Cross Lingual Extractive Question Answering

Linguists - Team 20

November 2024

1 Introduction

The problem focuses on on evaluating the cross lingual transfer capability of QA systems across multiple languages. Current QA systems are usually designed to work in just one language, but in many cases, people need answers in different languages or from sources in multiple languages. This is particularly important in multilingual societies or when translating content, such as legal documents, research papers, or customer support across regions.

The paper MLQA proposes two main tasks to evaluate how well a model can perform in cross-lingual QA, using a dataset called MLQA (Multilingual Question Answering).

- **Cross-lingual Transfer (XLT):** In this task, we train the model using data in one language (English, for example). The model learns from pairs of text (context), questions, and answers in this language. During testing, however, the model must answer questions in a different language (e.g., Spanish). This means it has to understand and extract answers from a different language than what it was trained on.
- **Generalized Cross-lingual Transfer (G-XLT):** In G-XLT, the model is trained using question answering data where all components—context (cx), question (qx), and answer (ax)—are in the same language, typically English. The challenge of G-XLT comes at test time, where the model is evaluated in a multilingual setting. Specifically, the model is required to answer questions presented in one language (qy) while extracting the correct answer (az) from a context provided in an entirely different language (cz).
- **Contextual Cross-lingual Question Answering (C-XQA):** In this task, the model is trained with data in multiple languages where the context, questions, and answers are in the same language. The model is tested in a multilingual setting as before.

These tasks were performed on the following languages: **English** (en), **German** (de) and **Spanish** (es).

2 Datasets

2.1 SQuAD

SQuAD 1.1 [3](Stanford Question Answering Dataset) is a benchmark dataset for training and evaluating question-answering models. It consists of over 100,000 question-answer pairs derived from a set of Wikipedia articles. It has 87599 training pairs and 10570 validation pairs. Each question is accompanied by a corresponding passage, and the goal is to identify a span of text within the passage that serves as the correct answer to the question. The dataset is widely used due to its high-quality annotations and diverse topics, making it a standard for evaluating reading comprehension capabilities of models.

2.2 PaXQA

The **PaXQA** [2] (Parallel Cross-lingual Question Answering) dataset is designed for cross-lingual question answering tasks containing question-answer pairs in multiple languages including English, Chinese, and

Arabic. The dataset was created to facilitate research in multilingual and cross-lingual question answering systems.

We fine-tuned the BertForQuestionAnswering model using pairs from the dataset: English-English (50k), Chinese-Chinese (28k), and Arabic-Arabic (45k), with BERT scores greater than 0.6. The results on en-en are as follows:

- **Exact Match (EM):** 1.60
- **F1 Score:** 6.82

2.2.1 Ambiguity in the PaXQA dataset

Upon examining sample entries from the dataset, we observed significant ambiguities that hinder the model's ability to generate accurate answers. Some of the ambiguous samples:

Example of Question Answering Dataset

QA Dataset Example 1

Context:
The convict "pulled the ladder on which the victim was standing, which brought him down. He then beat him with an iron rod on his head several times until he died. He then wrapped his body with a sheet, put it in the car, and threw it away in one of the mountains to hide his crime, following a disagreement between the two," according to the ministry. The sentence was implemented in Aseer (south). At least 153 people were executed in 2007 and at least 37 in 2006, and at least 83 were executed in 2007 and 35 in 2004, while 113 people were executed in 2000. Saudi Arabia penalizes with execution the crimes of rape, murder, conversion from Islam, armed assault, and drug smuggling.

Question: How many executions were there in 2007?

Answer: {
Answer start: [481],
Text: [83]
}

In this example regarding executions in Saudi Arabia, the context presents conflicting figures for the number of executions in 2007, citing both 153 and 83. While the answer "83" is included, the presence of two different numbers creates confusion regarding which figure is accurate.

QA Dataset Example 2

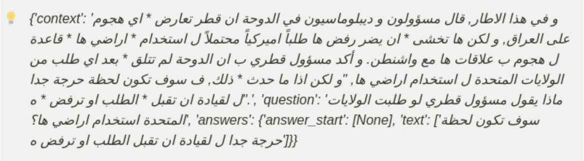
Context:
Wayne Whittaker, Beńs father, told the BBC that the physicians said that Ben would be in severe pain if he were a young teenager and had suffered this injury. ”Things Get More Difficult.” After PRO conducting a series of tests on Ben, it appeared that he was afflicted * with a congenital defect that * prevented him from PRO feeling pain, which * is an extremely rare condition. The experts hypothesize that the condition is attributable to a deficiency in a substance in the body called * PRO beta-endorphin, which * is a substance that * is spontaneously secreted * in the body PRO to make humans feel pain.

Question: What is Ben’s congenital defect?

Answer: {
  Answer start: [160],
  Text: [Things Get More Difficult]
}

In this instance involving Ben’s congenital defect, the provided answer, ”Things Get More Difficult,” does not correctly represent the condition itself. The accurate answer should pertain to his inability to feel pain, highlighting a significant gap in the dataset’s quality.

QA Dataset Example 3



{'context': 'و في هذا الاطار, قال مسؤولون و ديبلوماسيون في الدوحة ان قطر تعارض * اي هجوم
على العراق, و لكن ها تخشى * ان يضر رفض ها طلباً اميركياً محتملاً ل استخدام * اراضي ها * قاعدة
ل هجوم ب علاقات ها مع واشنطن. و أكد مسؤول قطري ب ان الدوحة لم تتلق * بعد اي طلب من
الولايات المتحدة ل استخدام اراضي ها, "و لكن اذا ما حدث * ذلك, ف سوف تكون لحظة حرجة جدا
ل لقيادة ان تقبل * الطلب او ترفض * ه".', 'question': 'ماذا يقول مسؤول قطري لو طلبت الولايات
المتحدة استخدام اراضي ها؟', 'answers': {'answer_start': [None], 'text': ['سوف تكون لحظة
حرجة جدا ل لقيادة ان تقبل الطلب او ترفض ه']}}

This example illustrates an annotation error, where the `answer_start` is marked as `None` but `text` is `not None`. This indicates a failure in the dataset’s annotation process, leading to ambiguity about where the answer can be found within the context.

QA Dataset Example 4

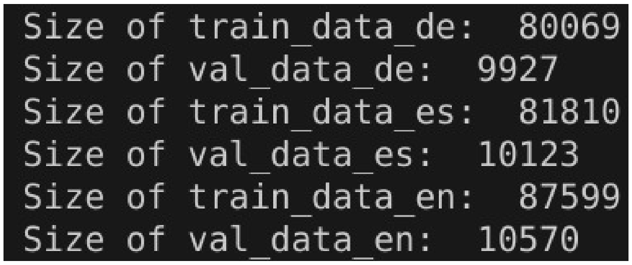
Context:
Shanghai, which * has entered into the post-industrialization mid-development stage, has been providing wide investment scope for foreign businesses from various countries. Shanghai is the largest modernized industrial city in China and has complete industry categories, developed infrastructure and perfect municipal functionality. The financial center, business center and shipping center of the inland of China are also here. The Pudong development and the Yangtze River Valley development strategy spearheaded * by Shanghai implemented * by the central government have redoubled the strength of this city. The good investment environment makes the success rate of investment by foreign businesses in Shanghai reach 80
Question: What is the name of China’s business center?

Answer: {
  Answer start: [337],
  Text: [financial center]
}

In this example about Shanghai, the question asks for the name of China’s business center, but the answer provided is just ”financial center,” which doesn’t fully answer the question.

In the PAXQA paper, it is mentioned that the test results improved after training on the SQUAD dataset. This indicates that the PAXQA dataset has significant ambiguity, which likely affects the quality of the results. Due to these inconsistencies, such as conflicting information and incorrect answers, the dataset does not provide reliable training data for question-answering models. Recognizing these limitations, we have decided not to use the PAXQA dataset for our work, as it would hinder our model's performance.

2.3 MT-SQuaD

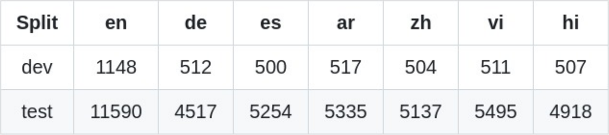


```
Size of train_data_de:  80069
Size of val_data_de:  9927
Size of train_data_es:  81810
Size of val_data_es:  10123
Size of train_data_en:  87599
Size of val_data_en:  10570
```

In the MLQA paper, the authors translated instances from the SQuAD training set into the target language using machine-translation using Facebook's production models.

2.4 MLQA

For testing, we used **MLQA**[1] (a multi-way aligned extractive QA evaluation benchmark). MLQA contains QA instances in 7 languages, English, Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese. MLQA has over 12K instances in English and 5K in each other language, with each instance parallel between 4 languages on average. Combined, there are over 46,000 QA annotations.



Split	en	de	es	ar	zh	vi	hi
dev	1148	512	500	517	504	511	507
test	11590	4517	5254	5335	5137	5495	4918

3 Experiments

To evaluate the performance of QA systems in cross-lingual settings, we explored three model architectures: encoder-only models, encoder-decoder models, and decoder-only models. Each model type is fine-tuned using the same training and testing setup, ensuring a consistent comparison of their capabilities across the tasks defined by MLQA: Cross-lingual Transfer (XLT), Generalized Cross-lingual Transfer (G-XLT), and Contextual Cross-lingual Question Answering (C-XQA).

3.1 Models and Approaches

- **Encoder-Only Models**
  - **Models Used**: `Multilingual BERT`, `XLMRoberta`
  - These models are fine-tuned for question answering by learning to predict start and end token positions of answers within a given context.
- **Encoder-Decoder Models**
  - **Models Used**: `MultiLingual T5`, `Multilingual Bart`

  - Encoder-decoder models are trained to generate answers as text sequences conditioned on the input question and context.

- **Decoder-Only Models**
  - **Models Used**: `GPT2`
  - Decoder-only models are fine-tuned as generative models, focusing on completing the input question and context to generate the correct answer.

3.2 Training

- **Task 1: Training on SQuAD**
  In this task, the models are fine-tuned on the SQuAD dataset, a widely used English-language dataset containing context, questions, and corresponding answers. The training data included context-question-answer triples exclusively in English (en-en), which allows for testing the generalization capabilities of the models in cross-lingual settings.

- **Task 2: Training on MT-SQuAD**
  In this task, the models are fine-tuned on the MT-SQuAD dataset by combining the context-question-answer triples of the three different languages (English, Spanish and German). This allows for testing if there is an improvement in models' performance compared to the ones trained only on English(SQuAD).

3.3 Testing

The models trained as above are evaluated on the MLQA dataset using the following strategies:

- **Same-Language Testing (XLT)**
  The models are tested on mono-lingual pairs (`en-en`, `de-de`, `es-es`). This corresponds to Cross-Lingual Transfer (XLT) and evaluates how well the models perform when the context-question-answer triples are in the same language.

- **Cross-Language Testing (G-XLT)**
  The models are tested on cross-lingual pairs, including `en-de`, `en-es`, `de-en`, `de-es`, `es-en`, and `es-de`. In these cases, the context and question are in different languages, and the answer has to be given in the corresponding context language, requiring the models to demonstrate robust cross-lingual understanding and generalization.

4 Metrics

4.1 Encoder-only Models

We used **EM** (exact match) and **F1** score as performance metrics because its predictions are span based.

EM (Exact Match): It is a metric used to evaluate tasks like question answering. It measures the percentage of predictions that exactly match the ground truth answer, ensuring both the content and format are identical.

F1 Score: It is a metric that combines precision and recall to measure the accuracy of a model's predictions. It is particularly useful in scenarios like question answering, where partial matches are rewarded, as it balances the trade-off between precision and recall.

4.2 Decoder-only & Encoder-Decoder Models

We used **Rouge-l** and **BERT** score as performance metrics because its predictions are generative.

ROUGE-l (Recall-Oriented Understudy for Gisting Evaluation): It is a metric used to evaluate text generation tasks. It compares the overlap of n-grams, or word pairs between the generated text and the reference text, focusing on recall to measure content similarity.

$$\text{RC} = \sum_{i \in S_e, j \in S_o} a_{ij} \Big/ \sum_{i \in S_e} a_{i*}$$
$$\text{PR} = \sum_{i \in S_e, j \in S_o} a_{ij} \Big/ \sum_{j \in S_o} a_{*j}$$
$$\text{F1} = 2 * \text{RC} * \text{PR} \Big/ \text{RC} + \text{PR}$$

Figure 1: F1 Score

BERT Score: It is a metric for evaluating text generation tasks that leverages pre-trained language models like BERT. It computes similarity scores between token embeddings of the generated and reference texts, capturing semantic alignment beyond exact matches or n-gram overlap.

4.3 Normalizing answers

Before calculating metrics both the generated answers and ground truths are normalized using below steps:

- Lowercasing
- Removing Punctuation
- Removing Articles
- Whitespace Normalization

5 Results

5.1 Encoder only Models

We have utilized SQuAD 1.1 and MT SQuAD for training the `BertForQuestionAnswering` and `XLMRobertaForQuestionAnswering` are fine-tuned using the hyperparameters **batch_size** = 128, **lr**=1e-4 and **epochs** = 3,**optimizer** = optim.AdamW(model.parameters().

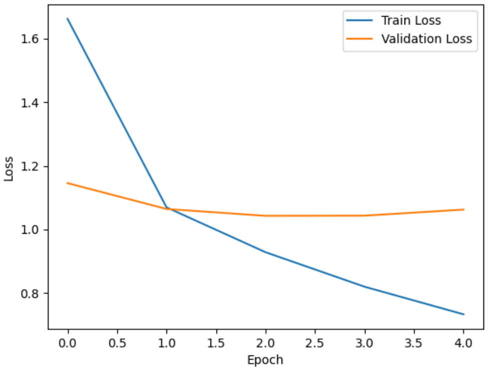


Figure 2: Bert on SQuAD

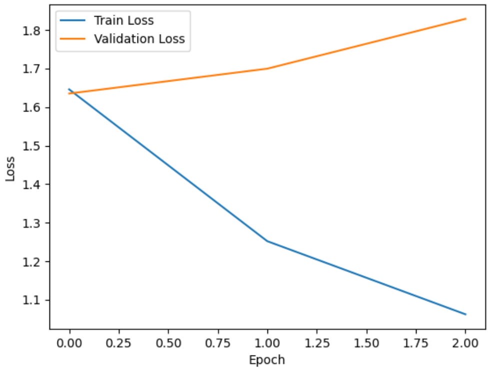


Figure 3: Bert on MT-SQuAD

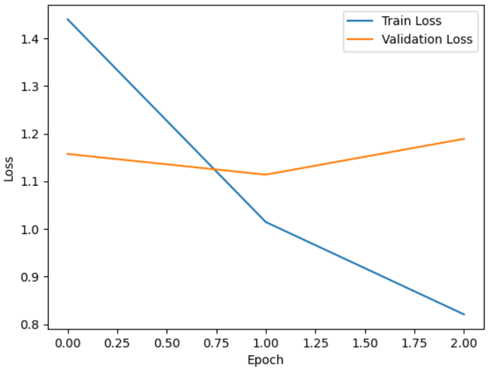


Figure 4: Roberta on SQuAD

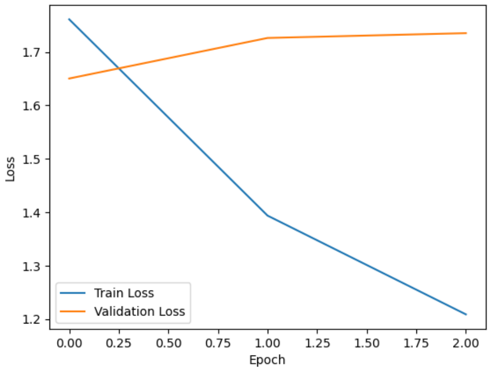


Figure 5: Roberta on MT-SQuAD

	en-en[1]	en-de	en-es	de-en	de-de	de-es	es-en	es-de	es-es
mBERT-v1[2]	63.71	53.71	51.44	45.85	45.89	40.32	48.18	41.84	46.64
mBERT-v2[3]	61.53	55.01	52.66	38.88	39.43	37.56	42.98	42.00	44.13
XLM-Roberta-v1	61.88	37.55	31.43	42.40	42.24	19.93	45.36	26.07	43.58
XLM-Roberta-v2	61.98	48.64	49.84	40.12	41.33	36.26	44.93	39.02	46.13

Table 1: EM

	en-en	en-de	en-es	de-en	de-de	de-es	es-en	es-de	es-es
mBERT-v1	76.95	66.14	64.76	59.85	59.96	53.87	65.50	59.51	64.32
mBERT-v2	75.23	67.26	67.00	55.72	56.48	54.13	62.49	60.32	64.00
XLM-Roberta-v1	74.70	48.18	42.63	55.56	56.39	29.82	61.76	39.56	60.41
XLM-Roberta-v2	74.51	59.63	62.44	55.18	56.87	50.33	62.60	56.42	64.07

Table 2: F1 Score

Cross-Lingual Transfer in v1 Models (English-Only Training)

XLT (x-x):

- The highest performance is observed for *en-en* (mBERT-v1: 63.71, XLM-Roberta-v1: 61.88), which aligns with the training language. However:
  - For *de-de* and *es-es*, performance is considerably lower (e.g., mBERT-v1: 45.89 for de-de and 46.64 for es-es). Similarly, XLM-Roberta-v1 struggles with *de-de (42.24)* and *es-es (43.58).*
- The significant drop in non-English x-x performance suggests that v1 models do not generalize well to unseen languages in a monolingual context.

G-XLT (x-y):

- The performance on x-y pairs (e.g., *en-de, en-es, de-en, de-es, es-en, es-de*) is generally lower than en-en but better than non-English x-x scores:
  - For mBERT-v1:

[1] An x-y pair refers to a context-question pair, where x represents the context and y represents the question.
[2] v1 models are trained exclusively on English
[3] v2 models are trained on the combined data (English, Spanish and German)

        - *en-de: 53.71, en-es: 51.44, de-en: 45.85, de-es: 40.32, es-en: 48.18, es-de: 41.84.*
        - These scores indicate moderate transfer from English to other languages, particularly for target languages (de, es). Reversed pairs (e.g., *de-en, es-en*) also perform relatively well due to English's strong representation in pretraining.
    - For XLM-Roberta-v1:
        - *en-de: 37.55, en-es: 31.43, de-en: 42.40, de-es: 19.93, es-en: 45.36, es-de: 26.07.*
        - The drop in x-y performance compared to mBERT-v1 highlights weaker cross-lingual transfer, particularly for pairs where the context is non-English (e.g., de-es, es-de).

Key Insight for v1 Models:

- Cross-lingual transfer is better when English is involved (e.g., *en-de, de-en*) but deteriorates significantly for pairs that do not include English (e.g., *de-es, es-de*). The reliance on English embeddings is evident, limiting transfer capabilities across non-English pairs.

Cross-Lingual Transfer in v2 Models (Multilingual Training on en, de, es)

XLT (x-x):

- Training on en, de, and es improves monolingual (x-x) scores across all languages:
    - For mBERT-v2:
        - *en-en: 61.53 (slightly lower than mBERT-v1), de-de: 39.43, es-es: 44.13.*
        - These scores suggest marginal improvement for non-English x-x pairs compared to mBERT-v1.
    - For XLM-Roberta-v2:
        - *en-en: 61.98, de-de: 41.33, es-es: 46.13.*
        - The improvements are more substantial for XLM-Roberta-v2, particularly in de-de and es-es cases.

G-XLT (x-y):

- Training on multilingual data significantly enhances cross-lingual performance across all x-y pairs:
    - For mBERT-v2:
        - *en-de: 55.01, en-es: 52.66, de-en: 38.88, de-es: 37.56, es-en: 42.98, es-de: 42.00.*
        - Compared to v1, mBERT-v2 shows slight improvements in en-de and en-es and notable gains in de-es and es-de.
    - For XLM-Roberta-v2:
        - *en-de: 48.64, en-es: 49.84, de-en: 40.12, de-es: 36.26, es-en: 44.93, es-de: 39.02.*
        - XLM-Roberta-v2 outperforms its v1 counterpart across all x-y pairs, with particularly strong gains for non-English cross-lingual pairs like de-es and es-de.

Key Insight for v2 Models:

- The multilingual training in v2 models bridges the gap between English and non-English embeddings, leading to noticeable improvements in x-y settings, especially for non-English pairs (e.g., *de-es, es-de*).
- While en-x and x-en pairs (e.g., *en-de, de-en*) show incremental improvements, the most significant gains are seen in non-English pairs where v1 struggled.

Summary of Cross-Lingual Transfer for v1 and v2 Models

The v2 models demonstrate that multilingual training significantly enhances cross-lingual generalization for both x-x and x-y tasks, with the most dramatic gains in non-English cross-language pairs.

Pair Type	**v1 Performance (Limited Transfer)**	**v2 Performance (Improved Transfer)**
x-x (e.g., de-de)	Limited transfer for non-English x-x pairs (e.g., de-de, es-es). Performance is better in en-en due to training data.	Slight improvement in de-de and es-es, with better multilingual adaptation.
x-y (en-de)	Moderate transfer for en-x or x-en pairs, with English acting as a pivot.	Noticeable improvement for en-x, x-en, and non-English pairs like de-es and es-de.
x-y (de-es)	Very weak transfer for non-English x-y pairs (e.g., de-es, es-de).	Significant improvement for non-English x-y pairs.

Table 3: Summary of Cross-Lingual Transfer for Encoder only Models

5.2 Encoder-Decoder Models

`MBartForConditionalGeneration` and `MT5ForConditionalGeneration` using the hyperparameters **batch_size** = 16, **lr**=1e-4 and **epochs** = 3,**optimizer** = optim.AdamW(model.parameters().

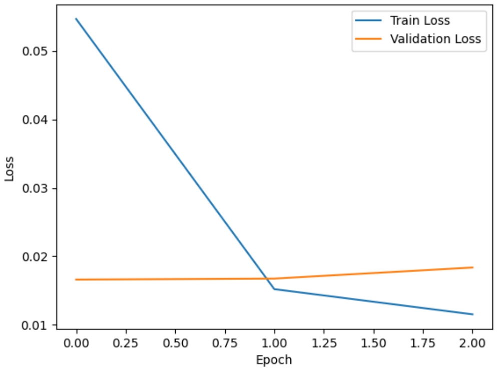


Figure 6: Bart on SQuAD

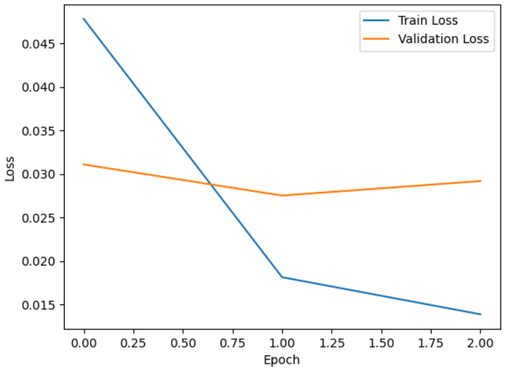


Figure 7: Bart on MT-SQuAD

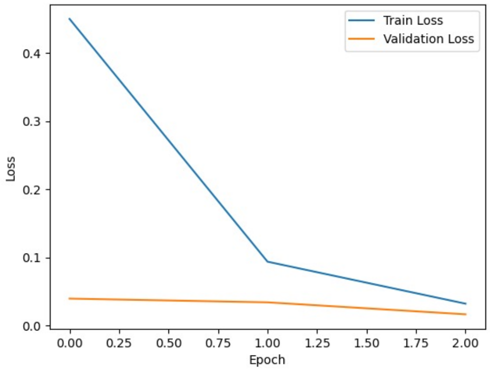


Figure 8: T5 on SQuAD

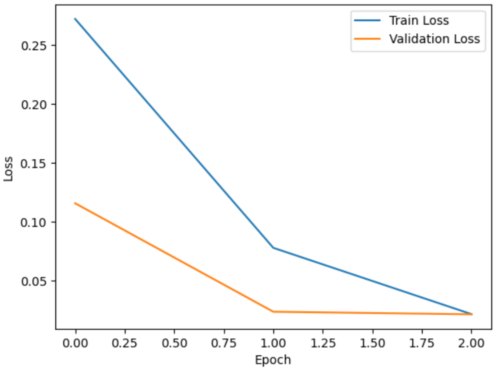


Figure 9: T5 on MT-SQuAD

	en-en	en-de	en-es	de-en	de-de	de-es	es-en	es-de	es-es
BART-v1	0.59	0.31	0.24	0.29	0.28	0.17	0.32	0.23	0.25
BART-v2	0.61	0.51	0.48	0.44	0.48	0.41	0.53	0.50	0.53
T5-v1	0.61	0.53	0.54	0.45	0.46	0.42	0.49	0.43	0.51
T5-v2	0.71	0.69	0.67	0.59	0.60	0.59	0.65	0.64	0.65

Table 4: Rouge-l

	en-en	en-de	en-es	de-en	de-de	de-es	es-en	es-de	es-es
BART-v1	0.87	0.78	0.74	0.78	0.76	0.68	0.78	0.74	0.71
BART-v2	0.88	0.85	0.84	0.82	0.83	0.81	0.85	0.84	0.84
T5-v1	0.88	0.85	0.85	0.82	0.82	0.81	0.84	0.81	0.84
T5-v2	0.91	0.90	0.89	0.86	0.86	0.86	0.88	0.88	0.88

Table 5: BERT Score

Cross-Lingual Transfer in BART

BART-v1 (Monolingual or Limited Multilingual Training):

XLT (x-x):

- **Scores:** en-en: 0.59, de-de: 0.28, es-es: 0.25.
- Performance is strongest for *en-en*, reflecting BART-v1's proficiency in English-based tasks.
- *de-de* and *es-es* scores are much lower (0.28 and 0.25), suggesting limited capacity for non-English generation due to lack of robust multilingual training.

G-XLT (x-y):

- **Scores:** en-de: 0.31, en-es: 0.24, de-en: 0.29, de-es: 0.17, es-en: 0.32, es-de: 0.23.
- Cross-lingual performance is weak overall, with the lowest scores for *de-es (0.17)* and *es-de (0.23)*, highlighting BART-v1's difficulty in reasoning between non-English languages.
- English-centric pairs (e.g., *en-de: 0.31, de-en: 0.29*) perform better, indicating a bias toward English-based alignment.

BART-v2 (Multilingual Training):

XLT (x-x):

- **Scores:** en-en: 0.61, de-de: 0.48, es-es: 0.53.
- Multilingual training boosts x-x performance across all languages, with *de-de (0.48)* and *es-es (0.53)* showing significant improvements over BART-v1.

G-XLT (x-y):

- **Scores:** en-de: 0.51, en-es: 0.48, de-en: 0.44, de-es: 0.41, es-en: 0.53, es-de: 0.50.
- Cross-lingual performance improves markedly compared to BART-v1:
  - English-to-non-English (e.g., *en-de: 0.51, en-es: 0.48*) sees significant gains, reflecting stronger alignment with non-English languages.
  - Non-English pairs (*de-es: 0.41, es-de: 0.50*) show improved alignment compared to BART-v1 but still lag behind English-centric pairs.

Cross-Lingual Transfer in T5

T5-v1 :

XLT (x-x):

- **Scores:** en-en: 0.61, de-de: 0.46, es-es: 0.51.
- Stronger x-x performance compared to BART-v1, particularly in *de-de (0.46)* and *es-es (0.51)*, indicating better multilingual capabilities.

G-XLT (x-y):

- **Scores:** en-de: 0.53, en-es: 0.54, de-en: 0.45, de-es: 0.42, es-en: 0.49, es-de: 0.43.
- T5-v1 outperforms BART-v2 in most x-y pairs:
  - English-centric pairs (*en-de: 0.53, en-es: 0.54*) show strong performance.
  - Non-English x-y pairs (*de-es: 0.42, es-de: 0.43*) exhibit moderate scores, still weaker than English-centric scenarios.

T5-v2:

XLT (x-x):

- **Scores:** en-en: 0.71, de-de: 0.60, es-es: 0.65.
- Substantial improvement across all x-x scenarios compared to T5-v1 and BART-v2.

G-XLT (x-y):

- **Scores:** en-de: 0.69, en-es: 0.67, de-en: 0.59, de-es: 0.59, es-en: 0.65, es-de: 0.64.
- T5-v2 achieves the best cross-lingual performance among all models:
  - Non-English x-y pairs (*de-es: 0.59, es-de: 0.64*) show significant gains, surpassing all other models.

Summary of Cross-Lingual Transfer for Encoder-Decoder Models

Scenario	**BART-v1 (Limited Transfer)**	**BART-v2 (Improved Transfer)**	**T5-v1 (Strong Transfer)**	**T5-v2 (Superior Transfer)**
x-x (e.g., de-de)	Weak for non-English pairs (**de-de: 0.28**).	Moderate improvement (**de-de: 0.48**).	Strong x-x performance (**de-de: 0.46**).	Exceptional x-x performance (**de-de: 0.60**).
x-y (e.g., en-de)	Weak cross-lingual performance (**en-de: 0.31, de-es: 0.17**).	Moderate improvement (**en-de: 0.51, de-es: 0.41**).	Strong transfer in English-centric and non-English pairs.	Best cross-lingual performance (**en-de: 0.69, de-es: 0.59**).

Table 6: Summary of Cross-Lingual Transfer for Encoder-Decoder Models

5.3 Decoder only Models

`GPT2LMHeadModel` if fine-tuned using the hyperparameters **batch_size** = 16, **lr**=1e-4 and **epochs** = 8,**optimizer** = optim.AdamW(model.parameters().

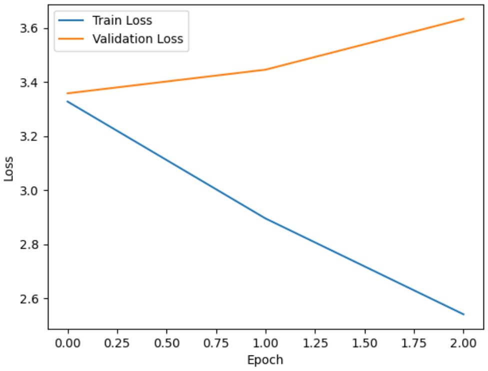


Figure 10: GPT2 on SQuAD

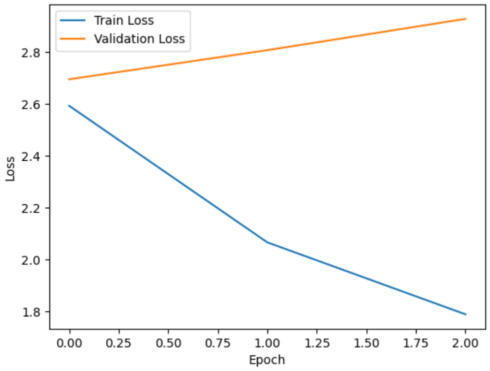


Figure 11: GPT2 on MT-SQuAD

	en-en	en-de	en-es	de-en	de-de	de-es	es-en	es-de	es-es
GPT2-v1	0.12	0.07	0.07	0.09	0.09	0.08	0.10	0.10	0.11
GPT2-v2	0.11	0.10	0.10	0.12	0.12	0.11	0.15	0.14	0.16

Table 7: Rouge-l

	en-en	en-de	en-es	de-en	de-de	de-es	es-en	es-de	es-es
GPT2-v1	0.62	0.60	0.60	0.56	0.56	0.56	0.59	0.62	0.62
GPT2-v2	0.63	0.63	0.63	0.60	0.60	0.60	0.65	0.65	0.66

Table 8: BERT Score

Cross-Lingual Transfer in GPT2-v1

XLT (x-x):

Scores:

- en-en: 0.12, de-de: 0.09, es-es: 0.11.

Performance is strongest for *en-en*, reflecting GPT2-v1's alignment with its training language (English). Non-English x-x scores (*de-de: 0.09, es-es: 0.11*) are notably weaker, suggesting that GPT2-v1 struggles to generate high-quality answers in non-English languages without multilingual training.

G-XLT (x-y):

Scores:

- en-de: 0.07, en-es: 0.07, de-en: 0.09, de-es: 0.08, es-en: 0.10, es-de: 0.10.

GPT2-v1 shows very limited ability to handle cross-lingual contexts:

- For *en-de* and *en-es* (context in English, question in German or Spanish, and answer in English), the scores (0.07 each) indicate poor generation quality when context and question are in different languages.
- *de-en* and *es-en* (context in non-English, question in English, and answer in English) perform slightly better (0.09 and 0.10), likely due to the model's proficiency in English-based generation.
- Non-English x-y pairs (*de-es, es-de*) show weak transfer, with scores between 0.08 and 0.10. This reflects the model's limited ability to handle cross-lingual reasoning without English involvement.

Key Insight for GPT2-v1:

Cross-lingual transfer is minimal, especially for non-English contexts and non-English questions. The model struggles most in scenarios where context and questions are in different languages.

Cross-Lingual Transfer in GPT2-v2

GPT2-v2, trained on multilingual data (en, de, es), shows substantial improvements across x-x and x-y scenarios.

XLT (x-x):

Scores:

- en-en: 0.11, de-de: 0.12, es-es: 0.16.

While *en-en* performance is slightly lower compared to GPT2-v1 (0.11 vs. 0.12), *de-de* and *es-es* scores improve (0.12 and 0.16, respectively). The highest score is achieved in *es-es* (0.16), suggesting that Spanish embeddings benefited most from multilingual training.

G-XLT (x-y):

Scores:

- en-de: 0.10, en-es: 0.10, de-en: 0.12, de-es: 0.11, es-en: 0.15, es-de: 0.14.

GPT2-v2 demonstrates marked improvements over GPT2-v1 in cross-lingual settings:

- For *en-de* and *en-es*, scores rise to 0.10 each. This suggests that the multilingual training allows GPT2-v2 to generate better answers in English despite non-English questions or contexts.
- For *de-en* and *es-en*, scores improve to 0.12 and 0.15, reflecting strong generation when context is in non-English and questions/answers are in English.
- For non-English cross-lingual pairs (*de-es, es-de*), scores rise to 0.11 and 0.14, showcasing better multilingual alignment and capacity to bridge non-English languages.

Key Insight for GPT2-v2:

Multilingual training significantly enhances cross-lingual transfer across all x-y pairs, particularly improving English generation when prompted with non-English contexts and questions (e.g., *de-en, es-en*). Non-English x-y pairs (e.g., *de-es, es-de*) see notable gains, indicating improved multilingual embeddings and cross-lingual alignment.

Summary of Cross-Lingual Transfer for GPT2 Models

Scenario	**GPT2-v1 Performance (Minimal Transfer)**	**GPT2-v2 Performance (Improved Transfer)**
x-x (e.g., de-de)	Limited performance for non-English monolingual tasks. Weak embeddings result in *de-de: 0.09, es-es: 0.11.*	Significant improvement for non-English x-x pairs (*de-de: 0.12, es-es: 0.16*).
x-y (e.g., en-de)	Weak performance when context and question languages differ. Scores like *en-de: 0.07* highlight difficulties with cross-lingual reasoning.	Improved scores for x-y pairs, particularly for *de-en (0.12), es-en (0.15)*, reflecting stronger multilingual generalization.
Non-English x-y (e.g., de-es)	Minimal transfer between non-English languages (*de-es: 0.08*).	Noticeable gains in non-English x-y pairs (*de-es: 0.11, es-de: 0.14*).

Table 9: Summary of Cross-Lingual Transfer for GPT2 Model

6 Conclusion

6.1 Comparison Across Families

Aspect	Encoder-Only	Encoder-Decoder	Decoder-Only
Intra-Language (x-x)	Strong for English; moderate for others after multilingual training.	Best performance, balanced across languages.	Decent but weaker compared to others, especially for non-English x-x pairs.
Cross-Language (x-y)	Moderate for English-centric tasks; weak for non-English pairs.	Superior transfer across all x-y pairs, especially in advanced multilingual models.	Limited transfer, struggles with non-English x-y tasks.
Multilingual Alignment	Relies on pretraining data; improves with multilingual fine-tuning.	Strong alignment across languages, especially in T5-v2.	Limited alignment, even with multilingual training.
Strengths	Efficient for retrieval tasks; good for monolingual QA.	Best for generation tasks requiring cross-lingual reasoning.	Excels in single-language generation tasks.
Weaknesses	Struggles with complex generation tasks.	Computationally expensive.	Limited cross-lingual capabilities.

6.2 Overall Analysis

- **Encoder-Decoder models** dominate cross-lingual transfer tasks, particularly T5-v2, which achieves robust performance across all x-x and x-y scenarios. Their bidirectional encoding and decoding architecture allows for effective multilingual alignment, making them ideal for tasks that require understanding and generating across languages.
- **Encoder-only models** perform well for monolingual tasks (x-x) and English-centric cross-lingual tasks (x-y). While their cross-lingual capabilities improve with multilingual training, they lag behind encoder-decoder models in generation tasks.
- **Decoder-only models** are least effective for cross-lingual transfer, with limited performance improvements even after multilingual training. They are better suited for single-language tasks where the focus is on generating coherent, fluent text.

7 Links

- Link to Models : Here
- Fine-Tuning T5 on SQuaD
- Fine-Tuning GPT2
- Fine-Tuning BERT

References

[1] Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering, 2020.

[2] Bryan Li and Chris Callison-Burch. Paxqa: Generating cross-lingual question answering examples at training scale, 2023.

[3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.