, such as incorporating contrastive learning, expanding the dataset, and experimenting with alternative model architectures to better address the challenges observed.

The project took longer than anticipated, particularly in defining the methodology. One of the key decisions was whether to approach the task as a simple binary classification problem or to use a masked language modeling approach. After careful consideration, I chose the masked approach, as it seemed more aligned with the model's architecture and offered a more interesting challenge. While this decision made the project more complex, it provided valuable insights into the intricacies of working with transformer-based models and multilingual datasets.

Reflecting on the experience, the task was both rewarding and challenging. Developing a deeper understanding of fine-tuning strategies and evaluating the model's embeddings was particularly interesting, but navigating the trade-offs in methodology required significant effort. If given the chance to revisit this project, I might consider experimenting with a larger model, such as `xlm-roberta-large`, or exploring a classification-based approach to compare its simplicity and effectiveness against the masked task. Nonetheless, this project provided a rich learning experience, underscoring the importance of adapting methodologies to suit the specific goals and constraints of a task.

References

Esteban Frossard, Mickaël Coustaty, Antoine Doucet, Adam Jatowt, and Simon Hengchen. 2020. Dataset for Temporal Analysis of English-French Cognates. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'20)*, Marseille.

Johann-Mattis List. 2014. *Sequence Comparison in Historical Linguistics*. Dissertations in Language and Cognition 1. Düsseldorf University Press, Düsseldorf.

Statutory Declaration

Hereby, I affirm that I have independently written the present work and have not used any aids other than those specified.

Passages of the work that have been taken verbatim or in essence from other sources are marked with citations. This also applies to drawings, sketches, and visual representations as well as sources from the internet.

Furthermore, I declare that I have not previously submitted this work, in whole or in part, as an examination performance.

Vienna, February 7, 2025

Fabian Schambeck BA