Automatic Translation of Span-Prediction Datasets

Ofri Masad
Efi Arazi School
of Computer Science
Reichman University
ofri.masad@gmail.com

Kfir Bar
Efi Arazi School
of Computer Science
Reichman University
kfirbarmail@gmail.com

Amir David Nissan Cohen
Bar-Ilan University
amirdnc@gmail.com

Abstract

Generating high-quality non-English language datasets is crucial for achieving high performance in various Natural Language Processing (NLP) tasks. In this paper, we propose a new approach for translating NLP datasets that relies on a two-phase pipeline and online translation services. Our approach focuses on solving the alignment problem that affects span prediction tasks and utilizes automatically labeled data for training an alignment model. We demonstrate that our model-based approach shows higher accuracy than any other alignment method and improves the average F1 score on several Question-Answering (QA) datasets, specifically on the XQuAD Translated-train dataset, achieving new state-of-the-art results.

1 Introduction

During the past ten years, natural language processing (NLP) has rapidly developed in all aspects of research. New models, datasets, training platforms, and techniques are published almost daily while accuracy and efficiency are soaring to new heights. Yet, most works focus mainly on English, given its global dominance, and maybe a small number of other languages, while most languages get less attention. This lack of focus subsequently leads to a scarcity of resources for the majority of languages, as demonstrated in Appendix C, resulting in less performing models. Nevertheless, real-world NLP applications are needed in other languages, as they are in English. Focusing mainly on English may divert the research community's attention away from addressing linguistic features not typically found in English, such as dealing with intricate morphological structures. Creating large labeled datasets is labor intensive; it requires the engagement of experts in the domain and language in focus. Thus, non-English datasets tend to be less abundant and usually smaller.

In this work, we focus on automatically generating datasets for the traditional question-answering (QA) task, written in many diverse languages. In this task, the input consists of a question, a context, and an answer. The context is a short passage, and the answer is defined as a specific span of text within that context. The model is expected to predict the span of the answer within the given context. This task is deemed one of the foundational tasks in NLP and is frequently used as a performance measure for various models (Wang et al., 2018; Liu et al., 2019; Lan et al., 2020). Additionally, QA has recently been employed as a preliminary training step for models before they are trained on downstream NLP tasks. These tasks include event extraction (Du and Cardie, 2020), named entity recognition (Li et al., 2019, 2020), relation classification (Cohen et al., 2021), information extraction (Pires et al., 2022), and other downstream tasks (Hashavit et al., 2018).

Constructing a QA dataset from the ground up or through manual translation from another language requires significant work, time, resources, and a substantial level of expertise in NLP. Recently, the idea of utilizing automated translation tools to generate these datasets has been suggested as a way to reduce both costs and the amount of manual labor required (Abadani et al., 2021; Mozannar et al., 2019; Macková and Straka, 2020). Using automated translation tools in such settings presents challenges, which can often compromise the overall data quality. A fundamental challenge that arises is the identification of specific text spans in the translated document. The translation of a QA instance written in English includes translating the question, the context and the original span of the answer. However, locating the translated version of the answer in the translated context is not a straightforward task since the answer may appear in a different translated surface form, as dictated by the context.

Figure 1 illustrates this problem. For example, consider the answer "Greek" (highlighted in blue).

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 160–173
November 1–4, 2023.

This word is inherently ambiguous as it can refer to both nationality and language, which may have different forms in other languages. Identifying the accurate span of Greek in the translated version of this context becomes challenging, as we may not have prior knowledge of the original meaning we are looking for. Moreover, in some languages, the translation may undergo morphological modifications to account for gender, person, number, case, or other language-specific affixations. When translating the answer in isolation, the contextual meaning associated with it can be lost, resulting in a noticeable difference between the translated answer and its counterpart within the translated context.

As we show, simple string matching proves insufficient to identify the translated answer's span within the translated context. Instead, an alignment process becomes necessary after translation to establish the most suitable semantic equivalent match in the translated context. If the answer is not properly aligned during the translation process, the corresponding instance may be excluded from the translated dataset, leading to a reduction in the overall number of samples available. Conversely, misaligned answers compromise the quality of the dataset, emphasizing the critical role of precision and recall in the alignment process to ensure high-quality datasets.

In this paper, we aim to enhance the use of automated translation tools for translating span-prediction datasets, with a specific focus on QA from English to other languages. We employ commercially available translation tools and introduce a novel alignment model that can be easily trained for any language, improving the coverage of the target language.

We propose a new span alignment method, described in greater detail in Section 3, which formulates the alignment problem as a span extraction problem and uses a multilingual language model to predict the alignment. We define an automatic data labeling process for creating data for training that model and show that this new approach can generate high-quality QA datasets. Finally, in Section 4, we evaluate our new approach by generating machine-translated QA datasets in 13 languages, training QA models on these datasets, and comparing our training results to models trained on existing datasets in the same languages. The average improvement of our approach over those baseline models is +3.3 in F1 score and +2.9 in exact-match (EM) score.

English original

C: Symbiosis (from Greek "together" and "living") is close and often long-term interaction between two different biological species...

Q: What language does the word "symbiosis" come from?
A: Greek

Czech translation

C: Symbióza (z řeckého „spolu“ a „živ-ing") je úzká a často dlouhodobá interakce mezi dvěma různými biologickými druhy...

Q: V jakém jazyce je slovo "symbióza" pocházet z?
A: řecký

Figure 1: Example of a QA instance consists of a context (C), a question (Q), and an answer (A) from the SQuAD v1.1 dataset. The answer in the original English sample is highlighted in blue as a span within the context and on its own. The answer in the translated Czech sample is highlighted in red and appears in different surface forms within the context and on its own

In summary, we make the following contributions:

1. We formulate the span alignment task as a span extraction task and suggest a new model-based approach to address it.
2. We achieve state-of-the-art results in QA in nine languages.
3. We have made our code and the generated QA datasets in 13 languages publicly available. They can be accessed at the following URL: https://github.com/ofrimasad/translated_qa.

2 Related Work

The most popular QA dataset is SQuAD v1.1 (Rajpurkar et al., 2016), containing 100K question-answer pairs in English. It has been extended in

SQuAD v2.0[1] (Rajpurkar et al., 2018), with 50K questions that have no answer in the given content. A popular non-English version of this benchmark is the XQuAD[2] benchmark dataset for evaluating cross-lingual QA performance. This dataset consists of only 1,190 question-answer pairs from the development set of SQuAD v1.1 translated into ten languages by professional translators (Artetxe et al., 2020). It also includes the XQuAD-Translate-train dataset, a machine-translated version of the full SQuAD v1.1 train set.

The current state-of-the-art results on SQuAD v1.1 were achieved by using Google's T5-11B model (Raffel et al., 2019), with F1 and EM scores of 96.22 and 91.26, respectively. In comparison, the equivalent multilingual model, mT5, trained on XQuAD-Translate-train dataset achieved only 85.2/71.3 F1 and EM scores (Xue et al., 2021, 2022) averaged on ten languages (Arabic, German, Greek, Spanish, Hindu, Russian, Thai, Turkish, Vietnamese, Chinese), highlighting a significant gap of 20% in EM performance.

Moreover, even when using large amounts of unlabeled data for pre-training large language models (LLMs), better results can be achieved in the QA task by fine-tuning the model using a dataset labeled explicitly for the task. GPT-3, as an example, achieves F1 of 69.8 on SQuAD v2.0 (Rajpurkar et al., 2018) when attempting few-shot predictions. In contrast, the fine-tuned current state-of-the-art models achieve F1 of 93 on the same dataset (Brown et al., 2020).

Both versions of SQuAD (1.1, 2.0) have already been manually or automatically translated into other languages: Spanish (Carrino et al., 2019), Czech (Macková and Straka, 2020), Arabic (Mozannar et al., 2019), Swedish (von Essen and Hesslow, 2020), Dutch (van Toledo et al., 2022), Finnish (Kylliäinen and Yangarber, 2022), Bangla (Bhattacharjee et al., 2021), and Persian (Abadani et al., 2021). Moreover, equivalent datasets were created in French (Heinrich et al., 2022; d'Hoffschmidt et al., 2020), Russian (Efimov et al., 2019), Hebrew (Keren and Levy, 2021; Cohen et al., 2023), and Korean (Lim et al., 2019).

Although the majority of translated versions have been generated using a consistent translation approach, such as utilizing pre-trained machine translation (MT) models or online MT services supported by these models, alternative approaches have been proposed to address the alignment challenge mentioned earlier.

A naive approach to handling this problem is to keep only samples where the translated answer can be found in the text, using either simple string matching (Heinrich et al., 2022) or more advanced fuzzy-matching methods that include some text normalization operations (e.g., white spaces and punctuation removal, lower-casing) (Kylliäinen and Yangarber, 2022) and some other basic heuristics. Adopting these approaches often leads to the removal of many instances from the translated dataset. for example, in (Heinrich et al., 2022), around 60% of the instances were discarded during this process. Furthermore, as the matching becomes fuzzier or less precise, the likelihood of generating incorrect alignments increases. This ultimately reduces the quality of the translated dataset. We later demonstrate that significant data loss incurred during the translation process has a significant impact on the accuracy of a QA model trained using the translated dataset.

Another approach involves utilizing word relation trees, often extracted by a model. In this method, both the context and answer are lemmatized, and the stem of each lemma is extracted (Macková and Straka, 2020; Kylliäinen and Yangarber, 2022). The stem of the answer is searched within the list of stems present in the context. Depending on the implementation, either the entire lemma is considered as the answer, or some simplified reductions are applied. This method can also be applied to a set of words aligned individually from the source to the target language using a DL model (the span of the aligned words is used in this case) (Zhen et al., 2021). This process still results in a significant data loss of 18%-28%, as shown by (Kylliäinen and Yangarber, 2022).

A third approach includes adding markers or tags directly to the text or as additional tokens (von Essen and Hesslow, 2020). The markers are short strings usually added before and after the appearance of the answer in the source text. The idea is that these markers are designed to survive the translation process by being copied as-is to the target text and could be located after translation so that the translated answer will be located in the translated text. Typically, markers used for this purpose are designed to be distinguishable from running text. They often take the form of non-standard text

[1]Both versions are under CC-BY-SA-4.0 license
[2]CC-BY-SA-4.0 license

patterns or symbols, such as <p>, «H», ##, @@, and &&&. This approach poses a new challenge: the markers should be resilient enough to survive the translation process and show up in the translated text, while having minimal effect on the context of the text. Figure 2 shows how different markers are preserved during translation vs. the effect of these markers on the context of the sentence. A very resilient marker, like "[34456]", has a very high chance of being included in the translation. However, in many cases, it significantly impacts the translation quality. On the other hand, a less resilient marker (i.e., "__") has little effect on the translation quality, but it is dropped from the translated text very often (more than 12% of the times).

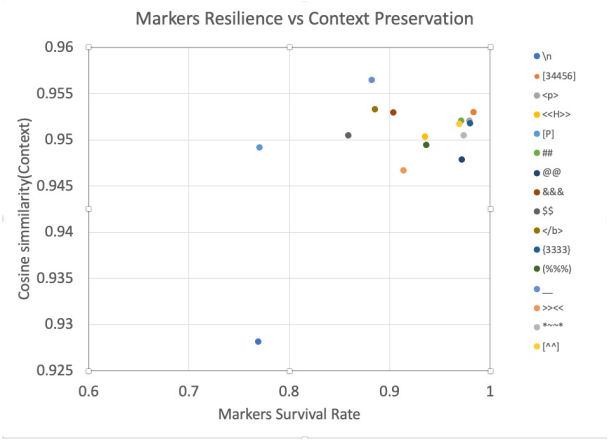

Figure 2: A plot of markers resilience vs. context preservation. Different markers are plotted in different colors. The survival rate is calculated as the percentage of times the marker has survived and appeared in the translation. Context is compared by calculating cosine similarity between the embeddings (generated by the multilingual model) of the text translated with and without the markers

Carrino et al. (2019) proposed another approach, using a Bayesian model with Markov-chain Monte Carlo (MCMC) inference for word alignments. The context is broken into sentences before translation, and each translated sentence is aligned with the original one. This process produced a complete word mapping from the source context to the translated context. Finally, they extracted the translated answer from the translated context using the mapping of the answer from the original context. This approach reduces data loss. However, it is less applicable to languages with a different morphology than English. To accommodate such languages, a work by von Essen and Hesslow (2020) presented two methods focusing on reordering the words. The initial word alignments are obtained using cosine similarity between the source and target texts, represented by embeddings generated using a multilingual model. Then, the Gromov-Wasserstein word distance matrix is minimized to force minimal word reordering while preserving the correct context of the answer. A work by (Lou et al., 2022) presented a similar approach with a different word distance matrix computation.

In a recent work proposed by von Essen and Hesslow (2020), a new approach was presented, in which a multilingual model was trained to align the translated answer and context. The model was trained using a contextual pyramid, holding a translated version of the span and its surroundings. When training, the task is to align this translated contextual pyramid to the correct span in the English text. During inference, the model is expected to carry out the same task, but to align the translated contextual pyramid to the translated Swedish text. The model is not directly trained on the task it is required to eventually perform due to a lack of training data. Instead, this method relies on the generalization abilities of a multilingual BERT model to solve this task by taking a zero-shot learning approach. The reported data loss using this method is only 8%.

Most methods depicted above apply some heuristics and basic statistical tools to solve the alignment problem. In this work, we show that the alignment problem should be solved using a more advanced algorithm, based on a language model, which has proven efficient in solving other NLP tasks.

3 Method

We propose a two-phase approach to tackle the alignment problem and generate span-prediction datasets of high quality. Our approach consists of the following two steps: 1) Train an alignment model for the target language; and 2) Translate the given dataset. We will now explain each step in more detail.

3.1 Step 1 - Train an Alignment Model for the Target Language

The alignment model is trained to accomplish the following task: given a translated sentence in the target language, and a phrase in English, find the span of the phrase within the translated sentence. Figure 3 illustrates the alignment process. The model predicts a span that closely aligns with the

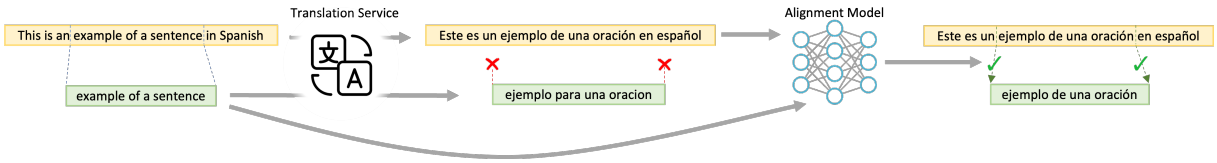

Figure 3: The translation and alignment process. Both the context (in yellow) and answer (in green) are translated. However, since the answer is translated out-of-context while its span inside the context is translated in-context, the translated answer does not appear in the translated context. The Alignment model takes the translated context and **English** answer, and predicts the span in the translated context

meaning of the translated phrase. We train the model to predict directly from the phrase in English. We train this model on a dataset that we automatically generate with labels. We start by translating only the contexts of SQuAD v1.1 (train) to the target language, using a machine-translation model, which we use as a black box. This allows us to utilize existing off-the-shelf models; in this work we use Google Translate.[3] We proceed to sample sentences from the translated dataset. Within each sentence, we randomly select a segment, defined as a few consecutive words. This segment will be our gold-label in the generated dataset. We then translate the selected segment into English. Note that the translation into English is done insensitive to the context. We make sure to select sentences and segments, according to the distribution of the length and position of the answers in SQuAD. The resulting dataset is formatted as a QA dataset, where each sentence we sample is considered as context, the selected segment as the answer, and the translated segment in English is treated as the question. Some examples from this dataset are listed in Table 5.

Following that, we proceed to fine-tune a pre-trained multilingual model[4] for the QA task on the dataset we have generated. To evaluate the performance of our model, we use a validation set that is created in a similar way to the one described earlier. The dataset also includes negative instances (questions that are impossible to answer based on the given context), as introduced in SQuAD v2.0. These instances are generated by choosing a phrase that does not appear in the sentence. By using negative instances, the model is trained to predict two values, the span of the answer, as well as the confidence level of the predicted span. When the confidence level is low it indicates that the model was unable to identify any span within the context that it deems as a suitable answer. We control the confidence level using a threshold value in order to reduce false predictions. We elaborate more on that in the next phase.

3.2 Phase 2 - Dataset Translation

We use the online Google Translate service to translate SQuAD v1.1 into other languages. Each context, question, and answer are translated together as a single unit to preserve as much context as possible in the translation process.

In Figure 2 we show how in-sentence markers may disappear in the translation, along with the impact they have on the translation quality. However, we noticed that if markers are placed between pairs of sentences, as sort of sentence delimiters, the markers predominantly remain intact, and the translation maintains its fidelity to the original text. Inserting such sentence-delimiter markers enables us to translate the context as one unit, as well as maintain sentence-level alignment between the source text of the context and its translated version. Using such markers is essential as the segmentation of the context into sentences in the source and target languages, does not always produce parallel sentences. We found that depending on the target language, between 7-15% of the contexts are divided into a different number of sentences in the target language. With the added markers, this was minimized to a range of 0.5-3%.

Following this, we would like to find the answer (written in English) in the translated version of the context. We refer to this procedure as answer alignment. To accomplish this, we employ three methods that essentially serve to complement one another:

1. First, if possible, we attempt to align the answer by locating it within the translated context using exact matching.

[3] https://translate.google.com

[4] We use a multilingual model since the phrase is in English while the span is searched in text written in the target language.

2. In cases where the answer cannot be aligned using exact matching, we use the alignment model described above to align the answer with the translated sentence.

3. Finally, when the alignment model predicts relatively low confidence, we segment the context into subsets of words with a total word count that approximates the word count of the answer. More formally:

$$\forall(w_i, \ldots, w_j) \subseteq (w_1, w_2, \ldots, w_m)$$

$$N_{ans} + 2 > j - i + 1 > N_{ans} - 2$$

where N_{ans} is the number of words in the answer, and m is the total number of words in the context. Then, we calculate the embeddings of the answer and all context segments using a pre-trained multilingual BERT model (cased), also known as mBERT, and use cosine similarity to find the closest segment to the answer. To prevent weak alignments, we set a threshold on the similarity value.

4 Experiments

We experiment on different languages and answer-alignment methods and compare our approach to other methods described in Section 2.

4.1 Experimental Settings

We evaluated our model on ten languages, using the Cross-lingual Question Answering Dataset (XQuAD) (Artetxe et al., 2019) as a benchmark. The creators of this dataset also released the Translate-train benchmark, in which SQuAD v1.1 (Rajpurkar et al., 2016) train set was automatically translated into ten languages using an automatic-translation model. We perform evaluations on the XQuAD dataset comprised of 240 paragraphs and 1,190 samples taken from the development set of SQuAD v1.1. Those instances were manually translated by professionals into the same ten languages. We use the Google Translate API service to translate the SQuAD v1.1 train set into ten languages. Both our alignment models and language-specific QA models are fine-tuned based on the multilingual BERT model (cased) (Devlin et al., 2019). We follow the XQuAD Translate-train benchmark, and assign the same values to all training hyperparameters. We train each model for the duration of three epochs, with a learning-rate value of $3.0e-5$, and a warm-up value of 6%. In all our training executions, we use a batch size of 8, a gradient accumulation of 8, and employ the widely-used AdamW optimizer. We set the model-based answer-alignment threshold to 0.05 and the cosine-similarity alignment threshold to 0.5.

Unfortunately, we did not have access to the same translation model used by the creators of XQuAD, as it seems the model used by XQuAD has better capabilities than the online Google Translation service provided by the vendor.

4.2 Results

We compare four groups of datasets, each containing datasets in ten languages translated from the original SQuAD v1.1 dataset:
Simple Alignment: created using Google Translate with only simple string matching
XQuAD Translate-train: created by the creators of XQuAD using a superior translation model than ours.
Simple + Cosine Alignment: created using Google Translate following two answer-alignment methods, string matching, and cosine similarity, as described in §3.2.
Full Alignment: our main approach. Created using Google Translate, and the full answer-alignment process described in §3.2.

Table 1 summarizes the F1 and EM scores on the XQuAD test set (in the same language as the train set) after training the multilingual BERT model on each one of the translated datasets. The table also provides the size of the translated train set as a percentage of the size of the SQuAD v1.1 original train set.

Compared to the baseline approach (Simple Alignment), which experiences a dataset size reduction of approximately 50%, our main approach (Full Alignment) maintains 93.4% of the original dataset size. Additionally, our main approach demonstrates improvements in both F1 and EM scores across all languages, with an average increase of 3.3% points in F1 and 2.98% points in EM. We attribute this improvement to the utilization of a larger number of samples in our approach.

When comparing our method to the XQuAD benchmark, we observe variations in performance across different languages. Nonetheless, our approach achieves an average improvement of 3.4% points in F1 and 2% points in EM over XQuAD. It is worth noting that XQuAD Translate-train outper-

	XQuAD Translate-train			Simple Alignment			Simple + Cosine Alignment			Full Alignment		
Language	**F1**	**EM**	**Size**	**F1**	**EM**	**Size**	**F1**	**EM**	**Size**	**F1**	**EM**	**Size**
Arabic (ar)	68.01	51.51	99.1%	68.06	51.51	56.8%	70.17	52.02	97.6%	**70.71**	**53.70**	94.2%
German (de)	74.67	60.08	94.3%	73.34	58.24	55.1%	75.13	57.90	97.8%	**76.84**	**61.18**	95.4%
Greek (el)	71.63	53.95	91.3%	68.43	50.34	48.2%	**73.43**	55.46	97.5%	72.83	**56.22**	89.2%
Spanish (es)	**79.30**	**62.27**	99.9%	77.17	59.66	55.2%	77.50	58.15	97.4%	79.27	**62.27**	96.1%
Hindi (hi)	70.08	**55.80**	98.0%	68.58	53.78	56.4%	70.13	53.53	97.8%	**70.82**	55.46	91.9%
Russian (ru)	**75.17**	**58.91**	96.9%	65.53	47.73	41.2%	73.22	54.29	97.1%	73.94	55.88	94.6%
Thai (th)	31.85	28.57	98.0%	59.01	51.26	52.2%	**63.59**	**55.97**	95.5%	61.95	53.45	86.8%
Turkish (tr)	69.51	**55.46**	98.8%	65.58	49.75	59.7%	66.75	49.92	97.8%	**69.54**	52.52	93.9%
Vietnamese (vi)	75.75	**56.55**	99.5%	75.85	55.21	58.1%	75.69	53.95	97.8%	**76.24**	54.20	95.5%
Chinese (zh-CN)	**66.20**	**56.60**	97.8%	61.13	53.78	56.3%	55.93	45.55	87.6%	63.56	55.29	96.1%
Average	68.2	54.0	97.3%	68.27	53.13	53.9%	70.15	53.67	96.4%	**71.57**	**56.02**	93.4%

Table 1: The results of mBERT-cased on the XQuAD test set, using the XQuAD Translated-train set as well as our datasets.

forms our method in certain languages. Nevertheless, our approach outperforms XQuAD in seven out of the ten tested languages. Notably, upon examining the language that yielded better results on the XQuAD dataset, we discover a significant difference in translation quality between XQuAD and our translation model. Based on this observation, we hypothesize that employing the same translation model utilized by our approach could potentially yield improved results. However, due to the unavailability of detailed information regarding the translation process of the XQuAD Translat-train dataset, we cannot provide a comprehensive analysis.

Alignment from English vs. Alignment from the Target Language. We found that using the English phrase as an input to the alignment model produced better results (1.7% F1 on average) than using the target-language phrase. We hypothesize that this happens due to some contextual biases added to the translated text. To demonstrate that, consider the following sample for alignment in Spanish:

> **C**: Cuando se encuentran por primera vez, se considera de buena educación inclinarse.
> **Q**: bow
> **A**: inclinarse

The Spanish sentence says: *"When you first meet, it is considered polite to bow"*. The alignment should be between the word "bow" in English and "inclinarse" in Spanish. But if we translate the word "bow" back to Spanish, we get the word "arco", which refers to another meaning of bow, a weapon (e.g., bow and arrow). Suppose the alignment model is trained using the translated phrase instead of the English one. In that case, it needs to handle a harder challenge, as the translated phrase "arco" and the target phrase "inclinarse" in the context, are completely different. The phrase "bow" in English is closer to the contextual meaning since it does not assume only one meaning of the word. The same concept can apply not only to words with multiple meanings, but also to words that appear in different surface forms depending on the context. All the results reported in this paper were achieved by training the alignment model using the English phrase. Table 3 shows a comparison between the results of the alignment model training with the two approaches.

Multiple Alignment Methods. We utilize a multi-method approach for aligning answers. Three different alignment techniques are applied based on the results of our experiments. Our findings indicate that basic string matching can successfully resolve 30-60% of the samples, contingent on the target language, consuming minimal resources and time. For the remaining unresolved samples, we utilize an alignment model along with the cosine similarity method described in Section 3.2. Preliminary testing showed that while cosine similarity achieves lower performance overall, it surpasses the alignment model on sequences with length ≥ 15 words. Thus, using cosine similarity as a complementary alignment method to model-based alignment improves overall results.

Size vs. Quality of the Dataset. An important insight that can be drawn from our results is that while preserving the maximal number of samples from the dataset in the translation is crucial, adding misleading samples counter-affects this. We observe that by using cosine similarity, we increased the average size of the dataset from 53.9% to 96.4%, but when using the model-based alignment process, we gain a larger improvement in F1 and EM scores, even though the average size of the generated dataset was only 93.4% of the original

one (smaller than the 96.4% we get by using cosine similarity). We illustrate this observation in Figure 4. While the size of the dataset grows monotonically when decreasing the threshold, we see that the F1 score is less predictable. When the threshold is over 0.2, the F1 gradually decreases correlated with the size of the dataset, but when the threshold is under 0.2, the F1 score is quite noisy. At this range of threshold values, the trade-off between adding more samples and adding misleading samples causes extreme changes in F1 over small changes in the threshold. In addition, we quantify the effect of using additional data on the overall model performance. We discuss this in Appendix A.

4.3 Using a Large Language Model to Generate the Dataset

To investigate the potential of LLMs in generating QA datasets in languages other than English, we employed the OpenAI GPT-3.5 API. Our initial experiment focused on generating samples comprising question-answer pairs within a context written in Hebrew. We sourced 50 distinct contexts from Hebrew Wikipedia. Each context was concatenated with five different English-written prefixes that were manually formulated to instruct GPT-3.5 (i.e., *Generate a SQuAD question and answer in Hebrew. Note that the answer must appear in the context itself. The context: <Hebrew context>*).

To evaluate the generated answers, we implemented a validation step by cross-referencing the answers with the corresponding context. Only answers present within the context were deemed usable. The obtained results revealed that the most successful prefix achieved a mere 18% usability rate. Notably, the handling of non-Latin languages by GPT-3.5 entails character-level tokenization, causing an elevated token count in requests and responses. Consequently, the cost estimation for producing an extensive dataset is prohibitively expensive due to the pricing structure based on price per 1k tokens.

In another experiment, we aimed to explore the feasibility of translating samples from the SQuAD v1.1 dataset into Hebrew, maintaining the alignment between the translated answer and the translated context. This experiment was designed to ascertain the LLM's effectiveness in performing high-quality translations while retaining the contextual coherence of the content.

Surprisingly, our efforts to identify an optimal prompt configuration to successfully translate the samples yielded discouraging results. The experiment recorded a success rate of less than 5%, indicating significant challenges in achieving accurate and reliable translations using the employed methodology.

These experiments collectively underscore the complexities involved in generating non-English QA datasets and performing accurate dataset translations using current LLMs.

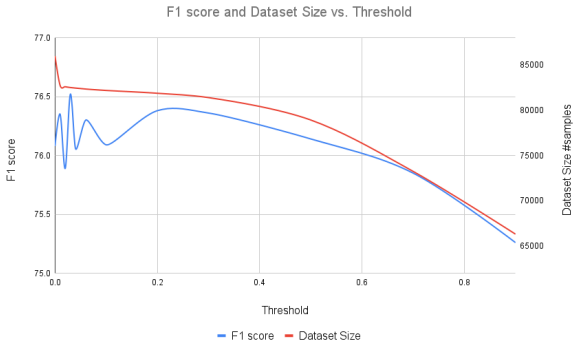

Figure 4: F1 score and dataset size vs. threshold value. Results of a BERT multilingual model trained on the SQuAD v1.1 dataset translated to German by using the alignment model with different thresholds and tested on the XQuAD-de test set.

Additional Languages. We conduct another set of experiments, to compare our general approach for dataset translation to the approaches described in Section 2. We collect the datasets and models used by previous works to translate QA datasets into Spanish, Swedish, Hebrew, and Czech. The results of these experiments are described in Table 2. The results show that our translation approach outperforms the other approaches in all languages except Swedish (von Essen and Hesslow, 2020).

These experiments facilitate a comparison between our automated, language-agnostic technique and some language-specific methodologies. We learned that the creators of the language-specific datasets possess intrinsic knowledge of the particular language, which can play a significant role when producing a dataset for a specified language. Some languages contain unique punctuation marks, different structures, and concepts that may be used to improve translation when handled correctly. Moreover, the evaluation process of a QA task includes a phase referred to as normalization. This process includes the removal of language-specific articles. In

Language	Model	Test Set	Train Set	F1	EM	Size
Spanish (es)	mBERT cased	XQuAD test set Spanish	SQuAD-es-TAR-train	77.6	61.8	100.0%
			Ours	**79.3**	**62.3**	96.1%
Swedish (sv)	mBERT cased	sv-dev-proj	proj-sv	**81.4**	**71.5**	99.1%
			Ours	80.1	70.7	96.8%
Hebrew (iw)	mBERT cased	ParaShoot test	ParaShoot train	56.1	32.0	1,792
			Ours	**66.6**	**41.2**	83,413
Czech (cs)	mBERT cased	SQuAD-cs v1.1[5]	SQuAD-cs (v1.1)	70.6	59.5	73.2%
	mBERT uncased		SQuAD-cs (v1.1)	73.9	62.1	73.2%
	mBERT cased		Ours	**76.4**	**65.8**	95.1%
	mBERT uncased		Ours	**76.7**	**66.3**	95.1%

Table 2: The results of mBERT trained on datasets translated by previous work (described in Section 2) and by our translation approach (described in Section 3). The results are reported on the dataset used by each study: Spanish (Carrino et al., 2019), Swedish (von Essen and Hesslow, 2020), and Czech (Macková and Straka, 2020). The Hebrew dataset, ParaShoot (Keren and Levy, 2021), was created manually (not a translation) and formatted similarly to SQuAD.

English, these articles are [*a, an, the*], in French [*le, la, less, l', du, des, au, aux, un, une*], and in German [*in, wine, einen, einem, wines, Weiner, der, die, das, den, dem, des*]. Not knowing the language-specific articles may dramatically affect the performance of the QA model, and unfortunately, there is not a source yet that outlines these articles for prevalent languages.

	English to Target		Target to Target	
Language	F1	EM	F1	EM
Arabic(ar)	**81.41**	**77.36**	80.44	74.13
German(de)	**81.80**	**78.79**	80.26	75.42
Greek(el)	78.34	**74.55**	**78.62**	73.73
Spanish(es)	76.42	**71.75**	**76.94**	70.53
Hindi(hi)	**80.06**	**73.55**	78.37	68.15
Russian(ru)	**81.74**	**77.93**	80.78	75.49
Thai(th)	**66.55**	**64.45**	64.79	61.91
Turkish(tr)	**82.49**	**77.45**	80.79	73.94
Vietnamese(vi)	**80.57**	**71.71**	80.35	70.14
Chinese(zh-CN)	**72.94**	**72.17**	70.44	69.14
Average	**78.23**	**73.97**	77.18	71.26

Table 3: The results of alignment model training in ten languages. Comparing two types of alignment, from English directly to the target language, and from the target language to itself.

5 Conclusion

In this paper, we presented a novel two-step approach for automatically translating span-prediction datasets. We have identified the alignment process as the differentiating component between different approaches, and formulated the alignment problem as a span extraction problem. We presented a method for training an alignment model and using such a model to obtain high-quality translations of instances of a QA dataset. The evaluation results show that our approach improves the quality of the datasets created through translation from English into 13 different languages, with an average F1 score improvement of 3.4%, achieving state-of-the-art results on XQuAD.

6 Limitations

Our approach is fully dependent on the quality of the machine translation system for the target language. Although machine translation systems are available for most languages, they might not yet be available for some less common languages. The quality of the said machine translation system might affect the result of our process and produce lesser results. Any biases in the machine translation system may be inherited by the resulting dataset, which may lead to bias confirmation.
Moreover, our approach requires either local computing resources or access to online services. These resources might be expensive or limited when used under the free usage terms.

7 CO2 Emission Related to Experiments

Experiments were conducted on a RTX 3090 GPU (TDP of 350W). A cumulative of 5-9 hours per language to train the alignment model + 3 hours to train and test the final model which evaluates the dataset's quality. Total emissions are estimated to be 1.68 kgCO_2eq per language or 30.3kgCO_2eq for all experiments.

8 Acknowledgements

We would like to express our gratitude to Gal Rapoport for his valuable contributions and technical discussions throughout the course of this project.

References

Negin Abadani, Jamshid Mozafari, Afsaneh Fatemi, Mohammd Ali Nematbakhsh, and Arefeh Kazemi. 2021. ParSQuAD: Machine translated SQuAD dataset for persian question answering. In *2021 7th Interna-*

tional Conference on Web Research (ICWR), pages 163–168.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. In *Annual Meeting of the Association for Computational Linguistics*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, Md. Saiful Islam, Wasi Uddin Ahmad, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2021. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *NAACL-HLT*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *CoRR*, abs/2005.14165.

Casimiro Pio Carrino, Marta Ruiz Costa-jussà, and José A. R. Fonollosa. 2019. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *International Conference on Language Resources and Evaluation*.

Amir DN Cohen, Hilla Merhav Fine, Yoav Goldberg, and Reut Tsarfaty. 2023. Heq: a large and diverse hebrew reading comprehension benchmark.

Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. 2021. Relation classification as two-way span-prediction.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Pavel Efimov, Leonid Boytsov, and Pavel Braslavski. 2019. SberQuAD - Russian reading comprehension dataset: Description and analysis. In *Conference and Labs of the Evaluation Forum*.

Anat Hashavit, Naama Tepper, Inbal Ronen, Lior Leiba, and Amir DN Cohen. 2018. Implicit user modeling in group chat. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, page 275–280, New York, NY, USA. Association for Computing Machinery.

Quentin Heinrich, Gautier Viaud, and Wacim Belblidia. 2022. FQuAD2.0: French question answering and learning when you don't know. In *International Conference on Language Resources and Evaluation*.

Omri Keren and Omer Levy. 2021. ParaShoot: A Hebrew question answering dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 106–112, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilmari Kylliäinen and Roman Yangarber. 2022. Question answering and question generation for Finnish. *ArXiv*, abs/2211.13794.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD1.0: Korean QA dataset for machine reading comprehension. *ArXiv*, abs/1909.07005.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chenwei Lou, Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, Weiwei Tu, and Ruifeng Xu. 2022. Translation-based implicit annotation projection for zero-shot cross-lingual event argument extraction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2076–2081, New York, NY, USA. Association for Computing Machinery.

Katerina Macková and Milan Straka. 2020. Reading comprehension in Czech via machine translation and cross-lingual transfer. In *Workshop on Time-Delay Systems*.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Ramon Pires, Fábio C. de Souza, Guilherme Rosa, Roberto A. Lotufo, and Rodrigo Nogueira. 2022. Sequence-to-sequence models for extracting information from registration and legal documents. In *Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings*, page 83–95, Berlin, Heidelberg. Springer-Verlag.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Chaïm van Toledo, Marijn Schraagen, Friso van Dijk, Matthieu Brinkhuis, and Marco Spruit. 2022. Exploring the utility of Dutch question answering datasets for human resource contact centres. *Information*, 13(11).

Hannes von Essen and Daniel Hesslow. 2020. Building a Swedish question-answering model. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 117–127, Gothenburg. Association for Computational Linguistics.

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.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ranran Zhen, Rui Wang, Guohong Fu, Chengguo Lv, and Meishan Zhang. 2021. Chinese opinion role labeling with corpus translation: A pivot study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10139–10149, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A F1 Scores on Subsets of SQuAD v1.1

Generally speaking, increasing the size of the dataset improves the model's performance, up to a certain threshold. It can be seen in Figure 5, which shows the F1 scores reached by the same model trained on different sizes of subsets of the SQuAD v1.1 dataset (in English). In this case, the model still benefits from increasing the number of samples up to 100% of the dataset.

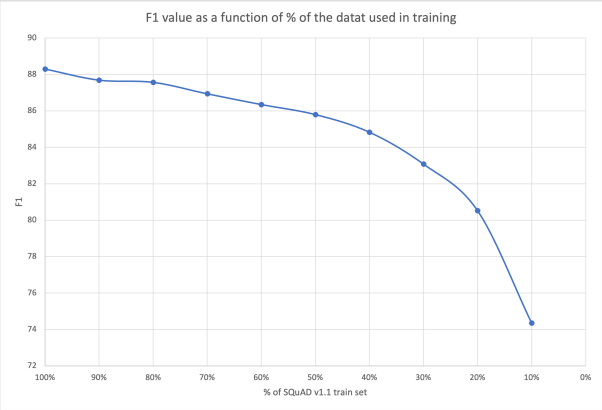

Figure 5: F1 score vs. size of train set (as a percentage of the full SQuAD v1.1 train set). Each data point represents the results of a different mBERT model trained on a subset of the SQuAD v1.1 train set and evaluated on the SQuAD v1.1 development set

B Results on Translated SQuAD v2.0

	Simple Alignmen		Full Alignment	
Language	**F1**	**EM**	**F1**	**EM**
Arabic(ar)	68.71	54.59	**72.68**	**58.69**
German(de)	74.58	61.03	**78.01**	**64.51**
Greek(el)	70.12	53.82	**74.27**	**59.96**
Spanish(es)	80.41	63.50	**82.98**	**66.97**
Hindi(hi)	70.92	56.54	**72.86**	**58.67**
Russian(ru)	65.04	47.96	**74.68**	**59.74**
Thai(th)	44.98	38.78	**49.05**	**43.96**
Turkish(tr)	70.96	55.62	**73.73**	**58.37**
Vietnamese(vi)	77.77	57.45	**79.61**	**58.77**
Chinese(zh-CN)	63.36	59.55	**65.60**	**62.41**
Czech(cs)	72.34	59.51	**77.92**	**66.19**
Hebrew(iw)	70.07	56.58	**72.27**	**58.01**
Swedish(sv)	78.63	68.07	**80.24**	**70.22**
Average	69.84	56.38	**73.38** (+3.54)	**60.50** (+4.11)

Table 4: Results of BERT-multilingual(cased) trained on SQuAD v2.0 train set and evaluated on SQuAD v2.0 development set, both translated to different languages. Simple alignment baseline refers to simple matching defined in §3.2. Full Alignment refers to our main approach. Created using Google Translate, and the full answer-alignment process described in §3.2. our approach achieves an average improvement of 3.4 percentage points in F1 and 2 percentage points in EM over the baseline

C Resource Availability of Common Languages

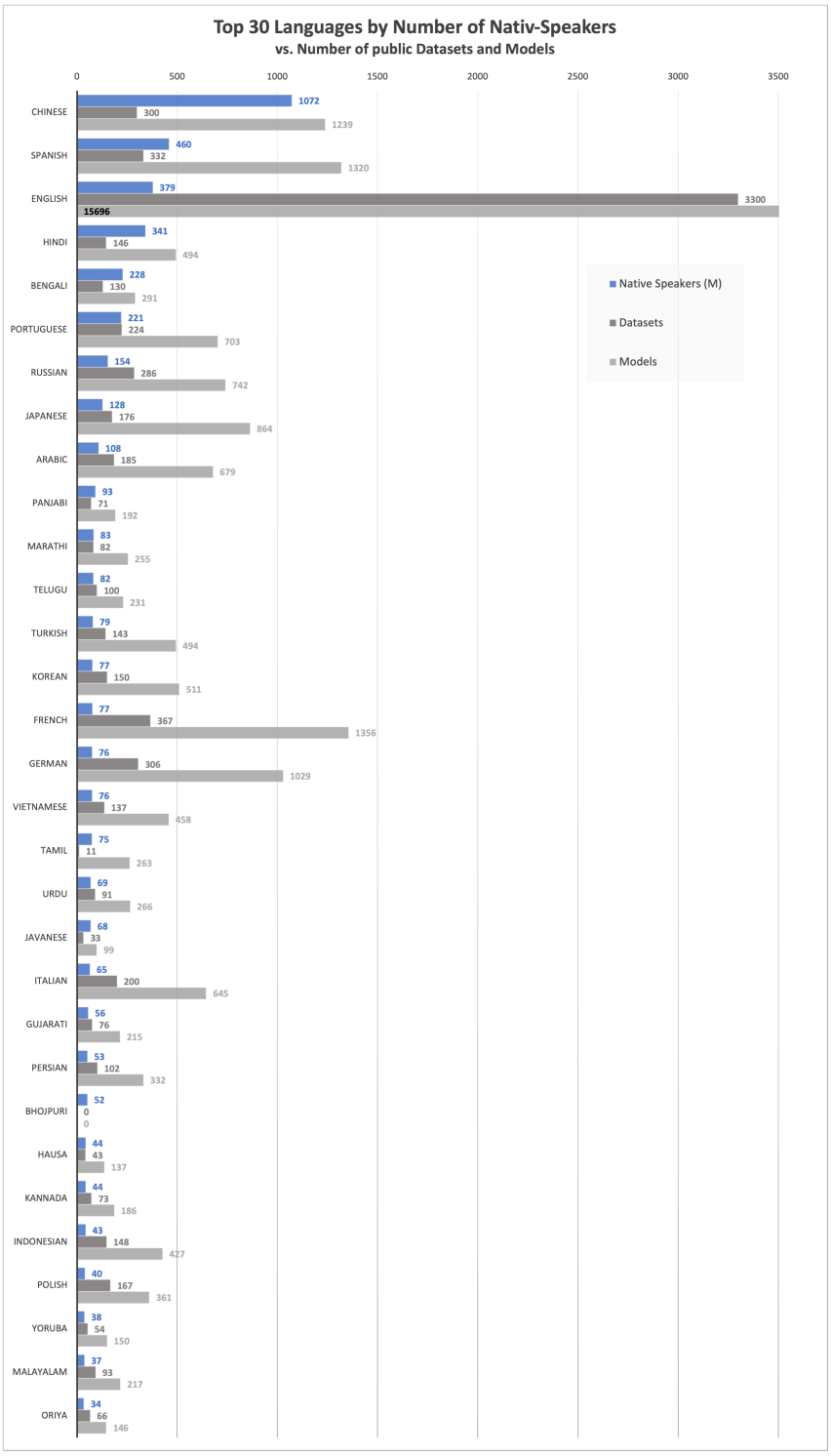

Figure 6: Top 30 languages sorted by native speakers vs. the number of public datasets and models available on the popular HuggingFace hub (Lhoest et al., 2021) (as of May 2023). There are 15696 models in English compared to only 14302 models in all other languages combined, and 3300 datasets in English compared to 4292 datasets in all other languages combined (English native speakers are 8.5% of all native speakers of the 30 languages)

D Samples from the Alignment Dataset

English Phrase	Target Language Phrase (span)	Context
Pointe-Noire and along the Atlantic coast.	Pointe-Noire und entlang der Atlantikküste. (103-146)	Die bedeutendsten Untergruppen des Kongo sind Laari in den Regionen Brazzaville und Pool sowie Vili um Pointe-Noire und entlang der Atlantikküste.
16 year olds, but in practice	16-Jährige, aber in der Praxis (78-108)	Die öffentliche Bildung ist theoretisch kostenlos und obligatorisch für unter 16-Jährige, aber in der Praxis fallen Kosten an.
less than the 79%	weniger als die 79% (72-92)	Die Netto-Einschulungsrate im Grundschulbereich lag 2005 bei 44%, viel weniger als die 79% im Jahr 1991.
of the boom and the	des Aufschwungs und der (186-209)	Die derzeitige Regierung herrscht über einen unruhigen inneren Frieden und steht trotz der seit 2003 rekordhohen Ölpreise vor schwierigen wirtschaftlichen Problemen bei der Stimulierung des Aufschwungs und der Verringerung der Armut.
government	Regierung (15-24)	Die derzeitige Regierung herrscht über einen unruhigen inneren Frieden und steht trotz der seit 2003 rekordhohen Ölpreise vor schwierigen wirtschaftlichen Problemen bei der Stimulierung des Aufschwungs und der Verringerung der Armut.
Peace	Frieden (63-70)	Die derzeitige Regierung herrscht über einen unruhigen inneren Frieden und steht trotz der seit 2003 rekordhohen Ölpreise vor schwierigen wirtschaftlichen Problemen bei der Stimulierung des Aufschwungs und der Verringerung der Armut.
he professed to be public	bekundete er öffentlich sein (82-110)	Als Sassou Nguesso am Ende des Krieges im Oktober 1997 an die Macht zurückkehrte, bekundete er öffentlich sein Interesse daran, Wirtschaftsreformen und Privatisierungen voranzutreiben und die Zusammenarbeit mit internationalen Finanzinstitutionen zu erneuern.

Table 5: Samples from the dataset used for training the alignment model (German in this example)