Comparing MultiLingual and Multiple MonoLingual Models for Intent Classification and Slot Filling

Cedric Lothritz[1 [0000−0002−5372−7970]], Kevin Allix[1 [0000−0003−3221−7266]], Bertrand Lebichot[1 [0000−0003−2188−0118]], Lisa Veiber[1 [0000−0002−3692−8308]], Tegawendé F. Bissyandé[1 [0000−0001−7270−9869]], and Jacques Klein[1 [0000−0003−4052−475X]]

University of Luxembourg, 6 rue Richard Coudenhove-Kalergi, 1359 Luxembourg

Abstract. With the momentum of conversational AI for enhancing client-to-business interactions, chatbots are sought in various domains, including FinTech where they can automatically handle requests for opening/-closing bank accounts or issuing/terminating credit cards. Since they are expected to replace emails and phone calls, chatbots must be capable to deal with diversities of client populations. In this work, we focus on the variety of languages, in particular in multilingual countries. Specifically, we investigate the strategies for training deep learning models of chatbots with multilingual data. We perform experiments for the specific tasks of Intent Classification and Slot Filling in financial domain chatbots and assess the performance of mBERT multilingual model vs multiple monolingual models.

Keywords: Chatbots · Multilingualism · Intent Classification · Slot Filling.

1 Introduction

Chatbots usually operate in a single language depending on where they are deployed (e.g., a chatbot for a British bank will only handle requests written in English). While deploying a single monolingual chatbot is usually sufficient in countries where the entire population speaks one language, this strategy presents challenges in multilingual areas where people do not necessarily speak the same language at a high level. In multilingual countries, such as Switzerland, Luxembourg, India, South Africa, etc. with two or more national languages, companies and banks need to be able to communicate with their clients in the language of the latter's choosing in order to stay competitive. The same holds true for client support chatbots, which have to support multiple languages to stay viable in a multilingual environment. This requirement presents a challenge as companies have to decide on a strategy for implementing a multilingual chatbot system. Two such strategies are as follows: (S1) For n languages, employ n chatbots, each of which is trained to handle requests in a single language. (S2) For n languages, employ one chatbot which is trained using data written in n languages. There are some immediate advantages for training a chatbot using mixed-language data as one would have to train only a single chatbot and maintain only one database as opposed to multiple. However, it is unclear how the performance of a singular

multilingual chatbot (S2) compares to a combination of multiple monolingual chatbots (S1). In this paper, we explore these two strategies for chatbots in a multilingual environment. Specifically, we investigate the performance of S1 and S2 on two tasks that represent fundamental blocks for chatbot systems: Intent Classification (IC), which is the task of identifying a user's intent based on a piece of text, and Slot Filling (SF), the task of identifying attributes that are relevant to a given intent. For this study, we use the Rasa chatbot framework, which uses the Dual Intent and Entity Transformer Classifier [2] for both the IC and SF tasks. Furthermore, we compare two techniques for text representation, namely bag-of-words (BOW) and multilingual BERT (mBERT) [5].
We aim to answer the following research questions:

- RQ1: How does the distribution of data samples per language influence the performance of multilingual chatbots?
- RQ2: How do S1 and S2 compare in terms of Intent Classification and Slot Filling?

For this study, we use a novel dataset for IC and SF in the financial domain, which we name the *Banking Client Support* (BCS) dataset. We also use the MultiATIS++ dataset published by Xu et al. [11].

This paper is structured as follows: In Section 2, we explain the datasets we use, the chatbot framework, and give a detailed description for S1 and S2. In section 3, we present the results of our experiments, answer the research questions, and show the merits of multilingual chatbots. Section 4 shows various papers related to this study, and we finally conclude our findings in Section 5.

2 Experimental Setup

2.1 Datasets

For this study, we use two multilingual datasets to evaluate the performance of multilingual chatbots. We created one dataset for client support bots in the banking domain as there are no public datasets available to the best of our knowledge. We also use a multilingual version of the well-known ATIS dataset to verify the results using a larger dataset.

Banking Client Support Dataset: The first dataset (which we refer to as banking client support dataset (BCS) throughout this paper) is based on a toy dataset provided by Rasa[1]. The original dataset contains 337 samples divided into 15 intents. We removed three of the intents together with 93 samples as they seemed too vague (*inform*) or were not directly related to the banking domain (*help&human_handoff*), and added 763 samples and introduced 16 new intents, resulting in 1003 samples across 28 intents with each intent being distributed quite equally. The intents cover basic conversational phrases such as *greet* or *affirm* and requests specific to the banking domain such as *make_bank_transfer*, *block_card* or *search_atm*. Additionally, the set contains 253 entities, divided into

[1] https://github.com/RasaHQ/financial-demo

6 unique entity types such as *account_type* or *credit_card_type*. We then translated the dataset into three languages (German, French and Luxembourgish) with Google Translate and manually corrected translation errors, resulting in a total of four distinct, but parallel datasets[2]. For this study, we use these four base datasets to construct mixed-language datasets containing equal numbers of samples from the base datasets, e.g., the English-French dataset consists of 50% English samples and 50% French data samples. There are 11 possible language combinations: six combinations with two languages, four with three languages, and one combination with all four languages, which gives us a total of 15 different datasets containing varying numbers of languages.

MultiATIS++ Dataset: The second dataset is based on the popular Airline Travel Information System (ATIS) dataset [4]. The original dataset contains a total of 5871 sentences divided into 26 intents. Furthermore, it contains 19 356 samples for slot filling, divided into 128 slot types. MultiATIS++ is a multilingual version of ATIS created by Upadhyay et al. [8] and Xu et al. [11]. For this study, we use the English, German and French versions of the MultiATIS++ dataset. Furthermore, we reduced the number of intents by removing intents with fewer than five samples, resulting in a total of 5860 sentences divided into 17 intents. It is to note that the distribution of the intents is highly imbalanced with 73.6% of the samples having the intent *atis_flight*. There are four possible language combinations, resulting in a total of seven datasets.

2.2 Chatbot framework Used in this Study

Rasa: Bocklisch et al. introduced the Rasa NLU and Rasa Core tools [2], with the objective of making a framework that is more accessible for creating conversational software. The modular design of a chatbot made with Rasa allows to swap out configuration files and training data. For this study, we created two different configurations: (C1) a bag-of-words (BOW) pipeline consisting of a WhitespaceTokenizer, RegexFeaturizer, LexicalSyntacticFeaturizer, and a CountVectorsFeaturizer. (C2) an mBERT pipeline which consists of the HFTransformersNLP model initializer using the cased multilingual BERT-base as its pretrained model as well as its accompanying tokenizer and featurizer[3].

mBERT: For this study, we will use the multilingual BERT [5] (mBERT) model as our datasets contain texts written in English, French, German, and Luxembourgish. However, as the number of Wikipedia articles varies greatly for every language of mBERT, there are significant disparities between the datasets used to train the different language components. Specifically, the English dataset is the largest with around 6 million articles, the German and French datasets have comparable sizes with 2.5 and 2.2 million articles respectively, and the Luxembourgish dataset is the smallest with only 59 000 articles.

For this study, we use the cased mBERT model with 12 transformer blocks, 768 hidden layers, attentions heads and 110 trainable parameters provided by Devlin et al.[4] [5].

[2] Available at https://github.com/Trustworthy-Software/BCS-dataset

[3] Further information on Rasa models: https://rasa.com/docs/rasa/components/

[4] https://github.com/google-research/bert/blob/master/multilingual.md

2.3 Implementation Strategies

S1: Pseudo-multilingual Chatbots For each monolingual dataset, we train two chatbots: one using an mBERT model, and one without. By combining a language-selector (LS) and monolingual chatbots, we can create pseudo - multilingual chatbots. This allows us to directly compare the performance between monolingual chatbots and multilingual chatbots. For the LS, we use langid[5].
S2: Multilingual Chatbots Based on the monolingual datasets, we construct mixed-language datasets. For every language combination, we extract a stratified subset from each monolingual dataset and combine them to create multilingual datasets. For each of these new datasets, we train two multilingual chatbots, one using a BOW model, and one using an mBERT model.

3 Experimental Results

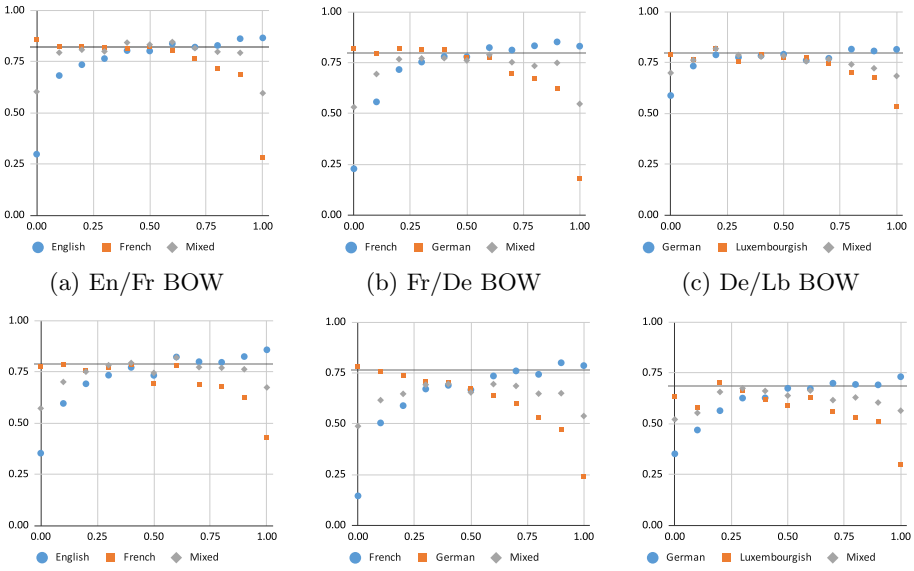
In this section, we will answer the two research questions that we formulated for this study (cf. Section 1) as well as discuss the results in Section 3.5.

3.1 RQ1: How does the distribution of data samples per language influence the performance of multilingual chatbots?

In order to answer this question, we create chatbots trained on bilingual datasets, vary the distribution of both languages in the sets, and evaluate their performance on various test sets. Specifically, we train 11 chatbot models on 11 mixed-language datasets where dataset 0 contains 0% samples from language A and 100% samples of language B, dataset 1 contains 10% samples of language A, 90% samples of language B, etc. These models are tested on three test sets: (1) a monolingual test set containing samples from language A, (2) a test set containing samples from language B, (3) a stratified test set containing an equal number of samples from both languages A and B.

Intent Classification Fig. 1 shows the performances of three language combinations in terms of F1 score. These combinations are: English/French (En/Fr), French/German (Fr/De) being two languages that are very dissimilar in terms of syntax and vocabulary, and German/Luxembourgish (De/Lb) being syntactically very similar. When varying the distribution of per-language data samples, we can make several observations: (1) when tested on a monolingual test set, we tend to observe very low performances if the training set does not contain the tested language at all, while we can see very high performances for the opposite case. This performance drop is less apparent for the De/Lb combinations (cf Fig. 1c and Fig. 1f). Furthermore, the Fr/De combinations (cf Fig. 1b and Fig. 1e) show the highest performance drop for these extreme cases. (2) When testing on the mixed-language test set, we can observe comparable performances for every training set, except for the models that were trained on monolingual training sets. (3) Models that are trained on sets containing 50% samples from each language tend to perform similarly for each test set. When performing the same experiment on the MultiATIS++ dataset, we observed that the performance remained stable except for the models trained on monolingual data.

[5] https://github.com/saffsd/langid.py



(a) En/Fr BOW (b) Fr/De BOW (c) De/Lb BOW

(d) En/Fr mBERT (e) Fr/De mBERT (f) De/Lb mBERT

Fig. 1: Evolution of the F1 score for bilingual chatbots for IC task when varying the distribution of data samples per language. The horizontal line represents the performance of the LS+monolingual chatbots models.

Slot Filling For the task of slot filling, we can make similar observations as we did for IC: very high and low performances for chatbots that were trained on monolingual datasets, with less noticeable drops for the German/Luxembourgish language combinations. When tested on the mixed test sets, most models perform similarly well except for the monolingual ones. It is to note that this performance drop is smaller for the SF task than it is for the IC task.

When performing the same experiment on the MultiATIS++ dataset, the performance of the models fluctuated only slightly except for the models trained on monolingual data.

RQ1 Answer: There is a noticeable drop in performance if a language is absent from the training set. A 50/50 split in the training set tends to lead to the highest performances on the mixed-language test sets.

3.2 RQ2: How do S1 and S2 compare in terms of Intent Classification and Slot Filling?

In order to answer this question, we reuse the bilingual chatbot models that were trained on the datasets which contain 50% data samples from each language (S2) and compare their performance to pseudo-bilingual chatbots (S1).

Table 1 compares F1 scores for pseudo-bilingual chatbot models and bilingual chatbot models for the IC task. Our results show that the combination of a

language selector and two monolingual chatbots yields higher performances with regard to every performance measure used. It is to note that the English/French variant is an exception to the rule as the model with the S2 strategy significantly outperforms the S1 model. This trend can be observed for both the chatbot models with an mBERT and the ones with a BOW model. The performance differences between S1 and S2 models with mBERT are usually larger when compared to the performance differences between the models that do not use pretrained models. Furthermore, the models based on BOW consistently outperform the models with mBERT by several percentage points.

Table 2 shows the results of the same task on the MultiATIS++ datasets. In contrast to the BCS sets, the results are in favour of the S2 strategy. When comparing the MCC scores, we observe that the performance of the bilingual models either exceeds or matches that of the combinations of LS+monolingual chatbots.

	BOW						mBERT					
	Bilingual			LS + Monolingual			Bilingual			LS + Monolingual		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
En/Fr	0.851	0.835	**0.833**	0.864	0.805	0.823	0.779	0.753	0.745	0.830	0.771	**0.791**
En/De	0.810	0.801	0.797	0.867	0.835	**0.843**	0.744	0.708	0.706	0.796	0.766	**0.769**
En/Lb	0.807	0.797	0.794	0.845	0.810	**0.819**	0.712	0.679	0.676	0.747	0.697	**0.703**
Fr/De	0.787	0.764	0.761	0.835	0.788	**0.796**	0.691	0.664	0.654	0.800	0.753	**0.763**
Fr/Lb	0.805	0.780	0.778	0.824	0.777	**0.787**	0.703	0.677	0.662	0.728	0.674	**0.679**
De/Lb	0.794	0.788	0.783	0.826	0.784	**0.797**	0.668	0.640	0.638	0.725	0.678	**0.683**

Table 1: Test results for bilingual chatbots (S2) vs monolingual chatbots with language selector (S1) on Intent Classification task on BCS set.

	BOW								mBERT							
	Bilingual				LS + Monolingual				Bilingual				LS + Monolingual			
	Prec	Rec	F1	MCC	Prec	Rec	F1	MCC	Prec	Rec	F1	MCC	Prec	Rec	F1	MCC
En/Fr	0.973	0.967	0.969	**0.929**	0.976	0.961	0.967	0.914	0.971	0.970	0.97	**0.941**	0.979	0.968	0.973	0.929
En/De	0.977	0.974	0.975	**0.942**	0.930	0.966	0.968	0.924	0.973	0.972	0.972	**0.937**	0.978	0.972	0.974	**0.937**
Fr/De	0.964	0.959	0.961	0.911	0.971	0.962	0.966	**0.916**	0.974	0.97	0.971	**0.933**	0.974	0.965	0.968	0.922

Table 2: Test results for bilingual chatbots(S2) vs monolingual chatbots with language selector(S1) on Intent Classification task on MultiATIS++ set

In order to determine if pseudo-bilingual (S1) significantly outperform bilingual (S2) models, we perform a Wilcoxon test for both strategies over every dataset used. We find that the differences in performance for mBERT models are indeed significant, but in the case for BOW models, only the difference in precision is clearly significant.

For the SF task. We generally see better results for the mBERT model. Similarly to the IC task, the combination of monolingual chatbots and a language selector almost consistently outperforms the chatbots trained on bilingual datasets by a large margin. This is true for both the BCS and the MultiATIS++ datasets. We once again determine statistical significance of the obtained results through a Wilcoxon test. The resulting p-values show that the performance differences are significant except for recall and F1 score for the BOW models.

RQ2 Answer: In most cases, S1 performs better than S2, with IC on MultiATIS++ being a notable exception.

3.3 Discussion

When using a small dataset, the results of the conducted experiments are generally in favour of strategy S1 and by a significant margin. This is true for both the IC and the SF tasks. The results are less conclusive when training the chatbots on the larger MultiATIS++ dataset. For the IC task, neither strategy is consistently outperforming the other. On the other hand, strategy S1 is superior regardless of the dataset as it outperforms S2 for the BCS dataset as well as the MultiATIS++ dataset. The performances of the investigated models were significantly dependent on the task. While BOW-models generally performs better for the IC task, mBERT-models seems to be the favourable choice for the SF task, as strategy S1 with mBERT generally largely outperformed the BOW-models when compared directly.

4 Related Work

Multilingual IC and SF: Previous multilingual text classification systems are usually based on two different approaches: (1) machine translation systems that translate training data into the target language [10] or (2) parallel corpora that are used to learn embeddings jointly from multiple languages [6]. Such crosslingual embeddings prove useful for binary classification tasks such as sentiment classification [12,13] and churn intent detection [1]. Abbet et al. [1] use multilingual embeddings for the task of churn intent detection in social media. They show that bilingual embeddings trained on an English and German dataset outperform monolingual embeddings for this binary IC task. Furthermore, they show that models trained on social media data can be applied to chatbot conversations as well. Schuster et al. [7] evaluate three methods for multilingual IC and SF, namely translating the training data into the target language, using pretrained crosslingual embeddings, and using a novel pretrained translation encoder to generate embeddings.

Multilingual Datasets: One major challenge for multilingual IC and SF is the lack of textual data in languages other than English. Schuster et al. created a dataset containing 57 000 utterances divided into three languages [7]: 43 000 utterances in English, 8600 in Spanish and 5000 in Thai. Their data is annotated for 12 intent types, and 11 slot types in total. They use their dataset to evaluate various crosslingual transfer methods for IC and SF. The ATIS dataset [4] is one of the most popular datasets for IC and SF. Originally available only in English, it was partially translated into Hindi and Turkish [9], creating MultiATIS. Xu et al. further extended MultiATIS to six more languages [11], resulting in MultiATIS++, consisting of nine versions of the original ATIS dataset. Datasets related to banking are difficult to find as most of them are proprietary [3], making our BCS dataset one of the few public datasets related to that domain.

5 Conclusion

In this paper, we presented a study on multilingual chatbots, specifically on the Intent Classification and Slot Filling tasks.

We compared two implementation strategies and two embedding techniques. We noticed that training a chatbot on mixed-language data decreases the overall

performance. We concluded that, in the case of two languages, the combination of a language selector and two monolingual chatbots (S1) usually outperforms chatbots that are directly trained on bilingual datasets (S2). While the BOW models almost consistently outperform the mBERT models in the Intent Classification tasks, the mBERT models usually perform better in the Slot Filling tasks when using the S1 strategy.

References

1. Abbet, C., M'hamdi, M., Giannakopoulos, A., West, R., Hossmann, A., Baeriswyl, M., Musat, C.: Churn intent detection in multilingual chatbot conversations and social media. arXiv preprint arXiv:1808.08432 (2018)
2. Bocklisch, T., Faulkner, J., Pawlowski, N., Nichol, A.: Rasa: Open source language understanding and dialogue management. arXiv preprint arXiv:1712.05181 (2017)
3. Costello, C., Lin, R., Mruthyunjaya, V., Bolla, B., Jankowski, C.: Multi-layer ensembling techniques for multilingual intent classification. arXiv preprint arXiv:1806.07914 (2018)
4. Dahl, D.A., Bates, M., Brown, M.K., Fisher, W.M., Hunicke-Smith, K., Pallett, D.S., Pao, C., Rudnicky, A., Shriberg, E.: Expanding the scope of the atis task: The atis-3 corpus. In: HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference of the North American Chapter of the ACL: Human Language Technologies (2019)
6. Lauly, S., Larochelle, H., Khapra, M.M., Ravindran, B., Raykar, V., Saha, A., et al.: An autoencoder approach to learning bilingual word representations. arXiv preprint arXiv:1402.1454 (2014)
7. Schuster, S., Gupta, S., Shah, R., Lewis, M.: Cross-lingual transfer learning for multilingual task oriented dialog. 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). pp. 3795–3805 (2019)
8. Upadhyay, S., Faruqui, M., Tür, G., Dilek, H., Heck, L.: (almost) zero-shot cross-lingual spoken language understanding. In: 2018 IEEE ICASSP. pp. 6034–6038 (2018). https://doi.org/10.1109/ICASSP.2018.8461905
9. Upadhyay, S., Faruqui, M., Tür, G., Dilek, H.T., Heck, L.: (almost) zero-shot cross-lingual spoken language understanding. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6034–6038. IEEE (2018)
10. Wan, X.: Co-training for cross-lingual sentiment classification. In: Joint Conference of the 47th Annual Meeting of the ACL. pp. 235–243 (2009)
11. Xu, W., Haider, B., Mansour, S.: End-to-end slot alignment and recognition for cross-lingual NLU. In: Proceedings of EMNLP 2020. pp. 5052–5063. ACL, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.410
12. Zhou, G., He, T., Zhao, J.: Bridging the language gap: Learning distributed semantics for cross-lingual sentiment classification. In: International Conference on Natural Language Processing and Chinese Computing. pp. 138–149. Springer (2014)
13. Zhou, H., Chen, L., Shi, F., Huang, D.: Learning bilingual sentiment word embeddings for cross-language sentiment classification. In: 53rd Annual Meeting of the ACL and the 7th International Joint Conference on NLP. pp. 430–440 (2015)