Cross-Lingual Domain Classification of Task-Oriented Dialog (EN-FR)
[SI 630 Final Project Report]

Anusha Pathuri
anupath@umich.edu

Abstract

Accurate identification of the domain of user commands is a crucial first step in digital voice assistant systems. Consider domain identification as a multi-class classification problem where the inputs are transcribed utterances. Supervised learning is limited by the availability of annotated datasets, an issue that is exacerbated for non-English languages. In this project, I tackle the problem of domain classification for French, a *relatively* lower resource language than English. With access to a parallel annotated dataset in French, I set out to compare the performance gap between fully-supervised training in the target language and cross-lingual zero-shot transfer from the source language using massively pre-trained Transformer-based masked language models (PLMs). The results reinforce that given the recent advancements in PLMs, domain identification is a somewhat trivial task with fully-supervised fine-tuning in the target language achieving near-perfect results (0.98 F1) and even zero-shot transfer from the source not lagging far behind (0.95 F1). The full potential of these models in such systems is better realized through more fine-grained applications such as precise intent detection from voice commands.

1 Introduction

The modern day ubiquity of digital voice assistants calls for systems that effectively understand and respond to user requests. Building systems that can do so for a wide range of languages is a step towards inclusive technology that is accessible to all, including non-native English speakers.

The ultimate goal of this project is to build a text classification model that accurately predicts the domain of utterances in French for task-oriented dialogue systems. Accurate domain prediction would enable the system to route requests to the correct functionality, leading to efficient interactions and a seamless user experience. This task will be accomplished by leveraging readily available annotated English data. In general, building models that can adapt to other languages would help overcome the issue of data scarcity that hinders development of applications for low-resource languages.

Domain	Full		Final		
	EN	FR	Train	Val	Test
alarm	2,006	1,581	1,113	138	330
calling	3,129	2,797	1,968	278	551
event	1,249	1,050	799	92	159
messaging	1,682	1,239	861	125	253
music	1,929	1,499	1,082	137	280
news	1,762	905	668	78	159
people	1,768	1,392	975	145	272
recipes	1,845	1,002	697	124	181
reminder	3,058	2,321	1,656	203	462
timer	1,488	1,013	734	89	190
weather	2,372	1,785	1,261	168	356
Total	22,288	16,584	11,814	1,577	3,193

Table 1: Number of utterances in the MTOP dataset (Li et al., 2021) which uses a 70:10:20 train/val/test split.

2 Data

MTOP (Li et al., 2021) is a multi-lingual, multi-domain task-oriented dialog dataset consisting of synthetic utterances in 6 languages across 11 domains with fine-grained intent labels. Different from the authors' work, this project will focus on the more coarse-grained domain prediction. The high-resource source language is English and the relatively low-resource target language is French. As described by Li et al., the English texts were human translated to French and low quality annotations were discarded. In this project, only utterances that have a French translation are included in the dataset (shown under "Final" in Table 1). Some sample utterances and their domain labels[1] are shown in Table 2.

[1] https://huggingface.co/datasets/mteb/mtop_domain

English Text	French Text	Label	Class
Which of my friends went to Louisiana Tech?	Qui parmi mes amis est allé à la Louisiana Tech ?	people	7
delete reminder to pick up milk	supprime le rappel d'aller chercher le lait	reminder	8
Give me some chicken recipes	Donnez-moi des recettes de poulet	recipes	9
Call Teresa please	Veuillez appeler Teresa	calling	1
Temperature tomorrow morning around Dublin Fah...	Température demain matin aux environs de Dubli...	weather	5

Table 2: Sample utterances from MTOP (Li et al., 2021) for the domain prediction task.

3 Related Work

A major challenge in the field of cross-lingual natural language understanding (XLU) is the scarcity of annotated datasets. Eriguchi et al. (2018) leverage representations learned by a multilingual neural machine translation system and a task-specific classifier to achieve competitive results on benchmarks in a zero-shot approach to classification (inference on a new language that was never seen during training). Conneau et al. (2018) highlight the challenges of translation-based approaches[2] with TRANSLATE-TRAIN requiring a new classifier for every target language and TRANSLATE-TEST being computationally-intensive at test time. Gerz et al. (2021) tackle intent detection by combining machine translation with state-of-the-art multilingual sentence encoders and report an improvement over the zero-shot baseline through few-shot learning (fine-tuning on a small subset of in-domain training data). Razumovskaia et al. (2022) observe that multilingual task-oriented dialogue systems are best modeled using zero- or few-shot cross-lingual transfer from EN by either machine translation or multilingual representations. Artetxe et al. (2023) show that TRANSLATE-TEST with a strong machine translation system can outperform both TRANSLATE-TRAIN and zero-shot approaches in most cross-lingual classification tasks.

4 Methods

4.1 Multi-Layer Perceptron (MLP)

A simple MLP with one hidden layer consisting of 512 hidden units will be trained on the EN train set and evaluated on the machine translated FR test set (translated using NLLB-200 (Koishekenov et al., 2023)). This is the TRANSLATE-TEST approach in literature (see **MLP (EN, TT)** in Table 5). A Bag-of-Words (BoW) representation with word counts is used as input to the model.

4.2 Zero-Shot Prompting

GPT-3.5[3] is prompted with instructions to classify FR sentences into one of the 12 MTOP domains. This is a zero-shot prompting approach (see **GPT-3.5 Prompt** in Table 5).

Sample prompt: *"Your task is to determine what category a phrase is related to. Both the phrase and the category will be in French. Possible categories: la messagerie, l'appel, l'événement, la minuterie, la musique, le temps, l'alarme, les personnes, le rappel, les recettes, les nouvelles. Phrase: Où est-ce que l'aide fédérale américaine sera-t-elle envoyée ? Your answer (a single category from the list of possible categories): "*

4.3 Masked Language Models (MLMs)

Following the general approach in recent literature for many NLP tasks, I will use Transformer-based masked language models (MLMs) that have been trained on massive amounts of unlabeled data to learn representations and fine-tune them on the downstream task of multi-class classification. The pre-trained language models (PLMs) I will explore can be grouped into two categories:

- **Monolingual PLMs**: Trained on a single language. Consider RoBERTa (Zhuang et al., 2021) for English and CamemBERT (Martin et al., 2020) for French (Note: French is an outlier; most non-English languages do not have a specialized monolingual model).
- **Multilingual PLMs**: Trained on multiple languages together. Consider XLM-R (Conneau et al., 2020) trained on 100 languages including English and French.

As discussed in Section 3, we often do not have access to annotated data in the target language. In order to compare cross-lingual zero-shot transfer

[2] TRANSLATE-TRAIN: EN data is machine translated to the target language and the model is trained on this. TRANSLATE-TEST: model is trained on EN data, inference is on target language data machine translated to EN.

[3] https://platform.openai.com/docs/models/gpt-3-5-turbo

from English to French with monolingual supervision in French, two experimental settings are proposed. The approaches described below are adapted from Li et al. (2021) and Adelani et al. (2024).

4.3.1 Only source training data

Fine-tune on the EN train set and evaluate on the FR test set (zero-shot transfer).

1. Monolingual PLM (EN) + Monolingual fine-tuning (EN). Evaluated on the machine translated FR test set (see **RoBERTa (EN, TT)**).
2. Multilingual PLM + Monolingual fine-tuning (EN). Evaluated both on the machine translated FR test set (see **XLM-R (EN, TT)**) and directly on the FR test set (see **XLM-R (EN)**: *cross-lingual transfer*).

4.3.2 Only target training data

Fine-tune on the FR train set and evaluate on the FR test set.

1. Monolingual PLM (FR) + Monolingual fine-tuning (FR) (see **CamemBERT (FR)**).
2. Multilingual PLM + Monolingual fine-tuning (FR) (see **XLM-R (FR)**).

All the experiments will be conducted using HuggingFace's Transformers library (Wolf et al., 2020). All MLMs will be fine-tuned for 20 epochs using the AdamW optimizer with an initial learning rate of 2e-5.

5 Evaluation and Results

Metrics

- **Classification**: F1 score (micro - gives equal importance to each sample, macro - gives equal importance to each class, weighted - macro weighted by the proportion of each class' support in the dataset). Micro precision and recall.
- **Machine Translation**: BLEU score[4] and BERTScore[5] to evaluate the quality of translations in the TRANSLATE-TEST (TT) approach.

Baselines

1. **Most Frequent**: Always predict "calling", the class with the largest number of instances in the training set.

Label	Keyword_EN	Keyword_FR
alarm	alarm	alarme
calling	call	appel
event	event	événement
messaging	message	message
music	play	joue
news	news	nouvelles
people	who	qui
recipes	recipe	recette
reminder	remind	rappel
timer	timer	minuterie
weather	weather	temps

Table 3: Keywords for baseline 3

2. **Stratified**: Randomly sample from a prior defined by the training set class distribution shown in Figure 1.
3. **Keywords**: Naive rule-based classifier. Define a list of keywords per class based on the most frequently occurring words in the training corpus (Table 3). Check for the presence of a keyword in the query e.g. for "what's the weather like?", predict "weather". If keywords corresponding to multiple classes are present in the query, rare classes (less frequent in the training set) are given more priority. If none of the keywords corresponding to any class is present, predict "unknown".

The F1 scores on the EN and FR train and val sets for the three baselines are reported in Table 4. Not surprisingly, B3 is the best baseline. Figures 2 and 3 give us a rough idea of the complexity of different classes for the text classification task in each language. Inter-class misclassifications (e.g. "alarm" predicted as "reminder") make up less than 1% of the errors. The vast majority of errors seem to be texts classified as "unknown". It is reasonable to infer that higher the fraction of texts predicted as "unknown", more complex is that class. In general, classes "alarm", "reminder", and "timer" achieve high accuracy and seem to be relatively straightforward. For most classes, the naive method does much better for English than French, the difference being particularly large for classes "alarm", "timer", "news" and "music".

Results

The training loss and validation micro F1 score for the experiments in 4.1 and 4.3 are shown in Figure 4. Results of domain classification for the FR test

[4] https://huggingface.co/spaces/evaluate-metric/sacrebleu
[5] https://huggingface.co/spaces/evaluate-metric/bertscore

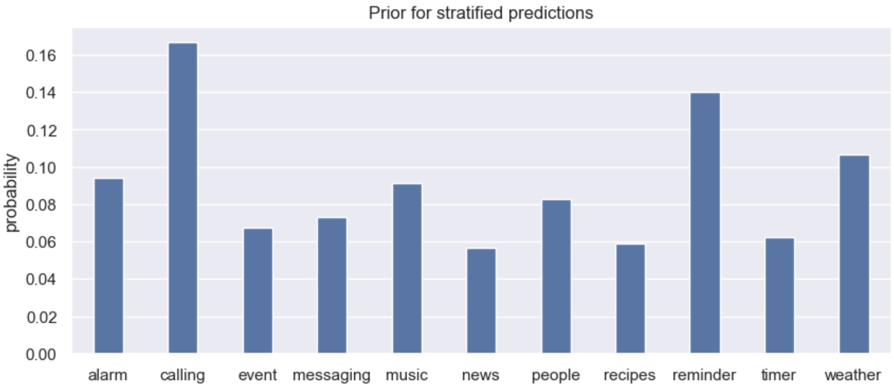

Figure 1: Prior for baseline 2

		Train			Val		
		Macro F1	Micro F1	Weighted F1	Macro F1	Micro F1	Weighted F1
Model	**Lang**						
B1: MostFreq	EN	0.03	0.17	0.05	0.03	0.18	0.05
	FR	0.03	0.17	0.05	0.03	0.18	0.05
B2: Stratified	EN	0.09	0.10	0.10	0.08	0.09	0.09
	FR	0.09	0.10	0.10	0.08	0.09	0.09
B3: Keywords	EN	0.65	0.63	0.74	0.64	0.62	0.72
	FR	0.56	0.54	0.65	0.56	0.55	0.65

Table 4: Baseline Results on EN & FR Train & Val Sets

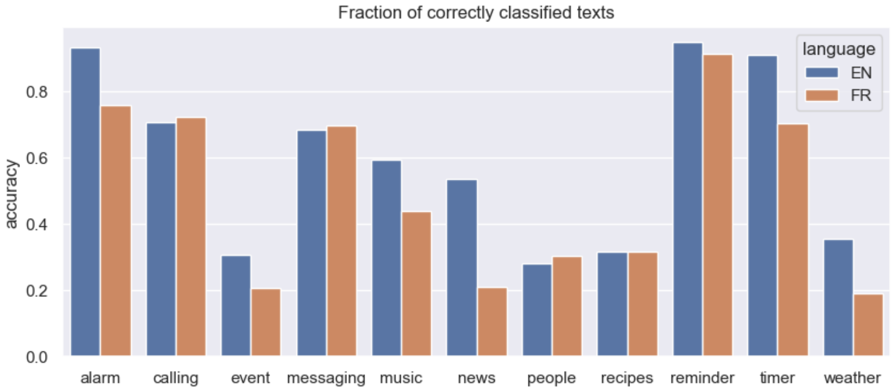

Figure 2: Baseline 3: Class-wise accuracy (train + val)

set are shown in 5.

- The BERTScore for the FR-to-EN translated test set is 0.92, indicating that the quality of machine translation is not a cause of concern in the TRANSLATE-TEST experiments.
- In the absence of an extensive prompt engineering exercise, GPT-3.5 underperforms in the zero-shot prompt setting which highlights the need for instruction refinement.
- The simple BoW MLP achieves an F1 score of over 0.92 in the zero-shot TRANSLATE-TEST setting, speaking to the relatively low complexity of the domain classification task.
- Not surprisingly, supervised fine-tuning on the FR train set achieves an almost perfect F1-score with both the monolingual and multilingual models being comparable. It is reasonable to consider this as the upper bound of accuracy for this task.

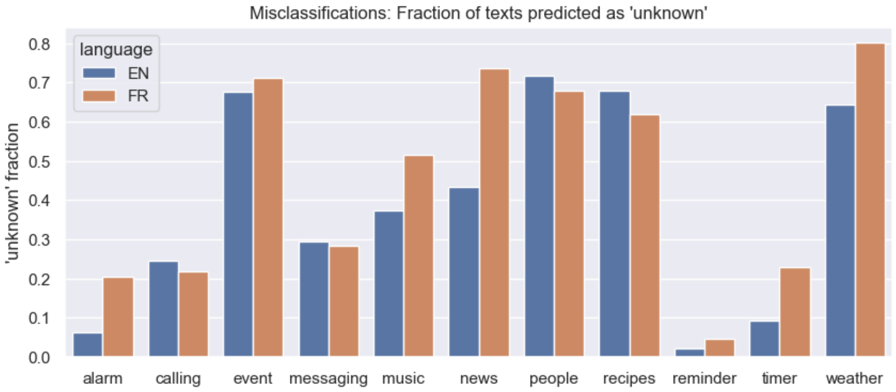

Figure 3: Baseline 3: Instances misclassified as "unknown" (train + val)

- In the TRANSLATE-TEST setting, the multilingual model achieves good zero-shot cross-lingual transfer (0.95 F1) and the monolingual model achieves comparable zero-shot transfer (0.948 F1).
- It is interesting to compare the two evaluation results of XLM-R fine-tuned on the EN train set. Having seen in-domain EN data, the multilingual model does better on the FR-to-EN machine translated test set (0.95 F1) than on the original FR test set (0.919 F1).

6 Discussion

The results tell a simple story: the domain classification task considered here is not a very complex problem. The utterances in the dataset are short (not more than 40 tokens) and do not use complicated language. A naive rule-based classifier implementing keyword matching already does quite well on some of the more trivial classes (e.g. alarm, reminder). It's not just the massively pre-trained MLMs that achieve near-perfect results overall; even the shallow MLP trained using a Bag-of-Words representation of the source language inputs beats the best baseline by a large margin and generalizes well on unseen data translated from the target language to the source. In many applications that run in real-time such as voice assistants, the speed-accuracy trade-off is a major factor governing development efforts. In such scenarios, the smaller MLP model (with a lightweight machine translation engine) would be preferable to a massive Transformer-based language model. With some basic feature engineering and hyperparameter tuning, the MLP could be further improved.

7 Conclusion

In this project, I explored the problem of multi-class classification for task-oriented dialog. Cross-lingual transfer from a high resource source language such as English to lower resource target languages that suffer from data scarcity can help accelerate application development for the latter. I compared the cross-lingual transfer performance to fully-supervised fine-tuning in the target language as I had access to a parallel dataset in the target language. The results show that this performance gap is not very large, so in the absence of a parallel dataset in the target language, zero-shot cross-lingual transfer is a viable option, provided the quality of machine translation is controlled for. In general, for state-of-the-art massively pre-trained language models, the domain classification task is quite trivial and they easily achieve near-perfect generalization without any hyperparameter tuning.

8 Other Things I Tried

At first, I had considered trying the TRANSLATE-TRAIN approach in which the entire EN train set is machine translated to FR and a model is fine-tuned on that. But with the near-perfect results achieved by some of the approaches shown in Table 5, I did not feel the need to proceed with this experiment.

9 What I Would Have Done Differently or Next

Firstly, I would have chosen the more challenging problem of intent detection as opposed to domain classification. I would have also liked to try a simple attention based classifier similar to what we had implemented in Homework 2.

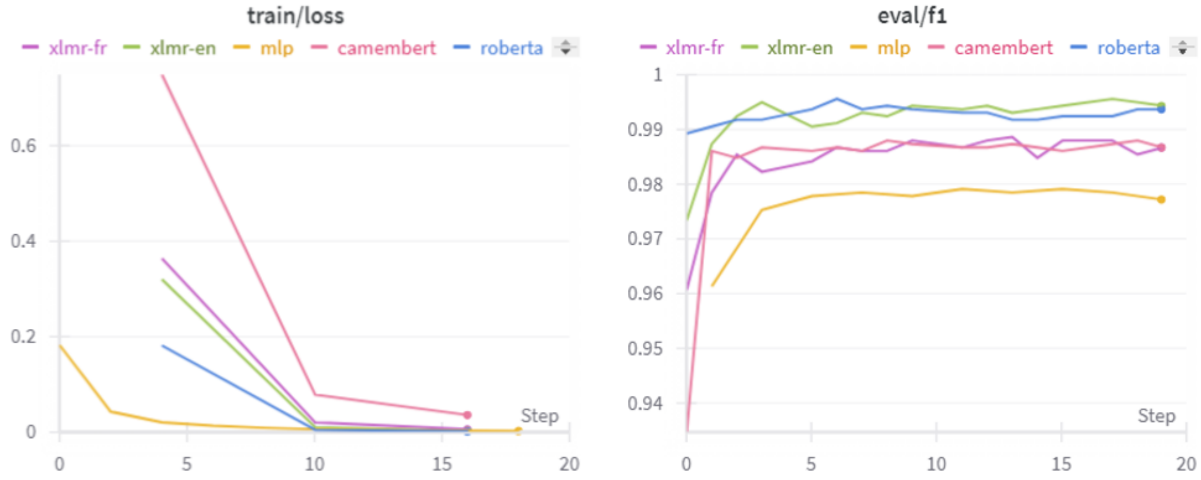

Figure 4: Supervised Fine-tuning (MLP, MLMs)

Model	F1	Precision	Recall	Macro F1	Weighted F1
B1: MostFreq	0.173	0.173	0.173	0.027	0.051
B2: Stratified	0.111	0.111	0.111	0.097	0.112
B3: Keywords	0.563	0.563	0.563	0.560	0.669
MLP (EN, TT)	0.928	0.928	0.928	0.930	0.928
GPT-3.5 Prompting	0.780	0.780	0.780	0.752	0.828
RoBERTa (EN, TT)	0.948	0.948	0.948	0.953	0.947
XLM-R (EN, TT)	0.950	0.950	0.950	0.955	0.949
XLM-R (EN)	0.919	0.919	0.919	0.914	0.918
CamemBERT (FR)	0.982	0.982	0.982	0.980	0.982
XLM-R (FR)	0.982	0.982	0.982	0.980	0.982

Table 5: Results on MTOP Domain FR Test Set

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7468–7475, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of* *the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Evgeniia Razumovskaia, Goran Glavas, Olga Majewska, Edoardo M. Ponti, Anna Korhonen, and Ivan Vulic. 2022. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *Journal of Artificial Intelligence Research*, 74:1351–1402.

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

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.