XNLIeu: a dataset for cross-lingual NLI in Basque

Anonymous ACL submission

Abstract

The XNLI dataset, a benchmark for Natural Language Inference (NLI), is extensively used to assess cross-lingual Natural Language Understanding (NLU) capabilities across various languages. In this paper, we extend XNLI to include Basque, a low-resource language that can greatly benefit from transfer-learning approaches. XNLIeu has been developed by first machine-translating the English XNLI corpus to Basque, followed by a manual post-edition step. We conduct a series of experiments using mono- and multilingual LLMs to assess a) the effect of professional post-edition compared to the automatic MT system b) the best cross-lingual strategy for NLI in Basque and c) whether the choice of the best cross-lingual strategy is influenced by the fact that the dataset is built by translation. The results show that post-edition is crucial and that the translate-train cross-lingual strategy obtains better results overall, although the gain is lower when tested in a dataset that has been built natively from scratch.

1 Introduction

Natural Language Inference (NLI) is a task that consists of classifying pairs of sentences, a premise and a hypothesis, according to their semantic relation: *entailment*, when the meaning of the premise entails that of the hypothesis; *contradiction*, when the meanings of both sentences can not co-occur at the same time, they have opposing truth conditions; and *neutral*, when both sentences are not semantically related whatsoever and could either occur or not at the same time (see Table 1 for examples).

NLI is an important task towards Natural Language Understanding (NLU), and is often used to test the semantic understanding of language models. It provides a general framework where different NLP tasks can be reframed, including information retrieval (Dušek et al., 2023), metaphor detection (Stowe et al., 2022) or relation extraction (Sainz et al., 2021). The NLI paradigm has also been proposed as a way to detect hallucination in Natural Language Generation (NLG) (Ji et al., 2023).

premise	Yesterday I saw an octopus at the beach.
entailment	I was at the beach yesterday.
contradiction	Yesterday I spent the whole day at home.
neutral	Octopi are my favourite animals.

Table 1: Example of a premise and three different hypotheses with the three possible relations.

XNLI (Conneau et al., 2018) is a popular benchmark to evaluate cross-lingual NLI capabilities among languages, and is made of human translations of the development and test sets of the MNLI dataset (Williams et al., 2018) in 14 languages and English. In this paper, we extend XNLI to include Basque, a low-resource language spoken in Spain and France (ISO-code: *eu*). We present XNLIeu, a dataset that has been built by machine translating and post-editing the English version of XNLI. Apart from XNLIeu, we also release the machine-translated version, which we use to assess whether professional post-edition is needed to obtain a reliable dataset, even if is has been more costly to obtain.

Previous work has emphasized the importance of the origin of the train and test data in cross-lingual settings (whether they are original or come from translations). In particular, Artetxe et al. (2020) shows that a mismatch on the origin can have a serious impact on the results, particularly when comparing different cross-lingual strategies. Moreover, NLI datasets are known to be biased and contain artifacts that lead models to rely on superficial clues (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; McCoy et al., 2019). To analyze the impact of these factors in XNLIeu, we create a native test set completely from scratch with original premises extracted from real Basque sources and hypotheses provided by Basque speakers, which were specifically told to avoid such biases.

Label	Example
	premise: *Dena idazten saiatu nintzen* 'I tried to write everything.'
entailment	*Nire helburua gauzak idaztea zen.* 'My goal was to write things'
contradiction	*Ez nintzen ezer idazten saiatu ere egin.* 'I didn't even try to write anything.'
neutral	*Aipatu zuen lan bakoitza idatzi nuen.* 'I wrote every paper he mentioned.'

Table 2: Examples from the XNLIeu dataset

	MNLI	XNLIeu	XNLIeu$_{\text{MT}}$	native
entailment	9.89	8.15	7.81	8.95
contradiction	10.39	8.73	8.39	9.94
neutral	11.4	9.31	8.98	9.41

Table 3: Average length of hypotheses for each semantic relation type in our three datasets, as well as MNLI. the original English instances of XNLIeu and XNLIeu$_{\text{MT}}$.

Using these datasets, we conduct a series of experiments using mono- and multilingual language models for Basque, both discriminative and generative, and test different training variants for cross-lingual NLI in Basque. The experiments set a new baseline on NLI in Basque, and serves us to analyze the effect of professional post-edition compared to the automatic machine-translation system as well as he best cross-lingual strategy for NLI in Basque, both in translated and native sets.

This paper is structured as follows: section 2 covers some relevant research and resources related to the topic in hand, our dataset is further explained in section 3, the description of the experiments and experimental settings in section 4, the results are covered in section 5, section 6 includes the analysis of the errors of the outputs of our models, section 7 a summary of the research and its conclusions and, finally, section 8 expands on the limitations of our research.

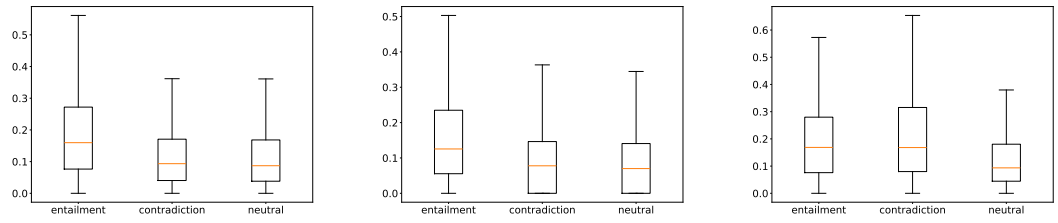
2 Related work

Cross-lingual NLI. The best results on NLI benchmarks to day are based on supervised learning, which requires large amounts of training data that are only available for resource-rich languages such as English. Examples of English NLI datasets are the Stanford NLI corpus (Bowman et al., 2015), the Multi-genre NLI corpus (Williams et al., 2018) or the Adversarial NLI corpus (Nie et al., 2020). The NLI task is also included among the tasks of the popular NLU benchmarks GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2020). Cross-lingual NLI is an alternative approach that leverages pre-trained multilingual models which are fine-tuned in resource-rich languages, and tested in the desired target language. This transfer approach, called *zero-shot*, is often compared to strategies that involve machine translation: *translate-train*, where the training set is translated to each target language and used to train the models on their respective language and *translate-test*, where the test set is translated to the source language, usually English. Alternatively, large multilingual autoregressive models are also known to perform well in cross-lingual settings, by providing them with a set of correct input/label pairs as prompts for new inputs (Brown et al., 2020).

XNLI. The Cross-lingual NLI corpus (XNLI) (Conneau et al., 2018) is a set of development and test sets in 15 high- and low-resource languages based on MNLI, meant as a cross-lingual benchmark for this task. Later, this corpus was expanded to include additional languages such as Korean (Ham et al., 2020).

NLI biases & artifacts. Most famous NLI datasets have also been reported to include biases and artifacts (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; McCoy et al., 2019) that should be considered when analyzing the results, as they seem to have critical effects on the performance of systems. Artetxe et al. (2020) analyzes the effect that translated datasets have in cross-lingual settings, due to the so-called *translationese*(Volansky et al., 2013), and concludes that mismatches between the origin of training and evaluation datasets cause an important impact on the performance of the models.

Evaluation of LLMs. Nowadays, the focus of the research on evaluation has shifted due to the outstanding growth of LLMs, which can achieve comparable results to fine-tuned pre-trained models with zero-shot and few-shot approaches for evaluation, and the center of attention is set on the global capabilities of the models rather than specific tasks (Guo et al., 2023). However, low-resource languages like Basque lag behind in NLP development, and can still benefit considerably from semantic datasets for tasks like NLI, that was not previously available for this language.



XNLIeu XNLIeu$_{\mathrm{MT}}$ Native

Figure 1: Box plots of the lexical overlap between premises and hypotheses calculated with cosine similarity of the three datasets.

3 The XNLIeu dataset

The XNLIeu dataset has been created following the steps of XNLI (Conneau et al., 2018), by machine-translating[1] the development and test sets of the MNLI corpus (Williams et al., 2018) and post-editing the machine-translated instances with assistance of professional translators. Table 2 shows examples of the XNLIeu dataset. Apart from XNLIeu, we also release the machine-translated version prior post-edition, dubbed XNLIeu$_{\mathrm{MT}}$, which we use to analyze the effect of post-edition.

Additionally, we have created an original test set from scratch for the Basque language, which will be referred to as *native*, primarily to compare the results with those of our other datasets, with detailed results discussed in section 5.2. Inspired by Bowman et al. (2015) and Artetxe et al. (2020), we performed the following steps:

- As a starting point, we extracted $5,000$ sentences from recent corpora containing local news in Basque that are not included in those used to train the models that will be fine-tuned in the experiments.
- From these initial sentences, we manually selected 207 sentences that we deemed appropriate for this task, which are the premises of our dataset.
- We redacted annotation guidelines that explain the task and provide examples to the annotators. In these guidelines, annotators are asked to be creative and to avoid as much as possible some of the annotation artifacts that have been found in the large datasets (Gururangan et al., 2018; Poliak et al., 2018), such as the use of negation to create contradictions.
- With the assistance of native Basque speakers, one hypothesis for each label was created per premise, resulting in three hypotheses per premise, with a total of 621 sentences.
- The dataset has undergone superficial corrections, focusing mainly on orthography and revising examples that featured ambiguous labels.

For the translation-train experiments, we also distribute a machine-translated version of the English MNLI training corpus to Basque, with a total of $392,702$ sentences.

3.1 Quantitative analysis

In this section we present a quantitative analysis of various aspects of the three developed datasets: the XNLIeu dataset, XNLIeu$_{\mathrm{MT}}$ and the native version.

Label distribution. Since there are three hypotheses for each premise in the dataset, the label distribution is perfectly balanced, resulting in no majority class and setting the baseline accuracy at 33%. This applies to the three datasets.

Sentence length. The average length for hypotheses for each semantic relation type, as shown in Table 3, indicates that there is a bias, as neutral hypotheses are longer on average, while entailed hypotheses tend to be shorter, likely because entailed sentences are often formed by omitting words from the premise (Gururangan et al., 2018). This bias is present in the MNLI dataset and in XNLIeu and XNLIeu$_{\mathrm{MT}}$, that have been translated from it. The hypotheses of the Basque datasets tend to be shorter than the original English ones, but the unbalance between the different semantic relation types is maintained. The native set is also skewed, but in this case, the contradictions are slightly longer

[1] All translations performed in the paper have been obtained using Elia at https://elia.eus/translator.

	XNLIeu		**XNLIeu$_{MT}$**		**native**	
entailment	no	0.58%	no	0.54%	in Basque	0.41%
	auxiliary[2]	0.24%	*auxiliary*	0.23%	film	0.24%
	something	0.19%	some	0.18%	*auxiliary*	0.24%
	some	0.18%	something	0.16%	movie	0.24%
	auxiliary	0.17%	like	0.13%	of the world	0.24%
contradiction	no	1.61%	no	1.65%	no	0.45%
	nobody	0.24%	nobody	0.23%	in Basque	0.34%
	never	0.2%	*auxiliary*	0.18%	Basque	0.28%
	auxiliary	0.18%	never	0.16%	my	0.23%
	my	0.16%	importance	0.14%	from Bilbao	0.23%
neutral	no	0.33%	no	0.31%	like	0.37%
	my	0.21%	dollar	0.2%	no	0.37%
	auxiliary	0.19%	my	0.2%	Basque	0.25%
	some	0.18%	*auxiliary*	0.16%	sometimes	0.25%
	like	0.15%	some	0.16%	people	0.25%

Table 4: Proportion of most frequent words of the three datasets, translated from Basque to English.

Discriminative		
Name	**Language**	**# of parameters**
IXAmBERT	Multi	177M
multilingual BERT	Multi	179M
XLM-RoBERTa (base)	Multi	279M
XLM-RoBERTa (large)	Multi	561M
BERTeus	Basque	124M
RoBERTa-eus Euscrawl	Basque	355M
Generative		
Llama-Eus (private)	Multi	7B
BLOOM	Multi	7.1B
XGLM	Multi	7.5B

Table 5: Details of the models that have been used in the experiments.

than neutral hypotheses, and entailments are still shorter on average.

Word frequency. Examining word frequency per label is insightful, especially since studies such as Gururangan et al. (2018) or Tsuchiya (2018) have reported that some NLI datasets exhibit a bias where the contradiction label is strongly associated with negation words. This seems to hold for the XNLIeu and XNLIeu$_{MT}$ datasets. As we can see in Table 4, the word *ez* 'no'appears much more frequently in contradictions, and so do some other negations like *inork* 'nobody'or *inoiz* 'never', it is plausible that models might be exploiting this feature as a form of shortcut learning for classification without even looking at the premise. The native dataset does not seem to be biased towards negation words, since the guidelines specifically asked the annotators to avoid using artifacts as much as possible. It is interesting to note that among the most frequent words of this dataset there are frequent references to the Basque culture: *euskaraz* 'in Basque', *euskara* 'Basque language', *euskal* 'Basque(adjective)'or *Bilboko* 'from Bilbao'.

Lexical overlap. The lexical overlap between the premise and hypothesis has been calculated as the cosine similarity between TF-IDF vector representations of both sentences. The results (Figure 1) show that on the XNLIeu and XNLIeu$_{MT}$ datasets, premises and hypothesis are more similar on average when their relationship is labeled as entailment than for the two other relation types, which has been once again previously detected as a typical NLI bias that can be harmful to the task, probably caused by how entailed hypotheses are easy to create by simply omitting parts of the premise. However, this bias is not reproduced on the native test set.

[2]Auxiliaries are further discussed in appendix A.

4 Experimental design

We conduct a series of experiments towards evaluating cross-lingual NLI for Basque, using different discriminative and generative language models, both mono- and multi-lingual. All models have been tested using the three datasets we have created and described in section 3, aiming to determine if post-edition introduces significant changes to the dataset that enhance its accuracy and reliability. We also want to compare the XNLI derived datasets to the native human-devised dataset and analyze the impact caused by biases and artifacts introduced in translation. Since there is no training set in Basque for NLI, we consider different cross-lingual alternatives[3]:

- *Zero-Shot transfer*: We use discriminative multilingual models that have been pretrained at least in English and Basque. These models are then fine-tuned using the English MNLI corpus. In a further experiment, we also consider source languages other than English for fine-tuning.

- *Translate-train*: We machine-translate the English MNLI dataset to Basque, and use it to fine-tune the discriminative models (both multilingual and Basque monolingual).

- *Zero-shot prompting*: Generative models (which include English and Basque at pretraining) have been directly tested without

[3]The translate-test approach has not been implemented since the datasets have been originally translated from English to Basque, so back-translating them to English would not allow us to draw meaningful conclusions.

zero-shot		
	XNLIeu	**XNLIeu$_{\text{MT}}$**
IXAmBERT	72.5 ($\pm 1.4e^{-3}$)	67.3 ($\pm 7.0e^{-3}$)
mBERT	60.1 ($\pm 5.7e^{-3}$)	57.9 ($\pm 1.2e^{-2}$)
XLM-RoBERTa base	73.4 ($\pm 3.5e^{-3}$)	69.0 ($\pm 9.0e^{-3}$)
XLM-RoBERTa large	**81.1** ($\pm 2.8e^{-3}$)	**75.4** ($\pm 2.0e^{-3}$)
translate-train		
	XNLIeu	**XNLIeu$_{\text{MT}}$**
IXAmBERT	75.9 ($\pm 6.4e^{-3}$)	71.3 ($\pm 4e^{-3}$)
mBERT	74.8 ($\pm 4.2e^{-3}$)	71.3 (± 0.0)
XLM-RoBERTa large	**83.8** ($\pm 6.0e^{-4}$)	**79.9** ($\pm 1.0e^{-3}$)
RoBERTa-euscrawl	83.0 ($\pm 7.1e^{-3}$)	78.6 ($\pm 2.0e^{-3}$)
BERTeus	79.0 ($\pm 4.2e^{-3}$)	74.9 ($\pm 8.0e^{-3}$)

Table 6: Accuracy of discriminative fine-tuned models tested with XNLIeu and XNLIeu$_{\text{MT}}$ datasets (mean and standard deviation of three runs).

fine-tuning. This has been done using *prompting*, which consists of framing any task as Natural Language Generation by creating a prompt that includes the input for the model.

Regarding the models, we have experimented with the following discriminative models: IXAmBERT (Otegi et al., 2020), multilingual BERT (Devlin et al., 2018), XLM-RoBERTa large (Conneau et al., 2019), BERTeus (Agerri et al., 2020) and RoBERTa-eus-large (Artetxe et al., 2022). Further details about these models can be found in Table 5. All of the models have been used in their cased version. For the BERT models we have used a learning rate of 5e^{-5}, and for the RoBERTa models we have used a smaller learning rate of 10e^{-6}. All models have been trained for 10 epochs, and the model selection has been performed on the development test. There has been no further attempt at hyperparameter optimization, since the goal was not to obtain the best possible model, but rather to compare the effects of the different sets and strategies. The models have been trained with three different random seeds to get the mean and the standard deviation and reduce the effects of randomness. The code used for the experiments with discriminative models has been adapted from the code examples for fine-tuning for different tasks provided by Wolf et al. (2020).

We have also tested three multilingual generative models that include Basque among their training languages: Bloom (BigScience Workshop et al., 2023), XGLM (Lin et al., 2022) and Llama-Eus, an in-house model based on Llama 2 (Touvron et al., 2023)[4]. The prompts used in our experiments can be seen in Appendix B. As for evaluation, we select the label whose log-likelihood is maximum, according to the model. The code used for testing the generative models is based on that included in the Language Model Evaluation Harness project (Gao et al., 2021).

[4]The model was built by continual pre-training of the English Llama 2 using the publicly available EusCrawl corpus (Artetxe et al., 2022) for 4 epochs.

	XNLIeu	**XNLIeu$_{\text{MT}}$**
llama-eus	**49.5**	46.9
Bloom	**49.5**	**47.4**
XGLM	48.1	46.7

Table 7: Accuracy of generative models tested with XNLIeu and XNLIeu$_{\text{MT}}$ datasets using a zero-shot prompting approach.

Following usual practice, we use accuracy as our evaluation metric, the ratio of instances that are correctly classified divided by the total number of instances.

5 Results

In this section, we show the main results of our experiments and provide discussion on the main findings. We start by analyzing the results on the datasets derived from XNLI (XNLIeu and XNLIeu$_{\text{MT}}$), followed by a comparison with those obtained using the native dataset. Finally, we describe the results of experiments using different source languages apart from English in fine-tuning.

5.1 Results for XNLIeu and XNLIeu$_{\text{MT}}$

The main results for the discriminative models can be seen in Table 6. All systems perform consistently better when evaluated on the post-edited XNLIeu compared to the machine-translated XNLIeu$_{\text{MT}}$, and in some cases, the relative ranking among the models change, as it is the case between multilingual BERT and IXAmBERT in the translation-test setting. Translate-train obtains better results overall on all models, and the difference is slightly higher in the XNLIeu$_{\text{MT}}$ dataset (7.3% accuracy points on average), where both training and test data have been created only through machine-translation. This result is consistent with the findings reported in (Artetxe et al., 2020). In general, it seems that the more the target language is present at training and fine-tuning time, the better the results are. For example, multilingual BERT does not perform well in a zero-shot setting, but

zero-shot transfer	
IXAmBERT	64.0 ($\pm 9.0e^{-3}$)
mBERT	52.4 ($\pm 1.6e^{-2}$)
XLM-RoBERTa base	65.3 ($\pm 7.0e^{-3}$)
XLM-RoBERTa large	**73.8** ($\pm 7.0e^{-3}$)
translate-train	
BERTeus	68.4 ($\pm 1.0e^{-2}$)
IXAmBERT	65.6 ($\pm 1.0e^{-2}$)
mBERT	62.8 ($\pm 9.0e^{-3}$)
RoBERTa-euscrawl	75.2 ($\pm 7.0e^{-3}$)
XLM-RoBERTa large	**76.4** ($\pm 1.3e^{-2}$)
zero-shot prompting	
Llama-Eus	47.2
BLOOM	**49.8**
XGLM	46.5

Table 8: Accuracy of discriminative (upper part) and generative (bottom part) models tested on the native test set.

	XNLIeu	**XNLIeu$_{\mathrm{MT}}$**	**native**
en	73.4 ($\pm 3.5e^{-3}$)	69.0 ($\pm 9.0e^{-3}$)	**65.3** ($\pm 7.0e^{-3}$)
ar	**73.9**($\pm 2.6e^{-3}$)	**71.2**($\pm 4.0e^{-3}$)	61.9($\pm 3.0e^{-3}$)
bg	73.2($\pm 8.9e^{-3}$)	<u>71.0</u>($\pm 2.1e^{-3}$)	62.7($\pm 9.0e^{-3}$)
de	<u>73.9</u>($\pm 5.3e^{-3}$)	70.4($\pm 7.0e^{-4}$)	63.5($\pm 8.0e^{-3}$)
el	73.7($\pm 1.7e^{-3}$)	70.7($\pm 7.0e^{-4}$)	63.6($\pm 7.0e^{-3}$)
es	73.7($\pm 5.2e^{-3}$)	70.3($\pm 7.0e^{-4}$)	<u>65.0</u>($\pm 7.0e^{-3}$)
fr	73.7($\pm 4.9e^{-3}$)	69.9($\pm 7.1e^{-3}$)	63.3($\pm 2.1e^{-2}$)
hi	73.3($\pm 7.0e^{-3}$)	70.7($\pm 4.2e^{-3}$)	62.3($\pm 5.0e^{-3}$)
ru	72.9($\pm 1.5e^{-3}$)	69.7($\pm 2.1e^{-3}$)	62.2($\pm 6.0e^{-3}$)
sw	71.8($\pm 3.1e^{-3}$)	68.3($\pm 7.1e^{-3}$)	63.1($\pm 6.0e^{-3}$)
th	73.0($\pm 6.7e^{-3}$)	70.2($\pm 4.2e^{-3}$)	64.1($\pm 6.0e^{-3}$)
tr	73.5($\pm 6.2e^{-3}$)	70.9($\pm 7.0e^{-4}$)	63.6($\pm 7.0e^{-3}$)
ur	66.5($\pm 4.6e^{-3}$)	65.0($\pm 1.4e^{-3}$)	56.0($\pm 1.1e^{-2}$)
vi	72.6($\pm 1.1e^{-2}$)	69.6($\pm 7.8e^{-3}$)	62.4($\pm 1.5e^{-2}$)
zh	71.8($\pm 7.0e^{-3}$)	69.7($\pm 2.1e^{-3}$)	62.0($\pm 6.0e^{-3}$)

Table 9: Zero-shot transfer accuracy of XLMRoBERTa fine-tuned in different languages. Best results in bold, second best underlined.

when trained in Basque, it is the model that improves the most.

Table 7 shows the results obtained by the generative models. Once again, the models perform better when evaluated on the post-edited XNLIeu, but the performance gap is smaller compared with fine-tuned approaches. In this case, the relative ranking among models varies depending on the evaluation dataset, which suggests that post-edition introduces essential changes to the dataset and is therefore important to obtain a reliable evaluation benchmark. We further analyze this aspect in Section 6.

5.2 Results for the native test set

Table 8 shows the results of the models when evaluated on the native dataset. Overall, the ranking of the models (both discriminative and generative) remains unchanged. The translate-train approach still yields better results than zero-shot transfer, but the difference in accuracy between both approaches is on average 2% percentage points smaller than those obtained in the translated sets. This is likely a consequence of the mismatch between train and test datasets, because in this setting the former is built by translating English text while the latter is natively written in Basque.

Discriminative models perform worse on the native test set, with approximately 10% lower accuracy on average. While comparing results among different datasets is not always meaningful, we attribute the performance drop to the fact that the native dataset is less biased, as seen in Section 3.1. As a consequence, the models cannot rely on superficial patterns to deduce the relation between sentences, which makes this dataset especially challenging. Another possible cause is the notable presence of references to the Basque culture as it was sourced from original Basque materials.

Generative models do not yield worse results compared to the machine-translated and post-edited sets. This result is a consequence of the zero-shot prompting strategy followed in generative models, which does not include fine-tuning, and therefore does not rely on examples that can induce bias in the model.

5.3 Choice of the source language

We have conducted additional typological experiments to test the impact of the choice of the source language in a zero-shot transfer setting for Basque. For this, we fine-tuned XLM-RoBERTa-base in 14 languages using machine-translated versions of the MNLI training data, as well as English, and tested them on XNLIeu, XNLIeu$_{\mathrm{MT}}$ and the native test set. The results of these experiments are depicted in Table 9.

The Table shows small differences in XNLIeu and XNLIeu$_{\mathrm{MT}}$. We attribute these results to the fact that in this setting, both the training and test data come from translations, which lessens the importance of which source language to use. This is not the case for English, whose train data is original and not translated, which is not among the languages that achieve the highest results. However,

when a native dataset is used, factors such as proximity between languages and loanword frequency gain relevance, as shown in the Table, where the difference among languages is higher. Choosing English or Spanish yields similar results, while the performance when any other language is selected is noticeably lower.

6 Analysis

This Section provides additional analyses of the results. We begin by considering the performance of the best model on a per-label basis, followed by a manual comparison of the model outputs on the XNLIeu and XNLIeu$_{\mathrm{TM}}$ datasets.

6.1 Results per label

Figure 2 shows the confusion matrices on each label (entailment/neutral/contradiction) corresponding to the model and setting that performed best, XLM-RoBERTa large fine-tuned in Basque. For both XNLIeu and XNLIeu$_{\mathrm{MT}}$, the label that gets the higher F1 score is contradiction (87.7 and 83.4 respectively), followed by entailment (83 and 79.1), while neutral instances obtain the worst F1 score overall (80.7 and 76.4). This is in accordance with the analysis performed in Section 3.1, which indicates the presence of biases in these datasets, as well as in the training dataset. The results suggest that the models do rely on those biases, for instance by classifying instances where the hypothesis contains negative words as contradictions, or those where the hypothesis is short and has large lexical similarity with premises as entailment. On the other hand, no specific biases were detected in neutral instances, and consequently, it is more difficult for models to correctly classify them.

Section 3.1 reveals that the native dataset does not suffer from such apparent biases, and this is again reflected in the results depicted in Figure 2 for this dataset (right part). While contradiction is still the label with the best F1 score (80.3), now the label that attains the worst F1 is entailment (71.2), and the second-best is neutral (73.9).

6.2 Effects of post-edition

Section 5 reveals that systems perform consistently worse when evaluated on the machine-translated XNLIeu$_{\mathrm{MT}}$ dataset compared to the post-edited XNLIeu. So as to get a deeper insight into this result, we performed an analysis on XNLIeu and XNLIeu$_{\mathrm{MT}}$ by selecting instances that have been correctly predicted in one dataset and wrongly in the other. The analysis reveals that XNLIeu$_{\mathrm{MT}}$ often contains translation errors that change the relation between premise and hypothesis, and that when post-editing the professional translators corrected those errors. The most frequent error converts entailment and contradiction hypotheses to neutral. Common translation errors include:

- Changing the polarity of a sentence from negative to positive or vice versa.

  No, I live off campus. → *ez naiz campusetik kanpo bizi* 'I don't live off campus'

- Using an incorrect auxiliary verb, which can have a detrimental effect and completely change the meaning of a sentence.

  I was still scared. → *eta oraindik beldurra ematen dit* 'I am still scared'

- Omitting crucial information from the original sentence or occasionally creating nonsensical sentences.

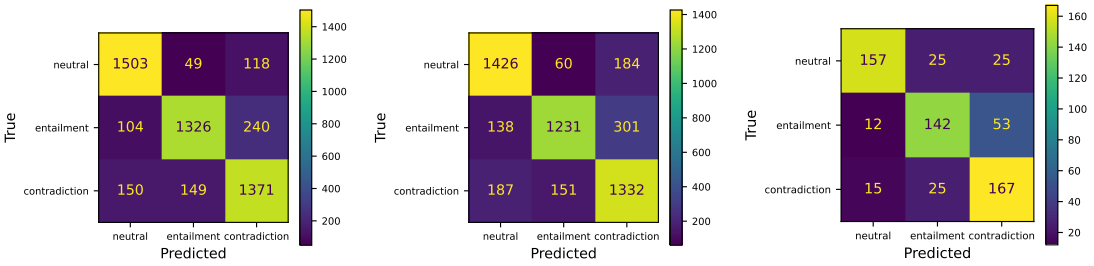
  I like feeling myself. → *Nik neuk gustuko dut ontzia.* 'I like the vessel myself'

On the other hand, there do not seem to be clear patterns in those instances that have been correctly predicted on XNLIeu$_{\mathrm{MT}}$ and incorrectly on XNLIeu. We have only found a handful of examples where the original label of XNLI is ambiguous and post-edition introduces necessary changes to make the translations accurate and fluent that can alter the relation between both sentences.

All in all, these differences lead us to the conclusion that the post-edition process is essential for the creation of a reliable benchmark.

7 Conclusions

In this work, we introduce XNLIeu, a new dataset for cross-lingual NLI in Basque. XNLIeu is developed by machine-translating the English part of XNLI followed by a post-edition step with the assistance of professional translators. Along with XNLIeu we release the machine-translated version, as well as a Basque native version carefully built to avoid known biases in NLI datasets. We conduct a series of cross-lingual Basque NLI experiments using a set of language models and different



XNLIeu XNLIeu$_{\text{MT}}$ Native

Figure 2: Confusion matrices for the XLM-RoBERTa large fine-tuned in Basque, our best model, tested with our three datasets. Best viewed in color.

cross-lingual strategies. The experiments show that translate-train is the best strategy, particularly when there is no mismatch between the origin of the train and test data. In the native dataset translate-train still yields the best results, but the difference is comparatively smaller. This result is consistent with previous works that analyze the impact of datasets based on translations. We also manually analyze the results of the models and find that machine-translation often introduces artifacts that change the meaning of the premises or hypotheses, and that professional translators correct those errors when post-editing. We conclude that post-edition is a crucial step towards reliable evaluation of cross-lingual NLI.

All of the datasets developed in this paper are publicly available under free licenses[5]. We believe that they are an important resource that will contribute to filling the gaps in resources that exist in Basque, that can hinder the development of research and applications with a focus on semantics in this language.

8 Limitations

Some limitations to this study should be taken into account, specially in the design of future research.

We have centered our work around the Basque language, which is considered to be a low-resource language. This means that, although some LLMs feature Basque in their training, there is not as much data and tools available as for other languages like English or Spanish. This was the main motivation for this research, but there is no prior work about NLI in Basque to be used as a reference, specifically in the experimental design and the interpretation of the results of the experiments.

Generative models are becoming more complex and versatile and are currently a popular subject of investigation. Most modern evaluation approaches are not focused on creating large corpora for specific tasks, but rather on testing generative models using prompt engineering and zero-shot or few-shot strategies. Our approach may seem outdated, as our research has focused mainly on the creation of our datasets and discriminative models, and generative models have only been tested with a zero-shot prompting approach. Future research for NLI in Basque should extend this line of research to account for the most recent developments and should include more insight into effective prompts and experiments performed with strategies other than zero-shot. However, we believe that the creation of our dataset and the approach we have followed are still pertinent for a low-resource language like Basque, which unfortunately does not include all the necessary resources to fully leverage the most recent advances brought by generative models, and can take advantage of a task like NLI, that enables the development of semantic applications and is useful for transfer-learning into a lot of different tasks.

[5]The dataset will be distributed with the same license as XNLI. The final URL will be available upon acceptance.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages?

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 7674–7684, Online. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Roman Dušek, Aleksander Wawer, Christopher Galias, and Lidia Wojciechowska. 2023. Improving domain-specific retrieval by nli fine-tuning.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages

180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Most frequent words in original Basque

Table 10 shows the original words that have been translated to English in Table 4.

	XNLIeu		**XNLIeu$_{\text{MT}}$**		**native**	
entailment	ez	0.58%	ez	0.54%	euskaraz	0.41%
	nuen	0.24%	nuen	0.23%	filma	0.24%
	zerbait	0.19%	batzuek	0.18%	dezakezu	0.24%
	batzuek	0.18%	zerbait	0.16%	pelikula	0.24%
	daitezke	0.17%	gustatzen	0.13%	munduko	0.24%
contradiction	ez	1.61%	ez	1.65%	ez	0.45%
	inork	0.24%	inork	0.23%	euskaraz	0.34%
	inoiz	0.2%	nuen	0.18%	euskara	0.28%
	nuen	0.18%	inoiz	0.16%	nire	0.23%
	nire	0.16%	axola	0.14%	bilboko	0.23%
neutral	ez	0.33%	ez	0.31%	gustatzen	0.37%
	nire	0.21%	dolar	0.2%	ez	0.37%
	nuen	0.19%	nire	0.2%	euskal	0.25%
	batzuek	0.18%	nuen	0.16%	batzuetan	0.25%
	gustatzen	0.15%	batzuek	0.16%	jende	0.25%

Table 10: Proportion of most frequent words in Basque.

Some common words (*nuen, daitezke, dezakezu*) have been translated to English as *auxiliary*. Auxiliaries are strictly grammatical words that do not hold semantic meaning. In Basque, verbal auxiliaries provide grammatical information about the tense, the mode and the person and number of the arguments of the action, the subject, the direct object and the indirect object.

B Prompts for the generative models

The prompst used for testing the generative models are based on the one by Gao et al. (2021) (Table 11).

prompt	**label**
[premise], right? Yes, [hypothesis]	entailment
[premise], right? No, [hypothesis]	contradiction
[premise], right? Also, [hypothesis]	neutral

Table 11: Prompts in English.

Table 12 shows the translation to Basque.

prompt	**label**
[premise], ezta? Bai, [hypothesis]	entailment
[premise], ezta? Ez, [hypothesis]	contradiction
[premise], ezta? Gainera, [hypothesis]	neutral

Table 12: Prompts in Basque.