Cheap Character Noise for OCR-Robust Multilingual Embeddings

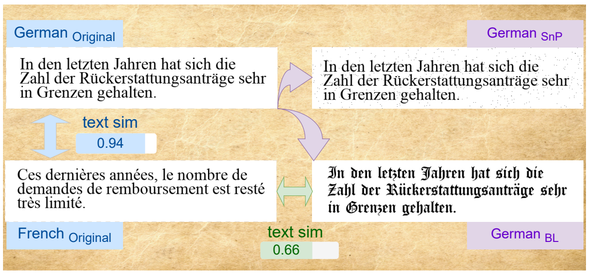
Andrianos Michail Juri Opitz Yining Wang Robin Meister
Rico Sennrich Simon Clematide
University of Zurich
<firstname>.<lastname>@uzh.ch

Abstract

The large amount of text collections digitized by imperfect OCR systems requires semantic search models that perform robustly on noisy input. Such collections are highly heterogeneous, with varying degrees of OCR quality, spelling conventions and other inconsistencies —all phenomena that are underrepresented in the training data of standard embedding models, with ramifications for their generalization. In our paper, we show that this problem can be alleviated with a simple and inexpensive method that does not require supervision or in-domain training. Specifically, we fine-tune existing multilingual models using noisy texts and a contrastive loss. We show that these models show considerable improvements across different noise conditions. Control experiments indicate minimal, and occasionally positive, impact on standard similarity tasks. These findings suggest that embedding models can be inexpensively adapted for cross-lingual semantic search in heterogeneous, digitized corpora. We publicly release our code, datasets, and models at https://github.com/impresso/ocr-robust-multilingual-embeddings.

1 Introduction

Optical Character Recognition (OCR) technology plays a central role in the digitization of historical documents, rendering large volumes of textual data accessible to Natural Language Processing (NLP) tools. To enable effective retrieval from such collections, semantic search based on text embeddings presents an appealing approach (Reimers and Gurevych, 2019; Gao et al., 2021). However, OCR output is inherently imperfect and often introduces errors due to factors such as poor image quality, complex layouts and fonts, low contrast and degradation of the source material (Dhingra et al., 2008). Consequentially, the noisy OCR output can severely degrade the performance of NLP systems by disrupting the syntactic and semantic structure of the text, possibly further exacerbated in multilingual contexts, where models show different performance degradations for different input languages (Lopresti, 2008; Todorov and Colavizza, 2022). A particular problem is that this noise disrupts tokenization processes, resulting in fragmented or incorrect sub-word units (van Strien et al., 2020; Todorov and Colavizza, 2022). Figure 1 showcases an example of ramifications for text embedding models: A multi-lingual embedding model assigns vastly different similarities to two parallel texts, just depending on their font before undergoing the digitization process.



In den letzten Jahren hat sich die Zahl der Rückerstattungsanträge sehr in Grenzen gehalten.

In den letzten Jahren hat sich die Zahl der Rückerstattungsanträge sehr in Grenzen gehalten.

Ces dernières années, le nombre de demandes de remboursement est resté très limité.

In den letzten Jahren hat sich die Zahl der Rückerstattungsanträge sehr in Grenzen gehalten.

Figure 1: Same meaning, but different similarity scores. The cosine similarity assigned by the *multilingual-gte-base* model differs by 28 points when comparing clean French–German sentence pairs to a version in which the German sentence has been OCRed from a Blackletter font. The distorted output contains errors such as “Ja**b**ren”, “**Nii**ckerstattungsanträge”, and “Gren**s**en ge**b**alten”.

In this work, we empirically analyze how OCR-induced noise affects multilingual embedding models and propose a method based on contrastive fine-tuning with randomly noised text. Our main contributions are:

1. We introduce a simple, cheap, and unsupervised adaptation strategy that improves the robustness of off-the-shelf multilingual embedding models to OCR noise in digitized text collections.

Findings of the Association for Computational Linguistics: ACL 2025, pages 11705–11716
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

2. We evaluate our method on challenging text similarity tasks in German and French, using digitized data with diverse types of OCR noise—including realistic, historically difficult conditions (e.g., Blackletter fonts) and minimal-noise, modern digitization. Despite its simplicity, our approach improves performance across all conditions, including on a language unseen during training. Moreover, it does not degrade performance on clean text similarity benchmarks.

3. We conduct detailed analyses of training strategies and investigate subtokenization mismatches as a potential source of performance degradation in the off-the-shelf models (an issue that our method then helps to mitigate).

2 Background and Related Work

Semantic search refers to a class of tasks aimed at retrieving texts that are semantically similar, or relevant to a given query. Today's dominant paradigm to address this tasks is embedding texts as vectors with pre-trained encoders fine-tuned on contrastive tasks (Reimers and Gurevych, 2019). In many tasks, this technique has been proven efficient and effective. However, these tasks typically involve contemporary and more normative text. Hence, it is unclear whether the models maintain their performance in noisy and highly heterogeneous text collections, as we would encounter in large digitized collections of text, e.g., within institutions such as libraries.

Typo Robustness. A related line of research investigates the robustness of dense retrieval and embedding models to typos and character-level perturbations. While OCR errors and human typos differ in origin, both introduce noise that disrupts tokenization and embedding similarity. Tasawong et al. (2023) propose a robust dense retrieval training method that aligns representations of pristine and misspelled queries while enforcing contrast to distinguish unrelated inputs. Their approach combines query augmentation with a dual self-teaching loss, significantly improving performance on both synthetic and real-world misspelled queries without sacrificing accuracy on clean inputs. Similarly, Sidiropoulos and Kanoulas (2022) examine the vulnerability of dual encoder architectures to misspellings, finding that even minor character-level changes can substantially degrade retrieval performance. They propose contrastive learning-based strategies to mitigate this sensitivity.

Training models on character-level noise has been shown to be effective for machine translation (Belinkov and Bisk, 2018; Vaibhav et al., 2019; Sperber et al., 2017) and cross-lingual transfer between closely related languages (Aepli and Sennrich, 2022). We demonstrate that multilingual embedding models have worse semantic search capabilities within digitized collections due to the presence of OCR errors. We then evaluate versions of further training approaches that aim to bring closer the representations of randomly noised text with its original version and show increased capabilities of semantic search in digitized collections even for languages pairs not included in our noised fine-tuning.

3 Notation and Preliminaries

Throughout this work, we make use of multilingual parallel datasets that may contain noise of different degree. In particular, the term *parallel* here does not necessarily only mean that there are pairs with texts from two different languages, but it can also mean, e.g., that the dataset is monolingual but contains pairs of texts where one part contains OCR noise. To denote such a dataset we use $\textbf{D}\textsc{\textbf{ata}}\textbf{S}\textsc{\textbf{et}}^{\boldsymbol{ns\rightarrow nt}}_{\boldsymbol{ls\rightarrow lt}}$. In this notation, $\boldsymbol{ns}$ ($\boldsymbol{nt}$) are variables that denote the type of observed noise, whereas $\boldsymbol{ls}$ ($\boldsymbol{lt}$) refer to the source (target) language within the dataset (if they differ). We formalize forms of OCR noise within the following categories:

- *Random Noise* (**RN**): A fully unsupervised stochastic noise. This will only be used for training in our proposed method.

- *Minimal Noise* (**MN**): Simplest of noise produced by OCR engines when the fonts, font size and resolution of the image is clear.

- *Salt and Pepper Noise* **(SnP)**: A noising condition which emulates tear and wear on paper by adding small (very subtle) sprinkles of ("Salt") and ("Pepper") throughout the scan, hence the name SnP. See also the example in Figure 1, top right.

- *BlackLetter* **(BL)**: Blackletter fonts, also known as Gothic script or Fraktur (Figure 1, bottom right), were prevalent in German-speaking regions until the early 20th century. These typefaces are characterized by their

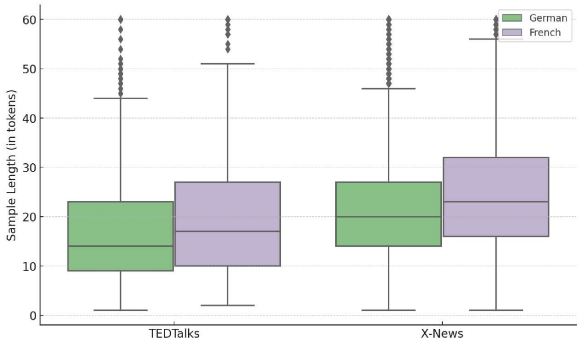


Figure 2: Lengths per sentence in tokens for German and French in the TED and X-News corpora.

dense, angular letterforms. Since Blackletter wasn't used in French texts, for French texts we assume an equivalent (in terms of OCR parsing difficulty) of *Scanned Distorted Noise **(SD)***. This is characterized by irregular distortions and artifacts introduced during the scanning or image processing stages, which make the text difficult to recognize accurately by OCR systems.

The languages that we mention in our paper are German (de), French (fr), English (en), Luxembourgish (lb), Turkish (tr) and Arabic (ar). For example, in a parallel DATASET, we consider a case where French source sentences in standard clean text are paired with German text OCRed from Blackletter font. We denote this dataset as $\text{DATASET}^{\varnothing\to\text{BL}}_{\text{fr}\to\text{de}}$. A full description of the generation process for our multilingual noised training data and a separate description of generating the realistic OCR evaluation data follows in §5.

4 Method Overview

The core idea of our method is straightforward: We generate a training set by inducing random character-level noise into a given dataset. This enables us to fine-tune a pre-trained embedding model to better align clean and character-noised versions of the same text using a contrastive loss objective. Our study focuses on lightweight adaptations of multilingual models, requiring only 20,000 positive pairs and modest computational resources (approximately 20 minutes on a single Tesla T4 16GB GPU). Because the method is fully unsupervised and language-agnostic, it can be easily applied to adapt multilingual embedding models for improved robustness to OCR-induced noise.

5 Experimental Setup

5.1 Creating Noised Training Data

To train our proposed models, we use two bilingual corpora. The first is the NeuLabs (**TED**) corpus (Qi et al., 2018), which contains French and German translations of TED Talk transcripts. The second is the ELRC-CORDIS News corpus (**X-News**), a parallel dataset of French and German sentences from the news domain—our primary application focus. For fine-tuning, we sample 20,000 positive pairs from each corpus. Descriptive statistics for both datasets are shown in Figure 2.

We augment the texts in these datasets by generating noisy variants through stochastic, unsupervised character-level modifications that simulate typical OCR errors. Specifically, we apply character-level perturbations to 5% of the characters in the original text. These modifications are randomly drawn from three types of noise: (1) **substitution**, where a character is replaced with a randomly selected alternative to mimic misrecognition; (2) **insertion**, where a random character is inserted at an arbitrary position; and (3) **deletion**, where a randomly chosen character is removed. This process produces a noisy version of each sentence, which we then pair with its original clean counterpart to form positive training examples.

We refer to the resulting datasets as ***noised TED*** and ***noised X-News***. In each dataset, one side contains the original (clean) text, while the other contains its stochastically noised counterpart. In our dataset notation, *noised TED* is defined as $\text{TED}^{\varnothing\to\text{RN}}_{\text{de}\to\text{de}} \cup \text{TED}^{\varnothing\to\text{RN}}_{\text{fr}\to\text{fr}}$, comprising monolingual sentence pairs. In contrast, *noised X-news* is defined as $\text{X-News}^{\varnothing\to\text{RN}}_{\text{de}\to\text{fr}} \cup \text{X-News}^{\varnothing\to\text{RN}}_{\text{fr}\to\text{de}}$, containing cross-lingual sentence pairs. Both datasets consist exclusively of positive pairs and are used to fine-tune embedding models for greater robustness to OCR noise.

5.2 Main Evaluation Datasets

We evaluate the effectiveness of our proposed adaptation method using cross-lingual datasets that include realistic and naturally occurring OCR noise. This evaluation is organized in two parts: **Evaluation task** and **Inducing realistic OCR noise**.

5.2.1 Evaluation task.

We adopt the Cross-Lingual Semantic Discrimination (CLSD) German-French evaluation benchmark introduced by Michail et al. (2025a). This task assesses whether multilingual embedding models can accurately identify the correct cross-lingual semantic match in the presence of challenging distractors. It is based on parallel sentence pairs from the WMT19 and WMT21 DE–FR news test sets (Barrault et al., 2019; Akhbardeh et al., 2021). Specifically, each datum consists of a source sentence, its correct translation in the target language, and a set of four semantically similar distractor sentences in the target language. The model must produce an embedding for the source sentence that is more similar to the target sentence than to any distractors. Performance is evaluated using accuracy, reported as Precision@1.

5.2.2 Inducing realistic OCR noise

We then construct our primary evaluation datasets by generating three realistic OCR noise variants of the CLSD dataset: **Minimal Noise (MN)**, **Blackletter (BL)**, and **Salt and Pepper (SnP)**.

To produce the Minimal Noise (**MN**) realistic OCR noise, we print the text in font *Times New Roman* at font size 10 and save it as an image at 300 pixels per inch (PPI). We then apply OCR using Tesseract 3[1], a widely used open-source OCR engine. This setup introduces light OCR errors, resulting in average character error rates (CER) of 0.4% for German and 0.6% for French.

For the **Blackletter (BL)** condition, we follow the same procedure but render the German text using Canterbury, a blackletter-style font commonly found in historical print. We pair this with a **Scanned Distorted (SD)** variant on the French side, which simulates scanning artifacts by introducing horizontal offsets and random spacing distortions within characters. This setting yields CERs of 2.8% for German and 2.4% for French.

For the **Salt and Pepper (SnP)** condition, we again follow the base procedure but add synthetic background noise by randomly scattering black and white pixels at a density of 0.45%. This visual noise mimics degradation due to paper aging or scanning artifacts, producing average CERs of 5.4% for German and 5.1% for French.

In evaluation, we annotate the source and target language noise types independently. For instance, evaluating CLSD WMT19 where both the French source and German target contain SnP noise is denoted as $\mathrm{WMT19}^{\mathrm{SnP}\rightarrow\mathrm{SnP}}_{\mathrm{de}\rightarrow\mathrm{fr}}$.

5.3 Control Task and Dataset

To ensure that our adaptation method does not negatively affect performance on standard semantic search tasks involving clean text, we conduct a set of control experiments. Specifically, we assess whether improvements on OCR-noisy inputs come at the cost of degrading the model's ability to handle more normative, noise-free text.

Thus, as a baseline for our evaluations, we use the clean (OCR-free) variant of the CLSD dataset, along with the multilingual Semantic Textual Similarity (STS) benchmark introduced by Cer et al. (2017). For STS, we focus on language pairs that are not present in our adaptation training data: $\mathrm{STS}_{\mathrm{en}\leftrightarrow\mathrm{tr}}$ (English–Turkish), $\mathrm{STS}_{\mathrm{en}\leftrightarrow\mathrm{es}}$ (English–Spanish), and $\mathrm{STS}_{\mathrm{en}\leftrightarrow\mathrm{ar}}$ (English–Arabic).

5.4 Historic Luxembourgish Bitext Mining

We further evaluate our adapted models using the Historical Bitext Mining task introduced by Michail et al. (2025b), referred to here as HISTLUX. This dataset consists of digitized historical Luxembourgish newspaper articles (published between 1840 and 1950), which have been segmented into sentences and manually translated into clean, modern German and French. As it contains naturally occurring digitization errors, HISTLUX provides a valuable testbed for assessing model robustness beyond simulated OCR noise.

5.5 Embedding Models

We evaluate our adaptation strategies using two state-of-the-art multilingual embedding models:

- **M-E5B**: Proposed by Wang et al. (2024a,b), this model is trained via weakly supervised contrastive pre-training on multilingual data, followed by supervised fine-tuning on information retrieval tasks. We use the multilingual-E5-base version, available on Hugging Face.
- **M-GTE**: Introduced by Zhang et al. (2024), this model is based on a multilingual long-context encoder and trained using a similar contrastive objective to M-E5B, but with a greater emphasis on difficult negatives. Our experiments use the gte-multilingual-base variant.

Our primary baseline to compare against is the unmodified base model (aka off-the-shelf model),

[1] https://github.com/tesseract-ocr/tesseract

		clean → OCR				OCR → OCR				**Averages**	
Model	**Approach**	$\text{WMT19}^{\varnothing\to\dots}_{\text{de}\to\text{fr}}$	$\text{WMT21}^{\varnothing\to\dots}_{\text{de}\to\text{fr}}$	$\text{WMT19}^{\varnothing\to\dots}_{\text{fr}\to\text{de}}$	$\text{WMT21}^{\varnothing\to\dots}_{\text{fr}\to\text{de}}$	$\text{WMT19}^{\dots\to\dots}_{\text{de}\to\text{fr}}$	$\text{WMT21}^{\dots\to\dots}_{\text{de}\to\text{fr}}$	$\text{WMT19}^{\dots\to\dots}_{\text{fr}\to\text{de}}$	$\text{WMT21}^{\dots\to\dots}_{\text{fr}\to\text{de}}$	clean → OCR	OCR → OCR
Blackletter / Scanned Distorted Noise Morgen Bonjour											
M-GTE	base	80.9	80.0	77.5	78.1	78.2	76.5	76.4	77.7	79.1	77.2
	clean X-News	80.6	79.6	80.6	77.8	77.7	75.5	75.9	76.4	79.6$_{+0.5}$	76.4$_{-0.8}$
	noised TED	85.3	84.3	82.0	81.9	82.3	81.2	80.3	80.7	83.4$_{+4.3}$	81.1$_{+3.9}$
	noised X-News	82.3	80.2	79.1	79.1	79.4	75.9	77.6	77.9	80.2$_{+1.1}$	77.7$_{+0.5}$
M-E5B	base	79.6	75.3	72.4	69.2	76.3	71.8	72.3	68.1	74.1	72.1
	clean X-News	78.3	77.3	75.4	74.3	75.0	74.2	74.7	73.0	76.3$_{+2.2}$	74.2$_{+2.1}$
	noised TED	81.0	78.7	78.7	76.3	77.1	76.3	77.5	75.8	78.7$_{+4.6}$	76.6$_{+4.5}$
	noised X-News	80.2	78.0	78.5	77.7	77.2	75.2	76.9	76.6	78.6$_{+4.5}$	76.4$_{+4.3}$
Salt and Pepper Noise Morgen Bonjour											
M-GTE	base	82.2	81.8	81.3	80.1	82.1	82.9	81.9	80.9	81.3	81.9
	clean X-News	81.8	82.1	80.2	80.9	80.8	82.1	81.8	80.6	81.2$_{-0.1}$	81.4$_{-0.5}$
	noised TED	85.0	84.5	82.9	81.4	85.4	85.0	84.5	83.1	83.5$_{+2.2}$	84.5$_{+2.6}$
	noised X-News	82.3	82.2	81.0	80.1	81.5	83.2	82.2	81.0	81.4$_{+0.1}$	82.0$_{+0.1}$
M-E5B	base	81.8	74.9	76.9	71.8	77.1	70.2	78.5	75.4	76.4	75.3
	clean X-News	77.9	78.4	79.6	77.2	77.9	78.6	79.6	77.4	78.3$_{+1.9}$	78.4$_{+3.1}$
	noised TED	80.3	79.1	82.1	78.0	81.1	78.9	82.5	79.9	79.9$_{+3.5}$	80.6$_{+5.3}$
	noised X-News	78.5	79.0	79.5	77.4	78.9	79.7	79.9	78.4	78.6$_{+2.2}$	79.2$_{+3.9}$

Table 1: Main German–French results. All values are averaged over five fine-tuning seeds. *Italicized* entries indicate results from our proposed adaptation approach.

which allows us to assess how well these embedding models perform on digitized text without any additional training—reflecting their typical use in real-world applications.

To isolate the effect of our noising strategy, we additionally fine-tune the model on the *X-News* corpus without any injected OCR noise. We refer to this condition as *clean X-News*.

All model training follows a consistent setup: we use the standard *MultipleNegativesRankingLoss* (Henderson et al., 2017), a batch size of 8, and train for one epoch using 20,000 positive pairs.

6 Experimental Results

We evaluate our approach in two scenarios: (1) clean → OCR, representing a typical user-query setting where clean input is used to search over noisy, digitized content; and (2) OCR → OCR, representing semantic search conducted entirely within OCR-processed corpora.

6.1 Main DE–FR Results

Our primary German–French results are presented in Table 1. Across both embedding models, our simple adaptation method yields substantial performance gains. Notably, these improvements persist even when training uses out-of-domain data. When

Control Tasks		X-Semantic Search (CLSD)				Similarity			**Averages**	
Model	**Approach**	$\text{WMT19}^{\varnothing\rightarrow\varnothing}_{\text{de}\rightarrow\text{fr}}$	$\text{WMT21}^{\varnothing\rightarrow\varnothing}_{\text{de}\rightarrow\text{fr}}$	$\text{WMT19}^{\varnothing\rightarrow\varnothing}_{\text{fr}\rightarrow\text{de}}$	$\text{WMT21}^{\varnothing\rightarrow\varnothing}_{\text{fr}\rightarrow\text{de}}$	$\text{STS}_{\text{en}\leftrightarrow\text{es}}$	$\text{STS}_{\text{en}\leftrightarrow\text{tr}}$	$\text{STS}_{\text{en}\leftrightarrow\text{ar}}$	CLSD	STS
M-GTE	base	90.2	90.5	89.7	91.6	83.9	76.4	76.4	90.5	78.9
	clean X-News	91.1	91.0	91.1	92.5	82.6	77.0	76.9	91.4$_{\mathbf{+0.9}}$	78.9$_{\mathbf{+0.0}}$
	noised TED	93.3	93.4	92.5	93.3	80.8	75.4	76.4	93.1$_{\mathbf{+2.6}}$	77.5$_{-1.4}$
	noised X-News	91.7	91.4	90.6	92.7	82.3	76.1	76.7	91.6$_{\mathbf{+1.1}}$	78.4$_{-0.5}$
M-E5B	base	91.5	86.3	88.5	82.0	76.6	63.3	71.3	87.1	70.4
	clean X-News	90.9	91.5	90.5	89.7	77.4	68.7	74.9	90.7$_{\mathbf{+3.6}}$	73.7$_{\mathbf{+3.3}}$
	noised TED	90.1	89.9	91.2	87.6	75.2	66.7	73.9	89.7$_{\mathbf{+2.6}}$	71.9$_{\mathbf{+1.5}}$
	noised X-News	91.3	92.0	91.0	90.0	76.3	69.1	74.1	91.1$_{\mathbf{+4.0}}$	73.2$_{\mathbf{+2.8}}$

Table 2: Results of control experiments on clean test data, i.e., without OCR noise. STS evaluation is based on Spearman correlation with human Likert-scale scores. All values are averaged over five fine-tuning seeds. *Italicized* entries indicate results from our proposed adaptation approach.

Approach	**lb↔fr**	**lb↔en**	**lb↔de**	**Average**
base	83.7	80.1	87.6	83.8
clean X-News	83.6	81.6	86.0	83.7$_{-0.1}$
noised TED	87.1	86.2	90.8	88.0$_{\mathbf{+4.2}}$
noised X-News	84.7	83.0	87.1	84.9$_{\mathbf{+1.1}}$

Table 3: Accuracy of our adapted models on the HISTLUX Bitext Mining evaluation set using M-GTE.

fine-tuning on the *noised TED* dataset and evaluating on Blackletter (BL) OCR text, we observe average improvements of +4.3 and +4.6 points (for M-E5B and M-GTE, respectively) when clean queries are matched against noisy candidates, and gains of +4.0 and +4.6 points when both query and candidate contain BL noise.

Similarly, for Salt and Pepper (SnP) noise, fine-tuning on *noised TED* results in average gains of +2.2 and +3.6 points when only the candidate text is noisy, and +2.6 and +5.3 points when noise is present in both query and candidate.

Fine-tuning on domain-aligned data (*noised X-News*) yields improvements in a comparable range, indicating the robustness and generalizability of our adaptation strategy.

Looking more closely at individual evaluations, we observe consistent improvements across all dataset variations and both directions of the language pair. The largest gain is observed in $\text{WMT19}^{\text{SnP}\rightarrow\text{SnP}}_{\text{de}\rightarrow\text{fr}}$, with an improvement of +8.7 points. Notably, some improvements also occur when fine-tuning on the *clean X-News* baseline, particularly for the less robust M-E5B model.

Finally, we perform statistical significance testing. Our null hypothesis is that the proposed random noise strategy yields correct predictions with equal likelihood as any of the baseline strategies (i.e., using the base model or training on clean data). Using Fisher's exact test, we reject the null hypothesis with high confidence ($p \ll 0.001$). This result holds consistently across both embedding models (M-E5B and M-GTE).

6.2 Performance on Historic Luxembourgish

In this experiment, we evaluate our models on HISTLUX, a dataset consisting of historical Luxembourgish text that has been OCR-processed by real digitization facilities. Notably, Luxembourgish is not included in any of our adaptation training data, making this a challenging generalization task for our method.[2]

The results are presented in Table 3. We observe consistent improvements with both of our noise-based adaptation strategies, particularly when fine-tuning on *noised TED*, which yields an average improvement of +4.2 points. In contrast, fine-

[2] Since the M-E5B model does not support Luxembourgish, we restrict this evaluation to the M-GTE model, which was exposed to 48,000 Luxembourgish sentence pairs during its pre-release contrastive pre-training (c.f. Table 10 in Zhang et al., 2024).

		clean → OCR				OCR → OCR				**Averages**	
Model	**Approach**	$\text{WMT19}^{\varnothing\to\dots}_{\text{de}\to\text{fr}}$	$\text{WMT21}^{\varnothing\to\dots}_{\text{de}\to\text{fr}}$	$\text{WMT19}^{\varnothing\to\dots}_{\text{fr}\to\text{de}}$	$\text{WMT21}^{\varnothing\to\dots}_{\text{fr}\to\text{de}}$	$\text{WMT19}^{\dots\to\dots}_{\text{de}\to\text{fr}}$	$\text{WMT21}^{\dots\to\dots}_{\text{de}\to\text{fr}}$	$\text{WMT19}^{\dots\to\dots}_{\text{fr}\to\text{de}}$	$\text{WMT21}^{\dots\to\dots}_{\text{fr}\to\text{de}}$	clean → OCR	OCR → OCR
Minimal Noise Morgen Bonjour											
M-GTE	base	87.3	88.7	89.5	90.6	87.2	88.7	88.8	90.4	89.0	88.8
	clean X-News	88.4	89.3	89.8	91.2	88.2	88.9	89.6	90.7	89.7$_{+0.7}$	89.3$_{+0.5}$
	noised TED	91.2	92.4	91.9	92.7	91.1	92.5	91.4	92.5	92.0$_{+3.0}$	91.9$_{+3.1}$
	noised X-News	89.3	89.9	90.4	91.8	89.2	89.6	90.0	91.4	90.3$_{+1.3}$	90.0$_{+1.2}$
M-E5B	base	88.9	84.9	86.7	80.4	88.7	84.5	86.2	80.7	85.2	85.0
	clean X-News	87.8	90.4	89.9	88.0	87.7	90.1	89.4	88.1	89.0$_{+3.8}$	88.8$_{+3.8}$
	noised TED	88.3	88.4	90.6	86.9	88.1	88.4	90.3	86.5	88.5$_{+3.3}$	88.3$_{+3.3}$
	noised X-News	89.0	90.8	90.5	89.1	88.9	90.9	90.0	89.0	89.8$_{+4.6}$	89.7$_{+4.7}$

Table 4: Results under minimal noise conditions. All values are averaged over five fine-tuning seeds. *Italicized* entries indicate results from our proposed adaptation approach.

tuning on clean, domain-aligned data (*clean X-News*) does not lead to any measurable gains. The most substantial improvement is observed for the Luxembourgish–English direction, with a gain of +6.1 points following adaptation on *noised TED*.

6.3 Control Experiments

Having observed substantial improvements on OCR-noisy texts, a key question arises: does this adaptation come at the cost of performance on standard, clean data? To evaluate this, we conduct control experiments on *contemporary, clean, cross-lingual, and out-of-language* semantic textual similarity tasks. Specifically, we use the English–Turkish, English–Spanish, and English–Arabic test sets from the multilingual STS benchmark: $\text{STS}_{\text{en}\leftrightarrow\text{tr}}$, $\text{STS}_{\text{en}\leftrightarrow\text{es}}$, and $\text{STS}_{\text{en}\leftrightarrow\text{ar}}$.

The results of these experiments are shown in Table 2. On the clean CLSD evaluation set, we observe no performance degradation. In fact, fine-tuning on *noised X-News* yields average gains of up to +4.0 points. For the STS benchmarks in previously unseen languages, results vary depending on the model: M-E5B shows modest improvements (up to +3.3 points), while M-GTE experiences small declines (up to –1.4 points).

Overall, we find that adaptation to OCR noise does not negatively impact performance on clean, cross-lingual tasks and remains stable across unrelated language pairs.

6.4 Minimal Noise Conditions

To assess whether our approach remains beneficial when applied to high-quality OCR text with minimal degradation, we evaluate the Minimal Noise (MN) variant of the CLSD dataset. This variant, which simulates modern digitized text with clean layout and high-resolution input, has a Character Error Rate (CER) below 0.6%.

We evaluate both the baseline and adapted models on this dataset, as shown in Table 4. As expected, the base models suffer significantly less performance degradation under MN conditions compared to the more challenging Blackletter (BL) and Salt-and-Pepper (SnP) scenarios. Nonetheless, our adaptation strategies continue to yield gains. Specifically, the M-GTE model improves by up to 3.0 points when fine-tuned on *noised TED*, and the M-E5B model improves by up to 4.7 points when trained on *noised X-News*.

These results indicate that even in cleaner OCR scenarios, our models benefit from the proposed adaptation method and maintain strong performance in modern digitization settings.

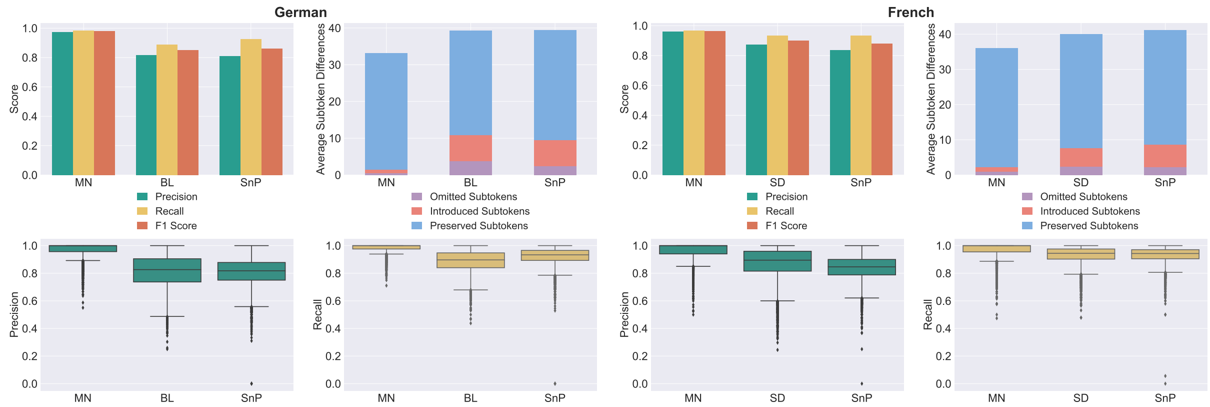


Figure 3: Precision, Recall and F1 scores of the sets of subtokens in the noisy text against their clean equivalent.

7 Discussion

Generalization issues in off-the-shelf models: Is it the tokenizer's fault? Our adaptation strategies reduce the performance drop that standard multilingual embedding models exhibit when processing OCR-noisy text. A central question, however, is what causes these models to generalize poorly to such noisy input.

We hypothesize that the tokenizer plays a crucial role in this degradation. Modern embedding models rely on subword tokenization, and OCR-induced character perturbations often result in token boundary mismatches. These mismatches propagate through the network layers, distorting semantic representations.

To investigate this hypothesis, we examine the XLM-RoBERTa tokenizer (Conneau et al., 2020) used by both M-E5B and M-GTE. This tokenizer is built on the SentencePiece unigram algorithm (Kudo, 2018; Kudo and Richardson, 2018). We start by analyzing words and short phrases from original noisy text pairs in the CLSD evaluation set (see Appendix Table 6 for examples). Our qualitative analysis reveals that even minor recognition errors can significantly disrupt the tokenization process. For instance, misreading the character "ö" as "ô" in the German word "können" (English: "can") causes what is normally a single meaningful subtoken to be split into three smaller, less interpretable fragments. Similarly, inserting an extra "i" in the French word "comme" (English: "like", used as a conjunction) results in the unintended subtokens "com" and "mie".

To quantify the effect of the subtokenization changes, we compare the tokenizations of clean and OCRed text by measuring their vocabulary overlap. Specifically, we compute precision, recall and F1 scores based on the sets of produced subtokens. The results are presented in Figure 3. We observe that the more noisy the conditions, the less of the original subtokens are preserved by the subtokenizer, a finding that seems to align well with the observed performance degradation of the baseline models in our main experiments. It is also interesting that while texts with BL noise have about half the character error rate of SnP, they preserve about the same amount of subtokens from the original texts. This suggests that misrecognitions, which occur commonly in difficult-to-parse fonts such as BL, alter the subtokenization more than insertion errors, which are more prevalent under SnP conditions.

These findings suggest that our adaptation strategy may help correct some of the harmful effects by misaligned tokenization, thereby reducing the propagation of errors through the model. A more detailed investigation of this mechanism is beyond the scope of the present study and is left for future work.

Method	**BL/SD**	**SnP**	**HistLUX**
Base	78.1	81.6	83.8
Random Noise	82.2	84.0	**87.9**
BL/SD Noise	**83.2**	86.0	84.3
SnP Noise	82.8	**86.1**	82.8

Table 5: Performance comparison of models trained with different noise types across three evaluation settings: CLSD with Blackletter/Scanned (BL/SD) noise, CLSD with Salt-and-Pepper (SnP) noise, and the historical HISTLUX benchmark. While training on realistic noise (BL/SD or SnP) improves performance on the corresponding noise type, models trained with random noise show more consistent generalization, including on naturally noisy historical data.

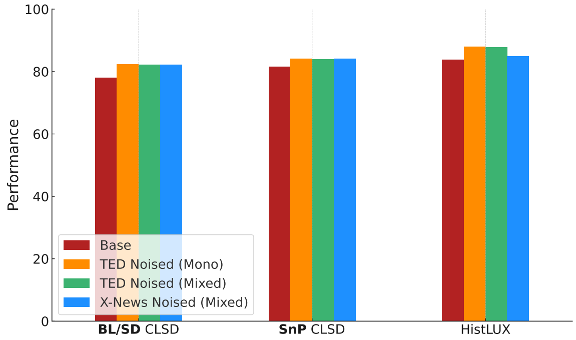


Figure 4: Ablation study comparing the effects of batch composition and training domain on model performance. "TED Noised (Mono)" refers to monolingual batching (one language per batch), while "TED Noised (Mixed)" uses mixed-language batches. "X-News Noised (Mixed)" applies the same mixed batching strategy on a different domain. Results are shown for three evaluation settings: CLSD with Blackletter/Scanned (BL/SD) noise, CLSD with Salt-and-Pepper (SnP) noise, and the historical HistLUX benchmark. In this experiment, X-News is used as a monolingual training set.

Training regime: Which strategy performs best? We examine how the batch mixing strategy affects performance in our best adaptation setup, *noised TED*. To do so, we replicate the results with monolingual batches[3], allowing us to isolate the effect of language mixing during training. In addition, to assess the impact of domain similarity, we conduct parallel experiment using the *noised X-News* corpus, also with monolingual batching, while keeping the training procedure identical to *noised TED*.

We show the results of this experiment in Figure 4. We observe negligible differences between mixed and monolingual batching strategies. However, training on the X-News dataset results in slightly lower performance than TED, suggesting that while domain may have some influence, the choice of training corpus plays only a minor role in the overall effectiveness of the noise adaptation.

Noise type: Does training on realistic OCR noise help? We investigate whether training models on more realistic OCR noise improves performance and generalization across different noise types. To this end, we train two variants of our models using the noised TED dataset, with noise generated via our realistic Blackletter (B)L and Salt-n-Pepper (SnP) pipelines.

As shown in Table 5, training on realistic noise improves robustness within the respective noise condition to a similar extent as training on random noise. However, it yields smaller gains on other types of noise. Notably, performance drops most clearly on the HistLUX benchmark, which contains naturally occurring OCR noise. In this setting, both realistic-noise models perform worse than the base model (by -0.4 and -2.0 points) and substantially worse than the model trained on random noise (by -3.8 and -5.4 points).

These results indicate that while training on (more costly) realistic noise can enhance robustness to that specific noise type, it reduces generalization to other types of noise and real-world OCR artifacts. In contrast, our random noise approach offers broader transferability at a lower cost.

8 Conclusions

We propose an inexpensive and effective strategy for adapting multilingual embedding models to be more robust to heterogeneous digitized text. Our adaptation approach improves performance across different noise conditions and even over historical digitized text in related languages. Through control experiments, we show that our method has minimal impact on the overall embedding quality. We find that OCR noise increases token fragmentation, which may explain the observed performance degradation. We believe that our work is an important step towards building and evaluating reliable semantic search systems for large and diverse digitized text collections.

Acknowledgments

We thank the anonymous reviewers for their suggestions. The authors received funding through the project *Impresso – Media Monitoring of the Past II Beyond Borders: Connecting Historical Newspapers and Radio*. Impresso is a research project funded by the Swiss National Science Foundation (SNSF 213585) and the Luxembourg National Research Fund (17498891).

Limitations

Our study demonstrates that multilingual embedding models can be made more robust to noise commonly found in digitized texts. However, such noise may not only consist of actual errors (e.g., from OCR systems), but it can also include other phenomena, such as typos in newspapers, regional

[3] "Mixed" refers to batches with samples from multiple languages; "Mono" to batches with samples from a single language.

or historical spelling variants. It is possible that the methodology proposed here improves the representation of such texts, but we are unable to assess this in our study. Next, we plan to scale our adaptation procedure to a massively multilingual version, with careful ablations about the effect of scaling the data, increasing the training noise, and mixing different noise adaptations. This could lead to a generalized multilingual model that is able to robustly represent highly heterogeneous texts across languages.

References

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. 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.

K. D. Dhingra, S. Sanyal, and P. K. Sharma. 2008. A robust ocr for degraded documents. In X. Huang, YS. Chen, and SI. Ao, editors, *Advances in Communication Systems and Electrical Engineering*, volume 4 of *Lecture Notes in Electrical Engineering*. Springer, Boston, MA.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *Preprint*, arXiv:1705.00652.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Daniel Lopresti. 2008. Optical character recognition errors and their effects on natural language processing. In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, AND '08, page 9–16, New York, NY, USA. Association for Computing Machinery.

Andrianos Michail, Simon Clematide, and Rico Sennrich. 2025a. Examining multilingual embedding models cross-lingually through llm-generated adversarial examples. *Preprint*, arXiv:2502.08638.

Andrianos Michail, Corina Julia Raclé, Juri Opitz, and Simon Clematide. 2025b. Adapting multilingual embedding models to historical luxembourgish. *Preprint*, arXiv:2502.07938.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Georgios Sidiropoulos and Evangelos Kanoulas. 2022. Analysing the robustness of dual encoders for dense retrieval against misspellings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2132–2136, New York, NY, USA. Association for Computing Machinery.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 90–96, Tokyo, Japan. International Workshop on Spoken Language Translation.

Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2023. Typo-robust representation learning for dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1106–1115, Toronto, Canada. Association for Computational Linguistics.

Konstantin Todorov and Giovanni Colavizza. 2022. An assessment of the impact of ocr noise on language models. In *International Conference on Agents and Artificial Intelligence*.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. 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 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Alexander van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream nlp tasks. In *International Conference on Agents and Artificial Intelligence*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

A Appendix

Noise Type	Original and Noisy Text	Subtokenization Changes
German Text from CLSD WMT19/21		

Simple Noise	**Original:** Staatssozialismus
	Noisy: Staatssozialiısmus
		Original Tokens:
		`[Staats][sozial][ismus]`
		Noisy Sub-Tokens:
		`[Staats][soziali][ı][s][mus]`

	Original: mit **Noisy:** mıt	**Original Sub-Tokens:** `[mit]` **Noisy Sub-Tokens:** `[mı][t]`

Blackletter	**Original:** historischen
	Noisy: bistorisehen
		OriginalSub-Tokens:
		`[historische][n]`
		Noisy Sub-Tokens:
		`[bis][tori][sehen]`

	Original: Europa **Noisy:** Curopa	**Original Sub-Tokens:** `[Europa]` **Noisy Sub-Tokens:** `[Cu][ropa]`
Salt and Pepper	**Original:** können **Noisy:** kônnen ; ;	**Original Sub-Tokens:** `[können]` **Noisy Sub-Tokens:** `[k][ôn][nen][;][;]`
	Original: andere **Noisy:** .ander.e	**Original Sub-Tokens:** `[andere]` **Noisy Sub-Tokens:** `[.][ander][.][e]`
French Text from CLSD WMT19/21		
Simple Noise	**Original:** présidente **Noisy:** preresidente	**Original Sub-Tokens:** `[président][e]` **Noisy Sub-Tokens:** `[presidente]`
	Original: leçon **Noisy:** lecon	**Original Sub-Tokens:** `[le][çon]` **Noisy Sub-Tokens:** `[le][con]`

Scanned Noise	**Original:** telle qu'elle
	Noisy: 'teHÊ': qu'elle
		Original Sub-Tokens:
		`[telle][qu]['][elle]`
		Noisy Sub-Tokens:
		`['][te][H][Ê]['][:][qu]['][elle]`

	Original: comme **Noisy:** commie	**Original Sub-Tokens:** `[comme]` **Noisy Sub-Tokens:** `[com][mie]`
Salt and Pepper	**Original:** autre Europe **Noisy:** autré ... Europe	**Original Sub-Tokens:** `[autre][Europe]` **Noisy Sub-Tokens:** `[au][tré][...][Europe]`
	Original: de gauche **Noisy:** de.gauche	**Original Sub-Tokens:** `[de][gauche]` **Noisy Sub-Tokens:** `[de][.][gau][che]`

Table 6: Examples of subtokenization changes in different noise conditions