Cross-Lingual False Friend Classification via LLM-based Vector Embedding Analysis

Mitko Nikov
mitko.nikov@student.um.si
Faculty of Electrical
Engineering and Computer Science,
University of Maribor
Koroška cesta 46
SI-2000 Maribor, Slovenia

Žan Tomaž Šprajc
zan.sprajc@student.um.si
Faculty of Electrical
Engineering and Computer Science,
University of Maribor
Koroška cesta 46
SI-2000 Maribor, Slovenia

Žan Bedrač
zan.bedrac@student.um.si
Faculty of Electrical
Engineering and Computer Science,
University of Maribor
Koroška cesta 46
SI-2000 Maribor, Slovenia

ABSTRACT

In this paper, we propose a novel approach to exploring cross-linguistic connections, with a focus on false friends, using Large Language Model embeddings and graph databases. We achieve a classification performance on the Spanish-Portuguese false friend dataset of F1 = 83.81% using BERT and a multi-layer perceptron neural network. Furthermore, using advanced translation models to match words between vocabularies, we also construct a ground truth false friends dataset between Slovenian and Macedonian - two languages with significant historical and cultural ties. Subsequently, we construct a graph-based representation using a Neo4j database, wherein nodes correspond to words, and various types of edges capture semantic relationships between them.

KEYWORDS

false friends, large language models, BERT, linguistics

1 INTRODUCTION

When observing individual languages, we come across homonyms, which are words that have the same spelling or pronunciation but varied meanings, such as the word "bat", which pertains to either the animal or the sports requisite. As we move from the confines of one language and observe two, we encounter chance false friends [10]. These have the same spelling but varied etymologies and meanings in different languages, such as the English word "in", which in Slovenian means "and". So, we decided to pivot our observation further and focus solely on words that have the same etymological origin and spelling whilst having different meanings in different languages, so-called semantic false friends [10, 12, 16].

A similar endeavour was undertaken by Ljubešić & Fišer [13], which attempted to identify true equivalents, partial false friends, and false friends in Slovenian and Croatian based on their spelling and semantic meaning. Our analysis will also touch on true equivalents (word pairs with the same meaning and usage [13]), partial false friends (pairs that alternate between polysemy and false friends [13]), and pure false friends.

An initial step to finding false friends could be lemmatization-based tagging [4], which is able to differentiate between parts of speech, reducing words to their root form. Which in practice means that a verb like "working" is reduced to its root of "work". Stemming is another alternative, which has already been applied to Czech together with a language-independent approach (n-gram) [9]. However, even though lemmatization proved effective for two other South Slavic languages, Croatian and Serbian [4], in our case, we expect the declension differences between Slovenian and Macedonian to be too significant for such a preprocessing step to be used.

A recently introduced method for automatic false friends detection in related languages [6] uses a linear transformation between the two vector spaces in both languages to isolate false friends. The linear transformation acts as a translation between the two languages. They [6] expect that one vector in one language should be close to its cognate partner [5] in the other language after the linear transformation, however, for false friends, this should not be the case. They use the Spanish and Portuguese Wikipedia as a corpus for the unsupervised learning of the Word2Vec models [15].

Since the linear transformation is a bijection, each vector in one of the languages is uniquely mapped to a vector in the other language. It is impossible for such a model to account for the different meanings one word can have. To solve this issue, we propose an improvement to this method by extending the vector space to use LLM embeddings [18] of meanings instead of single words.

Regarding the false friend classification between Macedonian and Slovenian, we needed to take a different approach to ground truth dataset creation. Our approach is based on finding words with the same spelling in Slovenian and Macedonian, translating them to English using a pre-trained bidirectional translator API and matching false friends accordingly. This approach also yields an unexpected amount of true friends, which are also useful to us. A prime example of a false friend would be the word "obraz", which in Slovenian means "face" and in Macedonian means "cheek". On the other hand, a true friend would be the word "jagoda", which means "strawberry" in both Slovenian and Macedonian. These and a few other examples are given in Table 1.

In the following sections, we will describe the methodology that we used to classify false friends, as well as the methodology used to create a ground truth dataset for Slovenian and Macedonian false friends. Our overview of the classification process will be based on BERT as a word to vector model, with special attention given to the embedding extraction process. Moreover, we will dive into our methodology for ground truth creation, with a special emphasis on the issues that result from the translation of words that have multiple meanings. We will also describe the graph database representation of the false friends dataset. Finally, we will present our results, comparing our methodology to an existing one. We will evaluate our results in terms of precision, recall, F1 scores, and provide a summary of our false friends ground truth dataset.

Table 1: Examples of true and false friends in the Slovenian and Macedonian language.

Slovenian word	Macedonian word	Slovenian meaning	Macedonian meaning	Type of word match
obraz	образ (obraz)	face	cheek	false friend
lice	лице (lice)	cheek	face	false friend
deka	дека (deka)	blanket	that	false friend
čas	час (čas)	time	time/hour	partial false friend
jagoda	јагода (jagoda)	strawberry	strawberry	true friend
kraj	крај (kraj)	edge/end/region	edge/end/region	true friend

2 METHODOLOGY

Recently, Large Language Models (LLMs) and advanced tokenizers have revolutionized our understanding of language technologies and made significant advancements in the field. Their ability to create incredibly complex and rich context-based vector spaces opens a new area of analysis. Now, we are no longer limited by Word2Vec models but can analyze the vast variety of contextual meanings of individual words.

Thus, our first improvement of the method presented by Castro et al. [6] comes with the introduction of LLM embeddings instead of Word2Vec models. We use the pre-trained BERT Multilingual LLM [8] to extract the embeddings of tokens in our training datasets as shown in Figure 1.

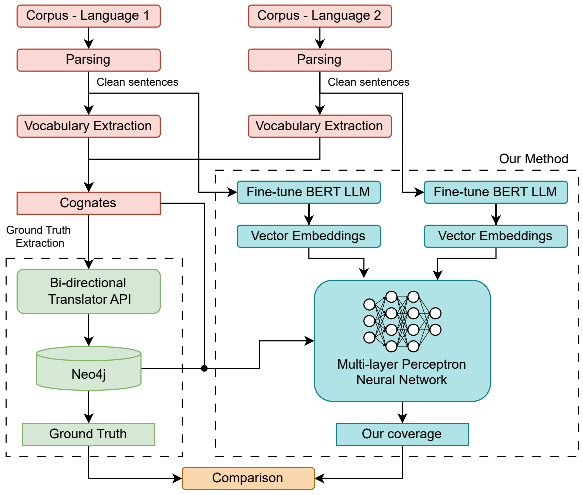


Figure 1: Our methodology

2.1 BERT as a Word to Vector Model

BERT (Bidirectional Encoder Representations from Transformers) [8] is a transformer-based model that has set new benchmarks in a variety of natural language and cross-language processing tasks [17]. Unlike previous models that processed text in a unidirectional way, BERT reads text bidirectionally, understanding the context of a word based on both its left and right surroundings. This bidirectional approach allows BERT to generate highly contextualized word embeddings.

The BERT transformer consists of multiple layers, where each layer is capable of capturing different aspects of the word's context. When we input a sentence into BERT, it tokenizes the sentence into subword units (tokens), processes these units through its multiple layers, and produces embeddings for each token at each layer. These embeddings are rich in context and can capture the nuances of word meanings in different sentences.

2.2 Embedding Extraction Process

To utilize BERT embeddings, we follow a systematic approach to extract and aggregate these embeddings:

(1) **Tokenization**: Using BERT's tokenizer, we split each word into its constituent subword tokens. This step ensures that even unfamiliar words or misspellings can be processed effectively by BERT.
(2) **Contextual Embedding Extraction**: We pass these tokens through the pre-trained BERT model to obtain embeddings. Since BERT embeddings are context-dependent, the same word can have different embeddings based on its surrounding words.
(3) **Averaging Token Embeddings**: For words split into multiple tokens, we compute the final word embedding by averaging the embeddings of all its constituent tokens. This aggregated embedding represents the word in its specific context within the sentence.

2.3 Classification of False-Friends

Instead of creating a linear transformation between vector spaces, we use the embeddings of pairs of words (correlated in both of the languages) and a training dataset of already classified pairs such as the Spanish-Portuguese dataset [7] to train a multi-layer perceptron neural network to classify pairs of unseen words as false or true friends.

The resulting embedding vector for each word is computed as the average of BERT's internal embeddings for each token comprising the word. Thus, leveraging the fixed internal embedding dimensions of BERT, where each token is represented by a vector $v \in \mathbb{R}^{768}$ space. We found that a simple dense neural network is enough for our methodology. This neural network has two hidden layers of 2000 neurons each, enough to get satisfiable results.

2.4 Creation of Ground Truth Dataset

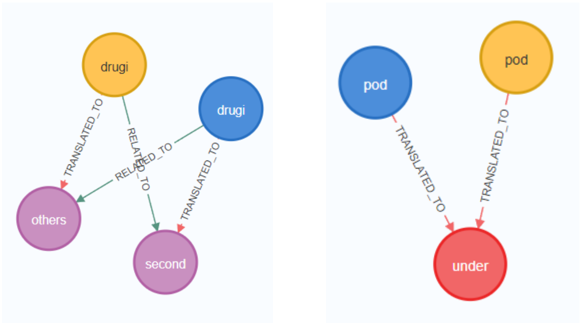
To extend the evaluation of our method, we needed to create a new ground truth dataset, which would consist of a collection of true and false friends. The prerequisite for obtaining said collection was processing a Slovenian [1] and a Macedonian corpus [3]. The former was obtained from The Slovenian Academy of Sciences and Arts, while the latter was obtained from the University of Leipzig. The Slovenian corpus was an official list of unique Slovenian words of 354205 different headwords, while the Macedonian corpus consisted of 350921 words obtained from Wikipedia. The latter was processed using our unique word extractor, which resulted in a unique word count of 248083.

These two lists of unique words were then further processed in order to extract homographs, which are words with the same spelling, in this case pertaining to Slovenian and Macedonian words. An initial hurdle was the difference in alphabets between the two languages. Slovenian uses the Latin alphabet, while Macedonian uses the Cyrillic alphabet. To overcome this, we transliterated the Macedonian corpus into the Latin alphabet. This allowed us to compare the two lists of unique words. We based our homograph extraction on a Levenshtein distance of 0, which meant that we only extracted homographs that were identical in spelling. Our extraction of homographs thus produced 21674 homographs and 226409 non-homographs. This stage of our corpus processing thus left us with 21674 candidates for false and true friends.

Our next step was to translate each Slovenian and Macedonian homograph to English and compare their English meanings. Those homographs that produced the same meaning were categorized as true friends, while the rest were categorized as false friends. This stage of our research made apparent a flaw in our translation API. The flaw being Google's translation API [2], which only returns one translation. Moreover, the limited scope of Slovenian and Macedonian, compounded by interjections, meant that some translations were inaccurate. Said inaccuracies then resulted in false positives, which were apparent in our Neo4j database.

The Neo4j database that we filled with Macedonian words, Slovenian words, true friends, and false friends was the backbone of our visualization. The latter helped us identify potential problems with our approach, such as improper false friend connections, example given in Figure 2a, and true friend connections due to limited responses from the Google Translate API, example given in Figure 2b.

Our analysis of the Neo4j database thus yielded a lot of food for thought. An appealing approach was the classification of false friends into segregation (pairs carrying absolutely different meanings), lexical pairs (both similar and dissimilar meanings), and inclusion (one dissimilar meaning on top of all other similar meanings) as outlined in [11]. But we decided to stick with the binary classification of false and true friends.



(a)

(b)

Figure 2: (a) The Slovenian word "drugi" could be translated as "others", which would match it with the Macedonian word "drugi" (други), or translated as "second", which results in a false friend. "Drugi" is, therefore, only 50% a false friend. (b) The Macedonian word "pod" (под) could, likewise, be alternatively translated as floor, which means that it could potentially be a false friend.

3 RESULTS

To compare and benchmark our approach, we recreated the results of the method used by Castro et al. [6]. Their method included acquiring the then-newest Wikipedia dumps (dated 20.03.2024), parsing them, and training the two Word2Vec models [14] in Spanish and Portuguese. Each model with a vector dimension of 100 took an hour and a half to train on 30 million sentences, after which we could finally derive the linear transformation necessary for translation. Their paper evaluates the method by classifying a pair of words as true or false friends given a ground truth dataset between Spanish and Portuguese [7]. Our re-testing of their method achieved the results shown in Table 2.

Our proposed method, however, showed a significant improvement on the Spanish-Portuguese dataset with an F1-score of 0.8381, shown in Table 3, without any fine-tuning on the pre-trained BERT Multilingual model. Moreover, because of the pretrained BERT model, our method including the extraction of embeddings and the training of the neural network to learn the classification of the words took under 10 minutes as opposed to the training time of the Word2Vec models of around 3 hours.

Table 2: Classification performance using the Castro et al. [6] approach on the Spanish and Portuguese dataset.

	Precision	Recall	F1-Score
False	0.7727	0.7730	0.7721
True	0.7446	0.7424	0.7427
Average	0.7586	0.7577	0.7574

Table 3: Comparison of our and Castro et al. methodology. Note that we are not using any additional sentences for fine-tuning the Multilingual BERT Model.

F1	Castro et al.		Ours
Sentences	**30M**	**200K**	**0**
False	0.7721	0.7324	**0.8505**
True	0.7427	0.5783	**0.8258**
Average	0.7574	0.6554	**0.8381**

Table 4: Classification performance of our approach on our Macedonian and Slovenian false and true friends datasets without fine-tuning of the BERT model.

	Precision	Recall	F1-Score
False	0.8077	0.8750	0.8400
True	0.8125	0.7222	0.7647
Average	0.8101	0.7986	0.8024

Our method of extracting homographs from our Slovenian and Macedonian corpus yielded 21674 candidates for false/true friends. Further analysis using comparisons of associated English meanings resulted in 14654 true and 7020 false friends. However, some of these were false positives due to the multi-meaning nature of various words. A manual review of the 7020 false friends gave us 151 ideal false friends that are largely free of true friend overlap. We used these 151 ideal false friends as our Slovenian-Macedonian false friends dataset. The false friends manual review was followed by an extraction of 268 true friends from our initial set of 14654 true friends. These 268 true friends then comprised our Slovenian-Macedonian true friends dataset.

Using our Slovenian-Macedonian dataset of false and true friends[1], we achieved similar classification capabilities as with the Spanish-Portuguese dataset. Our results can be seen in Table 4. All experiments were run on an Intel Core i7-9700K @ 3.60GHz and GeForce RTX 2070 SUPER GPU.

4 CONCLUSIONS

In this paper, we presented a novel approach to exploring cross-linguistic connections, specifically focusing on false friends, using Large Language Model embeddings and graph databases. Our methodology leverages the advanced capabilities of BERT for generating contextualized word embeddings and a graph-based representation to capture semantic relationships. We achieved classification performance on the Spanish-Portuguese false friend dataset with an F1 = 83.81% and classification performance on our Slovenian-Macedonian dataset of F1 = 80.24% using Multilingual BERT and a multi-layer perceptron neural network. BERT was not fine-tuned using any additional sentences.

Our results indicate that LLM embeddings significantly enhance the accuracy of false friend classification compared to traditional Word2Vec models. The use of a pretrained LLM also significantly reduced the time it takes to learn the classifications from 3 hours needed to train the Word2Vec models to under 10 minutes solely for the training of the multi-layer perceptron classifier. This highlights the potential of using sophisticated language models for even more complex linguistic tasks, paving the way for more accurate and insightful cross-linguistic analysis.

A natural next step to enhancing our methodology would be incorporating larger and more diverse corpora. These would fine-tune the pre-trained BERT model on specific language pairs or domains, improving the contextual accuracy of embeddings. Moreover, larger corpora would yield additional false and true friends in our ground truth dataset. More advanced translation APIs would be capable of providing multiple translations for each word, which would result in fewer false positives when creating such a dataset.

Furthermore, extending the methodology to other cross-linguistic phenomena, such as idiomatic expressions, cognates, and loanwords, would improve our understanding of language relationships. False and true friends are, therefore, the tip of a linguistic iceberg that calls for further exploration.

REFERENCES

[1] 2013. ISJ SAZU - List of Slovenian words. http://bos.zrc-sazu.si/sbsj_en.html [Online; accessed 2. Jun. 2024].

[2] 2024. Cloud Translation API | Google Cloud. https://cloud.google.com/translate/docs/reference/rest [Online; accessed 2. Jun. 2024].

[3] 2024. Wortschatz Leipzig Macedonian Corpora. https://wortschatz.uni-leipzig.de/en/download/Macedonian [Online; accessed 2. Jun. 2024].

[4] Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013. Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. *ACL Anthology* (Aug. 2013), 48–57. https://aclanthology.org/W13-2408

[5] E. Susanne Carroll. 1992. On cognates. *Sage Journals* 8 (jun 1992). Issue 2. https://journals.sagepub.com/doi/abs/10.1177/026765839200800201

[6] Santiago Castro, Jairo Bonanata, and Aiala Rosá. 2018. A High Coverage Method for Automatic False Friends Detection for Spanish and Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann, Shervin Malmasi, and Ahmed Ali (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 29–36. https://aclanthology.org/W18-3903

[7] María de Lourdes Otero Brabo Cruz. 2004. Diccionario de falsos amigos (espanol-portugues / portugues-espanol). https://ec.europa.eu/translation/portuguese/magazine/documents/folha47_lista_pt.pdf

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ACL Anthology* (June 2019), 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Ljiljana Dolamic and Jacques Savoy. 2009. Indexing and stemming approaches for the Czech language. *Information Processing & Management* 45, 6 (Nov. 2009), 714–720. https://doi.org/10.1016/j.ipm.2009.06.001

[10] Pedro Domínguez and Brigitte Nerlich. 2002. False friends: Their origin and semantics in some selected languages. *Journal of Pragmatics* 34 (Dec. 2002). https://doi.org/10.1016/S0378-2166(02)00024-3

[11] Ketevan Gochitashvili and Giuli Shabashvili. 2018. The issue if "false friends" in terms of learning a foreign language(Using the example of Georgian and English languages). *International Journal Of Multilingual Education* VI (July 2018), 33–41. https://doi.org/10.22333/ijme.2018.11006

[12] Diana Inkpen and Oana Frunza. 2005. Automatic Identification of Cognates and False Friends in French and English. *ResearchGate* (Jan. 2005). https://www.researchgate.net/publication/237129220_Automatic_Identification_of_Cognates_and_False_Friends_in_French_and_English

[13] Nikola Ljubešić and Darja Fišer. 2013. Identifying false friends between closely related languages. *ACL Anthology* (Aug. 2013), 69–77. https://aclanthology.org/W13-2411

[14] Long Ma and Zhang Yanqing. 2015. Using Word2Vec to Process Big Text Data. (Oct. 2015). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7364114

[15] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. 1–12. https://www.researchgate.net/publication/319770439_Efficient_Estimation_of_Word_Representations_in_Vector_Space

[16] Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation* 21, 1 (March 2007), 29–53. https://doi.org/10.1007/s10590-008-9034-5

[17] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4996–5001. https://doi.org/10.18653/v1/P19-1493

[18] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving Text Embeddings with Large Language Models. *arXiv* (Dec. 2023). https://doi.org/10.48550/arXiv.2401.00368 arXiv:2401.00368

[1] The datasets are available at https://github.com/mitkonikov/false-friends