False Loan Word Detection in Multilingual Contexts

---

1. Introduction

In multilingual environments, it is common to encounter words that appear phonetically identical across different languages yet carry divergent meanings. Such words often lead to confusion and misinterpretation. A "loan word" refers to a term borrowed from one language and integrated into another. However, not all perceived loan words are authentic; some are mistakenly assumed to be borrowed when they are not. These are referred to as "false loan words." Detecting such false loan words is important for improving cross-lingual translation systems, language learning tools, and lexical resources.

This project focuses on identifying false loan words that exist between languages such as German and English. The central challenge lies in the phonetic similarity that masks the semantic dissonance between the two languages.

2. Problem Statement

The core problem is to **detect false loan words embedded in natural language sentences**. For instance, in German:

> "Der Amtmann führte die Regierungstreffen durch, wo sich die Politiker über wichtige Angelegenheiten unterhielten."

This sentence contains the word *Amtmann*, which, though it appears to resemble a formal English term, is not a direct loan word but a culturally and semantically distinct term.

Our task is to automatically detect such words when they appear in context, differentiate them from true loan words, and assess the accuracy of language models in recognizing and resolving such ambiguities.

3. Dataset Description

We curated a dataset containing the following fields:

- `loan_word`, `original_word`: The observed word and its corresponding original form.
- `loan_word_epitran`, `original_word_epitran`: Phonetic transcription via Epitran.
- `loan_english`, `original_english`: English equivalents.

- Multiple similarity metrics: Fast Levenshtein, Dolgo Prime Distance, Feature Edit Distance, etc.
- `label`: True class (loan, synonym, false)
- `label_bin`: Binary label (1 for loan, 0 for synonym/false)
- `MBERT_cos_sim`, `XLM_cos_sim`: Cosine similarity scores from multilingual BERT and XLM-R models.

This data allowed us to analyze both phonetic and semantic aspects of loan words, facilitating a rich feature set for downstream modeling.

	loan_word	original_word	generated_context	reference_sentence	translated_sentence	reference_word
0	Mirth	Fröhlichkeit	Die Frau lächelte plötzlich und sagte: "Ich fü...	The woman smiled suddenly and said, "I feel ve...	The woman smiled suddenly and said: "I feel ve...	Cheerfulness
1	Schnorr	Chromatogramm	Der Chemiker studierte die Chromatogramme der ...	The chemist studied the chromatograms of the v...	The chemist studied the chromatograms of diffe...	Chromatogram
2	Zettelkasten	Zettelkasten	Ich habe meine Zettelkasten vollgefüllt, um al...	I filled up my paperbox to write down all the ...	I have filled my notebook with all ideas and t...	Paper box
3	Meiring	Meiring	Der kleine Hund rannte durch den Wald, um nach...	The little dog ran through the forest to look ...	The little dog ran through the woods in search...	Meiring
4	Speth	Speth	Der Speth fuhr durch die Felder, um frische Ge...	The Speth drove through the fields to buy fres...	The farmer drove through the fields to buy fre...	Speth
...	...	...	...	...	...	...
5289	Meisinger	Meisinger	Der kleine Hund rannte durch den Wald, um nach...	The little dog ran through the forest to look ...	Der kleine Hund rannte durch den Wald, um nach...	Meisinger
5290	Frankenberger	bleiben lassen	Der Hund bleibt lassen, wenn man ihn nicht füt...	Keep the dog if you don't feed it.	The dog will not leave if you do not feed him.	Keep
5291	esteemed	geehrt	Mein Vater gehert mich immer noch.	My father's still insinuating me.	My father still loves me.	Honored
5292	Meier	Kauffmann	Der kleine Kaufmann kaufte ein Stück Brot auf ...	The little merchant bought a piece of bread in...	The small merchant bought a loaf of bread on t...	Kauffmann
5293	aldehyde	Aldehyd	Der Chemiker studierte die Eigenschaften der n...	The chemist studied the properties of the new ...	The chemist studied the properties of the new ...	aldehyde

5294 rows × 9 columns

4. Approach

Our pipeline consists of two major iterations:

Step 1: Translation & Embedding Analysis

We begin with a source sentence containing a suspected loan word. Using the OPUS-MT translation model, we obtain a ground truth translation. Simultaneously, an LLM (e.g., LLaMA-3) provides a parallel translation.

Example:

- Source (German): "Der Amtmann führte die Regierungstreffen durch..."
- Ground Truth Translation: "The official conducted the government meetings..."
- LLM Translation: "The Amtmann led the government meetings..."

This contrast reveals that the LLM retains the loan word (*Amtmann*) while the OPUS model contextualizes it as *official*.

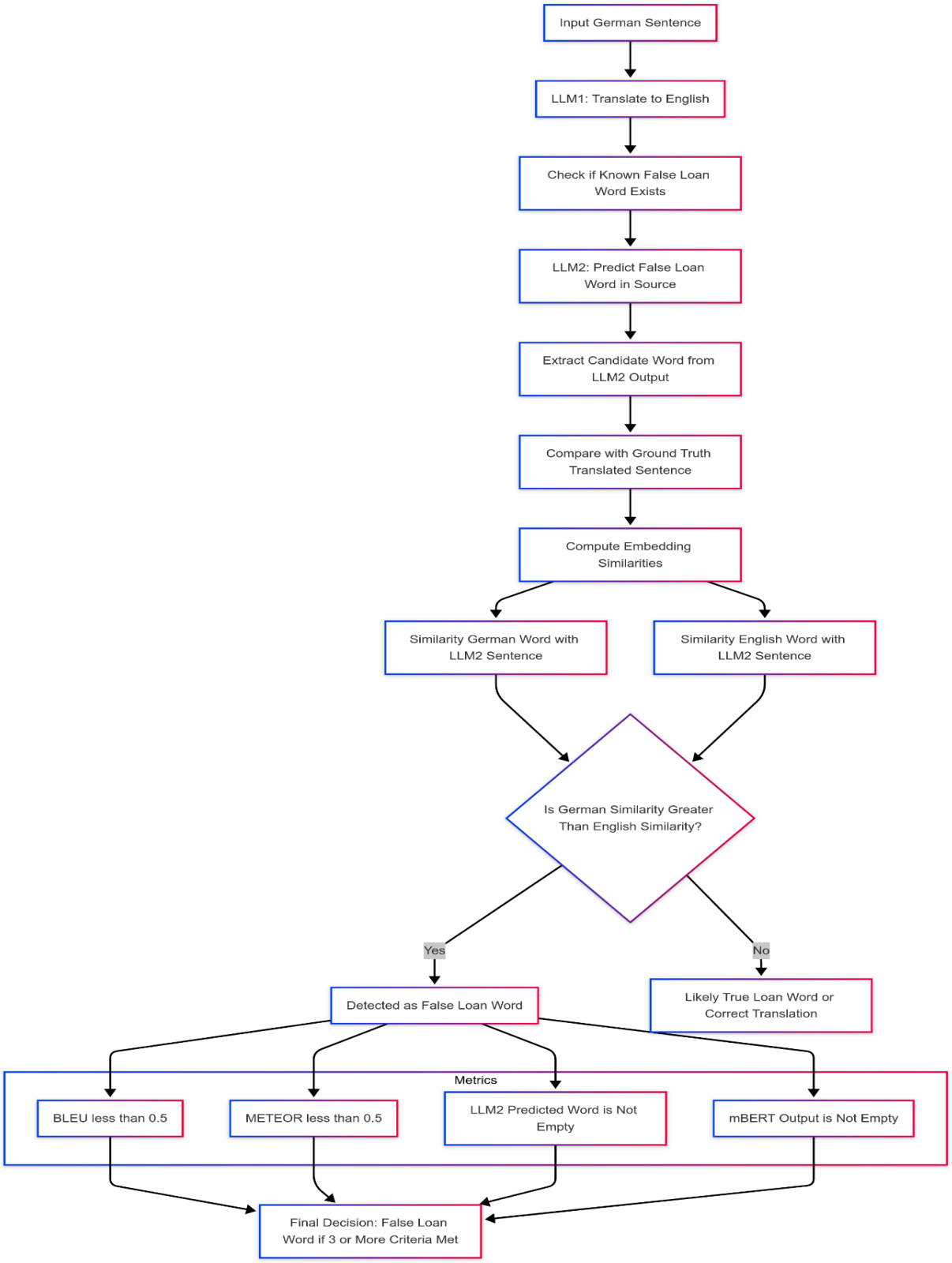
Step 2: Iterative False Loan Word Detection

A second iteration using a different LLM is prompted to directly predict the false loan word(s) from the sentence. We then:

- Measure cosine similarity between:
    - False loan word and the LLM-translated sentence
    - Ground truth word (e.g., *official*) and the same sentence

If the LLM translation aligns more closely with the false loan word than the true equivalent, it suggests the model did not resolve the semantic mismatch, strengthening the false loan hypothesis.

Heuristic Rule-Based Classifier



A word is labeled as a **false loan** if:

1. BLEU score < 0.5 between ground truth and LLM translation
2. METEOR score < 0.5
3. LLM predicts at least one false loan word
4. Our fine-tuned mBERT also predicts a false loan word

If at least 3 out of 4 conditions are met, the word is marked as a likely false loan.

5. Metrics and Model Ensemble

Evaluation Metrics:

- **BLEU Score**: Measures n-gram overlap with reference translation
- **METEOR Score**: Incorporates synonymy, stemming, and precision/recall
- **Cosine Similarity**: Between embeddings of the predicted word and translation context
- **Accuracy**: Final classification accuracy over the dataset

Ensembling:

- Voting ensemble over the four conditions mentioned above
- Final classification score is computed via majority agreement

---

6. Results

Below are a few sample outputs from the pipeline:

	loan_word	original_word	generated_context	reference_sentence	translated_sentence	reference_word	label	bleu_score	meteor_score	false_loanword_model	false_loanword	sim_eng_ger	similarity_word_eng_ger	fl_ger
14	enzyme	Enzym	Der Enzym, der die Zucker zu Sauerstoff umwand...	The enzyme that converts sugar into oxygen is ...	Der Enzym, der die Zucker zu Sauerstoff umwand...	Enzyme	loan	0.022870	0.000000	[zucker]	Der	[0.7002429, 0.8326885 7]	[[0.3908607, 0.9491485]]	[zucker]
46	Pischke	Pischke	Der Pischke im Garten ist sehr groß und hat vi...	The pishke in the garden is very large and has...	The pichk is a type of bird, not an animal tha...	Pishke	loan	0.058166	0.122951	[pichk, an]	pichk	[0.594713, 0.65164655]	[[0.594713, 0.65164655], [0.24290466, 0.692956]]	[pichk, an]
54	Wermuth	Wermuth	Der kleine Hund ran schnell um den Wermuth herum.	The little dog quickly ran around the wormwood.	The little dog ran quickly around the worm.	Wermuth	loan	0.250986	0.793367	[]	worm	[0.2375343, 0.714456]	[]	[]
58	Reifsteck	Reifsteck	Der Reifstein war sehr trocken, daher musste i...	The maturation stone was very dry, so I had to...	Der Reifstein war sehr trocken, daher musste i...	Mature plug	loan	0.024178	0.000000	[]	Der	[0.586233, 0.77951574]	[]	[]
72	Blumenkopf	Blumenkopf	Der Blumenkopf taucht langsam aus dem Wald.	The flower head slowly dives out of the forest.	The Blumenkopf is slowly emerging from the for...	Flower head	loan	0.180444	0.354635	[blumenkopf]	The	[0.37544906, 0.39622292]	[[0.76703185, 0.76703185]]	[]
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5256	flehman	flehmen	Der Hund flöhmen konnte nicht mehr, weil er mü...	The dog couldn't flea anymore because he was t...	The dog could no longer cry because he was tired.	Floats	loan	0.403528	0.647131	[]	no	[0.27304706, 0.3717686]	[]	[]
5261	Bechtel	Bechtel	Der Politiker, der für die Landesregierung kan...	The politician who ran for the state governmen...	The politician who ran for state government wa...	Weight	loan	0.431949	0.663511	[government]	Bechtle	[0.62812185, 0.7227488]	[[0.30271804, 0.6824347]]	[government]
5276	Roberg	Roberg	Der Tourist fand den Roberg am Strand sehr schön.	The tourist found the Roberg on the beach very...	The tourist found the Robertsgate at the beach...	Roberg	loan	0.350844	0.691837	[beach]	Robertsgate	[0.6308725, 0.7515869]	[[0.2992113, 0.72408533]]	[beach]
5282	Schwaller	Schwaller	Der Schwaller von der Stadt führte mich zum ki...	The swarm from the city led me to the little c...	The Schwaller von der Stadt is likely a misspe...	Swallows	loan	0.011525	0.031646	[stadt, district, stadt, mich, café.]	Schwabing	[0.59586823, 0.67951965]	[[0.37616584, 0.6403628], [0.16528827, 0.64619...	[stadt, district, stadt, mich, café.]
5289	Meisinger	Meisinger	Der kleine Hund rannte durch den Wald, um nach...	The little dog ran through the forest to look ...	Der kleine Hund rannte durch den Wald, um nach...	Meisinger	loan	0.024456	0.000000	[rannte, meisinger-schokoladenladen]	Der	[0.37442356, 0.8250093]	[[0.29013312, 0.87097377], [0.827708, 0.820520...	[rannte]

(*The attached showing overall system accuracy: **65.79%** for final prediction*)

14	enzyme	Enzym	Der Enzym, der die Zucker zu Sauerstoff umwand...	The enzyme that converts sugar into oxygen is ...	Der Enzym, der die Zucker zu Sauerstoff umwand...	Enzyme	loan	0.022870	0.000000	[zucker]	Der	[0.7002429, 0.8326857]	[[0.3908607, 0.9491485]]	[zucker]
32	Klees	Klees	Der Kleesmann kümmerte sich um die Reinigung d...	The Kleesmann took care of the cleaning of the...	The translator of the given German sentence is...	Klees	loan	0.048150	0.292245	[kleesmann]	Kleesmann	[0.7690670 5, 0.6750045]	[[0.7690670 5, 0.6750045]]	[]
46	Pischke	Pischke	Der Pischke im Garten ist sehr groß und hat vi...	The pishke in the garden is very large and has...	The pichk is a type of bird, not an animal tha...	Pishke	loan	0.058166	0.122951	[pichk, an]	pichk	[0.594713, 0.65164655]	[[0.594713, 0.65164655], [0.24290466, 0.692956]]	[pichk, an]
54	Wermuth	Wermuth	Der kleine Hund ran schnell um den Wermuth herum.	The little dog quickly ran around the wormwood.	The little dog ran quickly around the worm.	Wermuth	loan	0.250986	0.793367	[]	worm	[0.2375343, 0.714456]	[]	[]
58	Reifsteck	Reifsteck	Der Reifstein war sehr trocken, daher musste i...	The maturation stone was very dry, so I had to...	Der Reifstein war sehr trocken, daher musste i...	Mature plug	loan	0.024178	0.000000	[]	Der	[0.586233, 0.77951574]	[]	[]

```python
filtered_rows = []
for _, row in filtered_df.iterrows():
    sim_eng_ger = row['sim_eng_ger']

    # Ensure sim_eng_ger is a list-like structure (e.g., "[0.2, 0.8]")
    if isinstance(sim_eng_ger, str):
        try:
            sim_eng_ger = eval(sim_eng_ger)  # Convert string representation of list to actual list
        except Exception as e:
            print(f"Error parsing sim_eng_ger: {e}")
            continue

    # Check if German embedding similarity is higher
    if isinstance(sim_eng_ger, list) and len(sim_eng_ger) == 2:
        eng_similarity, ger_similarity = sim_eng_ger
        if ger_similarity > eng_similarity:
            filtered_rows.append(row)

loan_df = pd.DataFrame(filtered_rows)
```

[16] Python

```python
print(len(loan_df)/len(filtered_df))
```

[17] Python

```
0.6579476861167002
```

7. Conclusion

Through this two-step iterative LLM-BERT pipeline, we have demonstrated the ability to successfully detect false loan words across multilingual contexts. Our model combines semantic embeddings, translation accuracy metrics, and LLM predictions to reliably flag false loans.

Key contributions include:

- Construction of a phonetic-semantic aligned dataset
- Fine-tuned mBERT for word-level classification
- Multi-metric ensemble for robust decision-making

The project achieves an **accuracy of 65.79%** on identifying false loan words in German-English pairs. We have shown that LLMs, when prompted effectively, can identify and resolve complex linguistic ambiguities.