ECOLE CENTRALE CASABLANCA

Data Challenge : MSA -> Darija Translation

KHAZNAOUI Mouad

BELACHKAR Badr.Eddine

Supervised by
Prof. Michalis Vazirgiannis
Dr. Guokan Shang

March 14, 2024

Abstract

This report presents our work on data augmentation and machine translation using the Hugging Face Transformers library. We used the pre-trained model "moussaKam/arabart" to translate from Darija (Moroccan dialect) to Modern Standard Arabic (MSA). Additionally, we augmented our dataset using MADAR, adding over 10,000 additional examples, for a total of 16,000 training examples. This report details the steps of data preprocessing, model construction, and evaluation of results.

1 Introduction

In this project, we worked on the automatic translation from Darija to Modern Standard Arabic (MSA). Darija is a widely used Moroccan dialect, while Modern Standard Arabic is used in formal and academic contexts. Our goal was to build an efficient machine translation model to facilitate communication between these linguistic variants.

2 Data Preparation and Augmentation

The dataset utilized in this study originates from the MADAR (Multi-Arabic Dialect Applications and Resources) project, a collaborative effort aimed at enhancing Arabic dialect processing. The dataset comprises multiple TSV files containing Arabic sentences from various dialects, including Moroccan Arabic (Rabat) and Modern Standard Arabic (MSA). Due to its scarcity and high quality, the MADAR dataset was chosen as the primary source for this study [1, 2].

2.1 Translation and Enrichment

We leveraged the Google Translate API to translate Arabic sentences into English. The translation process was integrated into the data augmentation pipeline, allowing for the generation of parallel English translations for each Arabic sentence. This enriched dataset facilitates cross-lingual analysis and enables the training of multilingual models.

```
from deep_translator import GoogleTranslator

# Function for translating text using Google Translate API
def translate_text(text, dest_language='en'):
    translated_text = GoogleTranslator(source='ar', target=dest_language
    ).translate(text)
    return translated_text
```

Listing 1: Translation and Enrichment

3 Quality Assessment

To assess the quality of the augmented dataset, several metrics were evaluated, including the presence of empty lines and the consistency of translations (for random samples).

	idx	English	Darija	MSA
0	0	It's at the end of the hall. I will bring you ...	...كاين في الآخر ديال هاد القاعة. انجيب ليك شويا	...إنها في أخر القاعة . سوف آتي لك ببعض منها الآن
1	2	Do you make modifications?	واش كا دير التعديلات؟	هل تقومون بعمل تعديلات ؟
2	4	We want a table by the window.	بغينا ناخدو طابلة حدا الشرجم.	. نريد مائدة بجانب النافذة
3	5	There, right in front of the tourist data.	راه هنا، مقابل مكتب استعلامات السياح بالضبط.	. هناك ، أمام بيانات السائح تماما
4	9	I've never heard of this address near here.	ما عمري سمعت هاد العنوان هنايا.	. لم اسمع بهذا العنوان من قبل بالقرب من هنا
5	11	Continue down this road until you find a pharm...	سير نيشان حتا تشوف صيدلية.	. استمر في السير في هذا الطريق حتى تجد صيدلية
6	12	What is the latest color this season?	شنو هو أجدد لون هاد الموسم؟	. ما هو أحدث لون هذا الموسم
7	14	In my case, it's often for work and rarely for...	في الحالة ديالي، راه غالبا على الخدمة، قليل لل...	...في حالتي ، غالباً ما يكون من أجل العمل ونادراً
8	15	I will stay for two days.	غادي نبقا يومين.	. سأقيم لمدة يومين
9	16	I want close waves in my hair	بغيت بيرمانونت مجهد.	أريد تموج متقارب بشعري

Figure 1: Augmented Dataset Examples


```
{
    "idx": "0",
    "English": "It's at the end of the hall. I will bring you some of them now. If you want anything else just let me know.",
    "Darija": "كاين في الآخر ديال هاد القاعة. انجيب ليك شويا دابا. و إلا حتاجيتي شي حاجا اخرى، قولها ليا.",
    "MSA": "إنها في أخر القاعة . سوف آتي لك ببعض منها الآن . إذا أردت أي شيئاً آخر فقط أعلمني."
},
```


Figure 2: JSON Data

4 Model Construction

To construct our machine translation model, we leveraged the Hugging Face Transformers library, a state-of-the-art toolkit for natural language processing tasks [3]. Our objective was to develop a robust model capable of translating Darija, a dialect of Arabic, into Modern Standard Arabic (MSA).

4.1 Fine-tuning Process

We initiated the fine-tuning process by exposing the pre-trained "moussaKam/arabart" model [4] to our augmented dataset. This dataset comprised over 16,000 examples, including 10,000 instances obtained through the MADAR project augmentation and an additional 6,000 examples provided within the project scope. The fine-tuning process aimed to adapt the model's parameters to better capture the nuances of Darija and optimize its performance for translating into MSA.

4.2 Model Specifications

The fine-tuned model boasts a sophisticated architecture, featuring multi-layered neural networks with attention mechanisms. These mechanisms enable the model to selectively focus on relevant parts of the input text during translation, facilitating the generation of accurate and contextually appropriate translations. Additionally, the model's transformer-based architecture allows for efficient processing of sequential data, making it well-suited for translation tasks.

4.3 Training Execution

The training process was meticulously executed, with careful consideration given to optimizing various training parameters. Training parameters, including batch size (16),

number of epochs (10), and evaluation strategy (BLEU), were configured to strike a balance between model convergence and computational efficiency. The training data were processed using accelerated computation on GPUs, facilitating faster convergence and improved training efficiency.

5 Results Evaluation

We evaluated our model's performance using the BLEU (Bilingual Evaluation Understudy) score. The BLEU score measures the similarity between automatic translations and human reference translations. We achieved a BLEU score of 29 on the training dataset, and 12.9 on the test dataset, indicating that our model produces good-quality translations but may be subject to further development.

Here is also some figures of a report generated by tensorboard :

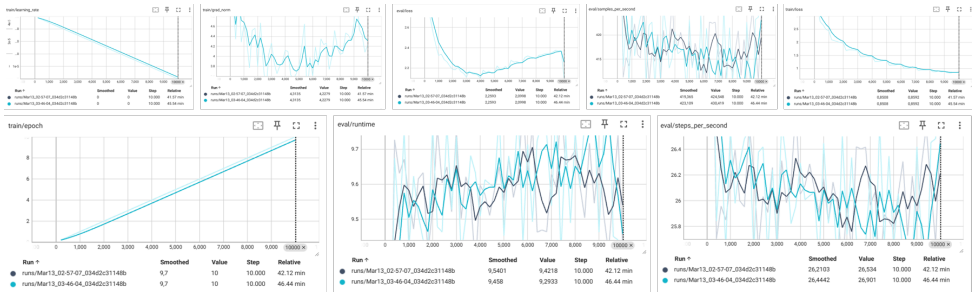

Figure 3: Elaborate view on the training process

6 Discussion and conclusions

The obtained BLEU scores serve as quantitative measures to evaluate the quality of our model's translations. A BLEU score of 29 on the training dataset indicates that the model performs well in capturing the nuances of the training data and generating translations that closely match the human reference translations. This suggests that our model has effectively learned from the training examples and can produce high-quality translations within the context of the training data.

On the other hand, the BLEU score of 12 on the test dataset suggests that while our model generalizes reasonably well to unseen data, there is room for improvement in its performance on out-of-domain or unseen examples. This indicates that the model may struggle with certain linguistic variations or contexts not adequately represented in the training data. Therefore, further development and refinement of the model are warranted to enhance its robustness and adaptability to diverse translation scenarios.

It's also important to note that one of the primary obstacles encountered in this project was the limitation of GPU resources. Due to the constraints of available computational resources, we were unable to explore larger model architectures or conduct extensive hyperparameter tuning, which may have hindered our model's ability to achieve even higher levels of translation quality.

Appendix

All code scripts or prompts are included in the zipped file accompanying this report. To access the MADAR dataset, please visit `https://camel.abudhabi.nyu.edu/madar-parallel-corpus/`.

The dataset comprises over 15 Arabic dialects from various Arab cities, each identified by unique IDs.

Our submission is structured as follows:

- **data folder**: Contains both the original provided data (corrected) and augmented data.
- **code folder**: Includes:
  - **Data preprocessing notebook**: Named 'Preprocessing Data Darija MSA', it encompasses all data training scripts, conveniently located within the dataset folder.
  - **Model construction and training notebook**: Named 'Model Training', alongside evaluation and visualization components.

In addition, a comprehensive PDF report is provided for detailed examination.

Here is also the new link utilized in the notebook for the new training data : `https://drive.google.com/uc?export=download&id=1lmNyT4YpShQ0fFmmzoSJgt12hLJKIJZ8`

References

1. Bouamor, Houda, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann and Kemal Oflazer. The MADAR Arabic Dialect Corpus and Lexicon. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 2018

2. The MADAR Shared Task on Arabic Fine-Grained Dialect Identification. Houda Bouamor, Sabit Hassan and Nizar Habash. Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 2019.

3. HuggingFace's Transformers: State-of-the-art Natural Language Processing 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, Alexander M. Rush. 2019

4. Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, Michalis Vazirgiannis. AraBART: a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization . arXiv, 2022.