Exploiting In-Domain Bilingual Corpora for Zero-Shot Transfer Learning in NLU of Intra-Sentential Code-Switching Chatbot Interactions

Maia Aguirre, Manex Serras, Laura García-Sardiña, Jacobo López-Fernández
Ariane Méndez and **Arantza del Pozo**
Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Parque Científico y Tecnológico de Gipuzkoa, Paseo Mikeletegi 57, Donostia / San Sebastián (Spain)
{magirre, mserras, lgarcias, jlopez, amendez, adelpozo}@vicomtech.org

Abstract

Code-switching (CS) is a very common phenomenon in regions with various co-existing languages. Since CS is such a frequent habit in informal communications, both spoken and written, it also arises naturally in Human-Machine Interactions. Therefore, in order for natural language understanding (NLU) not to be degraded, CS must be taken into account when developing chatbots. The co-existence of multiple languages in a single NLU model has become feasible with multilingual language representation models such as mBERT. In this paper, the efficacy of zero-shot cross-lingual transfer learning with mBERT for NLU is evaluated on a Basque-Spanish CS chatbot corpus, comparing the performance of NLU models trained using in-domain chatbot utterances in Basque and/or Spanish without CS. The results obtained indicate that training joint multi-intent classification and entity recognition models on both languages simultaneously achieves best performance, better capturing the CS patterns.

1 Introduction

Multilingual speakers outnumber monolingual speakers in the world (Tucker, 2001). In regions with various coexisting languages, a common feature of natural interactions amongst speakers is the continuous casual switching between the concerned languages or "codes". This phenomenon, known as Code-Switching (CS), is very frequent in both spoken and written informal interactions (Ahn et al., 2020). In fact, the percentage of CS in social networks ranges from 14.5% to 49.06% for the corpora explored in Gambäck and Das (2016). Moreover, in the study Al-Qaysi and Al-Emran that explores the educators and learners' attitudes towards using CS online, 86.40% of the students and 81% of the teachers claim to actively code-switch while chatting on social networks.

As is to be expected, CS also arises spontaneously during human-machine interactions such as conversing with chatbots (i.e., conversational agents). This being the case, Bawa et al. (2020) reveal that these interlocutors strongly prefer chatbots that do understand CS.

The most common strategy employed by chatbots to understand the interlocutor is to employ intent and entity-based annotation schemata. This involves intent –communicative purpose– detection and entity –key words– classification processes (Tur et al., 2010). For example, in the utterance "I want to have an Italian meal that does not exceed 15 euros", the *intent* behind the user's query is to *order food* given the *entity* labels *cheap* and *italian*.

Allowing multilingual interactions with chatbots involves running a language identifier prior to each speaker turn and executing language-specific Natural Language Understanding (NLU) models. However, this approach is only effective for inter-sentential CS, where the code alternation happens at utterance boundaries. In the case of intra-sentential CS, where the same utterance contains words or phrases belonging to two or more languages (Gumperz, 1982), detecting every intent and entity of the utterance poses a major challenge for existing algorithms (Banerjee et al., 2018). The previous sentence with intra-sentential CS would be "*Quiero* an Italian meal *que no supere los* 15 euros" (Spanish in italics). In this case, neither an NLU model in English nor an NLU model in Spanish would detect that the intent is *order food*.

One strategy to solve this problem would be treating CS as a language itself and training an NLU model with examples containing CS. However, collecting data with CS is quite challenging, as it hardly exists in written form and requires bilingual annotators. Instead, obtaining labeled monolingual data in multiple languages is easier and zero-shot cross-lingual transfer learning (TL) has been proved to perform well across different Natural Language Processing (NLP) tasks, including NLU (M'hamdi et al., 2021). Thus, this TL

Proceedings of EMNLP 2022 Industry Track, pages 148–154
December 9–11, 2020.

approach would be more practical to exploit in real-world industrial chatbot applications to be deployed in CS regions. Unfortunately, such methodology has not yet been explored in the literature, mainly due to the scarcity of real intra-sentential CS datasets.

In this paper, a Basque-Spanish CS corpus is exploited to evaluate different zero-shot cross-lingual TL experiments aiming to examine whether such multilingual training methodologies are capable of addressing the NLU problem of intra-sentential CS chatbot interactions. For this purpose, the effectiveness of three multilingual models is analysed in their ability to understand CS: one fine-tuned using a chatbot corpus in Basque, another one fine-tuned on a corpus of Spanish chatbot utterances and a third one fine-tuned using both corpora simultaneously. It is important to underline that none of the models was exposed to CS during training, and that the monolingual Basque and Spanish training corpora belong to the same domain of the CS corpus used for testing purposes. Through this comparison it has been determined that models fine-tuned on in-domain bilingual corpora simultaneously are able to generate cross-lingual bonds and perform better against CS. The fact that zero-shot fine-tuning of multilingual BERT models on both monolingual languages enhances their effectiveness in understanding CS stands as one of the main contributions of this work.

The remaining of the paper is structured as follows: Section 2 reviews recent work in the area of joint intent and entity detection and multilingual models. Section 3 analyzes the main characteristics of the corpora used both for training and evaluation purposes. Section 4 presents the architecture and specifications of the joint intent and entity detection implementation employed. Section 5 shows the results obtained in the different experiments carried out and, finally, Section 6 highlights the main conclusions and proposes tentative lines for future work.

2 Related Work

Intent Detection and Entity Classification

The traditional way of approaching intent detection and entity classification tasks for NLU is to address them separately. However, treating each task as an individual problem leads to inefficient usage of training resources. Among others, Chen et al. (2019); Lorenc (2021) have shown that combining intent and entity recognition in a single system achieves significant improvements in both tasks with lower computational resources. Cai et al. (2022) and Qin et al. (2020) propose novel methods that consider joint learning of both tasks by correlating the intents and entities and reach new state-of-the-art performance. In addition, Castellucci et al. (2019) have explored how these joint approaches also perform better in multilingual settings.

Multilingual models

Contextualised multilingual models, such as mBERT and XLM-R (Conneau and Lample, 2019), have achieved state-of-the-art results in monolingual and multilingual tasks on NLU benchmark tests (Wang et al., 2019; Hu et al., 2020; Liu et al., 2020). However, the effectiveness of NLU models on CS interactions remains unknown (Winata et al., 2021).

Still, there have been several attempts to use multilingual representations to encode CS sentences (Srinivasan, 2020; Aguilar et al., 2020; Khanuja et al., 2020), showing promising results and surpassing previously achieved performances (Aguilar et al., 2020; Khanuja et al., 2020).

Recent work has shown that, even if the embeddings across the 12 multi-head attention layers of mBERT are clustered across languages (Krishnan et al., 2021), they can be split into two components: a language-specific one and a language-neutral one (Krishnan et al., 2021; Libovický et al., 2020; Tanti et al., 2021). Pires et al. (2019) have also found that a shared subspace representing relevant linguistic information is common to cross-lingual BERT representations. Likewise, Chi et al. (2020) claim that part of the representation space of the syntactic level of mBERT is shared between languages and identify that mBERT has a cross-lingustic clustering of gramatical relations. In addition, Cao et al. (2019) suggest that mBERT also aligns semantics across languages. Libovický et al. (2020) use a set of semantic-oriented tasks to show that unsupervised multilingual contextual embeddings based on BERT capture similar semantic phenomena in very similar ways across languages.

Moreover, Krishnan et al. (2021) and Tanti et al. (2021) have also demonstrated that the cross-lingual capacity of mBERT models increases after fine-tuning as the models switch their ability to cluster embeddings by language to cluster them

according to the needs of the task. For example, regarding the intent detection task, the embeddings will be grouped by intents after fine-tuning.

On the other hand, several benchmarking experiments in NLP tasks other than NLU against CS test sets have shown that mBERT fine-tuning achieves the best performance compared to alternative multilingual models. Khanuja et al. (2020) presented the first model evaluation benchmark against CS. After testing various embedding techniques for all tasks and datasets, they concluded that the multilingual BERT model performs the best. They also demonstrate that, for most datasets, a modified version of mBERT that has been subsequently fine-tuned with synthetically generated CS data performs consistently better. Aguilar et al. (2020) propose another benchmark metric that combines ten corpora covering four different CS language pairs and four NLP tasks for the evaluation of linguistic CS. Superior performance of the mBERT models for each available language pair is observed across the vast majority of the tasks.

3 Data

3.1 Training and validation corpus

Three corpora have been used to train and validate the models:

1. A Basque corpus, consisting of utterances in the Basque language
2. A Spanish corpus, formed by utterances in Spanish language
3. A Bilingual corpus, grouping both corpora together

The Basque and Spanish corpora comprise a collection of text samples used to train the NLU modules of four bilingual (Basque-Spanish) chatbots. These chatbots were designed to answer specific questions related to the fields of administration, taxation, and transport. In addition, they were able to respond to greetings, requests for help, and some common social questions such as "Are you a robot?", etc. Besides their domain label, the examples in each corpus were annotated with semantic information regarding their intents and entity values. The preprocessing of the training data involved removing stress marks, capitalisation and punctuation marks, considering that users tend to write without respecting spelling rules while chatting.

The three training corpora are divided into a training set and a validation set, with a ratio of 75/25 and with an even distribution of intents and entities. The total number of unique entities is 39 and the number of unique intents is 90.

The partition size of each corpus in the training and validation sets is reflected in Table 1.

	Training	Validation
Basque corpus	1662	555
Spanish corpus	1452	485
Bilingual corpus	3114	1040

Table 1: Number of utterances associated with the training and validation set of each corpus.

3.2 Test corpus

In our experiments, The BaSCo –*Ba*sque-*S*panish *Co*de-Switching– corpus (Aguirre et al., 2022) is used to evaluate the robustness of the different NLU models against Basque-Spanish CS.

It is a compendium of 1377 utterances containing Basque-Spanish intra-sentential CS that belong to the same domain as the corpora used for training (i.e. chatbot interactions related to the fields of administration, taxation and transport) and also share the same set of intent and entity labels.

4 Implementation Strategies

The joint intent detection and entity classification NLU model developed in this work takes the one presented by Chen et al. (2019) and its corresponding implementation[1] as a baseline. For our experiments, the baseline model architecture has been adapted as shown in Figure 1 to support the detection of multiple intents in the same utterance.

This adaptation has involved two major changes. On the one hand, intent label representation has been adapted to allow assigning more than one intent to each utterance. For this purpose, *one-hot* [2] encoding has been adopted and consequently the activation function of the final layer has been changed from softmax to sigmoid. On the other hand, the loss function selected for multi-intent classification optimisation has been changed to *binary cross-entropy*, as it allows each utterance to have more than one associated intent (Ho and Wookey, 2019).

[1] https://github.com/90217/joint-intent-classification-and-slot-filling-based-on-BERT

[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html

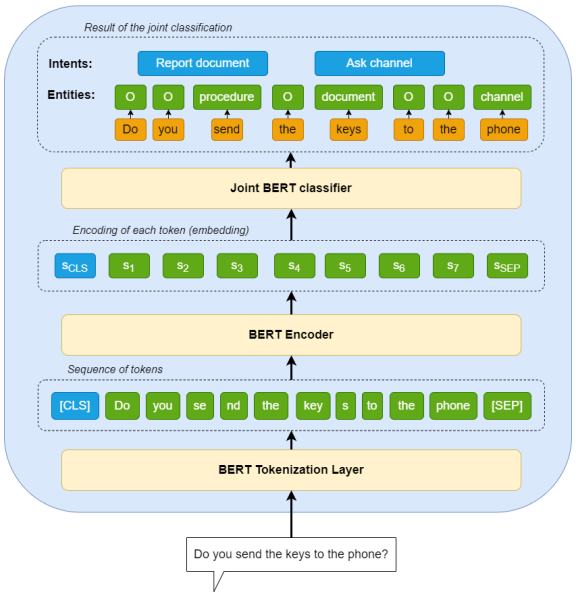


Figure 1: Architecture of the joint multi-intent and entity detection model. It consists of a first module that tokenises the input sentence and translates the information to the BERT model, which returns an embedding for each element of the sequence and a global embedding of the utterance. Entities are predicted using the individual embeddings of the sequence and intents are predicted using sentence embeddings.

This loss function predicts whether each possible intent appears in the utterance regardless of the rest of the intents. The loss function used to classify entities remains *sparse categorical cross-entropy*, since it meets the assumption that each token belongs to a single entity category. The goal of the implemented joint multi-intent and entity classifier is to minimise the sum of the two individual loss functions.

The BERT model employed is BERT-Base-Multilingual-Cased or mBERT, which has 12 layers, 768 hidden states and 12 heads. And the best performing hyperparameters are a maximum length and batch size of 128; the Adam optimiser (Kingma and Ba, 2015); a learning-rate of 9e-5; a dropout probability of 0.1; and 50 epochs.

5 Results

5.1 Validation results

Table 2 collects the results returned by the different models when evaluated over the Basque and Spanish validation sets. This table allows to directly compare the performance of the model trained on the bilingual corpus versus the performances of the models trained on the monolingual corpora.

The model trained with the bilingual corpus uses the same train/validation partition as the models trained on the monolingual corpora (i.e., the utterances used for training and evaluating the model are identical). Still, the results obtained with the model trained on the bilingual corpus outperform the results of the models trained on the monolingual corpora as more intents are properly classified. Therefore, it can be determined that, as expected, cross-lingual learning actually happens in the model trained with both languages.

This learning improvement occurs because in the fine-tuning process the model is trained with the same set of intent and entity labels for the Basque and Spanish corpora. In this way, it learns to relate and project text entries in different languages onto a common label space.

5.2 Test results

To better assess the cross-lingual learning capabilities of the trained models, the BaSCo corpus of intrasentential Basque-Spanish CS utterances is used as test set.

The results obtained for each model are shown in Table 3. As it can be seen, the cross-lingual comprehension acquired by the model trained on the Bilingual corpus is also evidenced against the CS test set, showing a clear improvement over the models trained on the monolingual corpora with results such as:

- +38 and +17 points in F1 micro and macro metrics respectively for intent classification.
- +2 and +15 points in F1 micro and macro metrics respectively for entity classification.

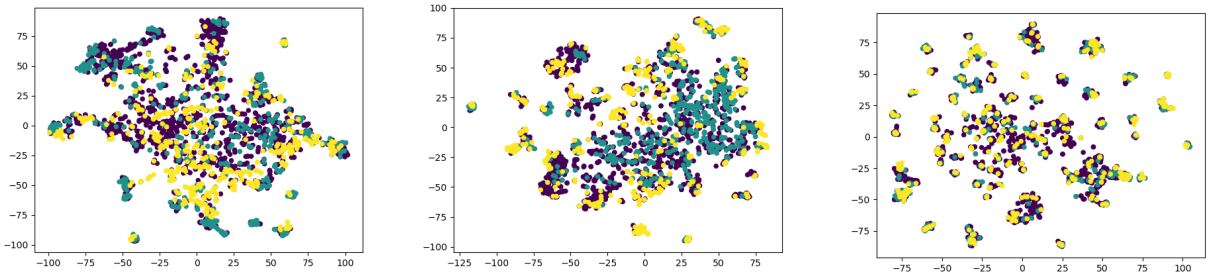
A more intuitive way of visualising the cross-lingual learning of the model trained on the Bilingual corpus is by means of a two-dimensional representation of the embeddings (i.e. the result given by mBERT's *pooling layer* for each sentence input to the model). For this purpose, the t-SNE method (t-distributed Stochastic Neighbor Embedding) is used to assign each high-dimensional data vector a position in a two-dimensional map (Van der Maaten and Hinton, 2008). In this way, Figure 2 shows the two-dimensional representation that each of the three models assigns to each of the following sets: the validation set of the Basque corpus, the validation set of the Spanish corpus and the BaSCo intra-sentential CS test set.

Metrics	Validation set	Models		
		Basque	**Spanish**	**Bilingual**
Intent classifier loss	Basque corpus	0.0292	/	**0.0241**
	Spanish corpus	/	0.0274	**0.0214**

Table 2: Loss metrics of the three multilingual models: (i) fine-tuned on the Basque corpus, (ii) fine-tuned on the Spanish corpus and (iii) fine-tuned on the Bilingual corpus, in their ability to understand multiple intents and entities. The results obtained at the end of training (epoch 50) are shown for the Basque and Spanish corpus validation sets separately.

Metrics		Model		
		Basque	**Spanish**	**Bilingual**
Multiple Intent Classification	Intent classifier loss	0.0594	0.0676	**0.0351**
	F1 Score micro	0.4339	0.3958	**0.8148**
	F1 Score macro	0.2889	0.3081	**0.4714**
Entity Classification	F1 Score micro	0.7162	0.6567	**0.7358**
	F1 Score macro	0.3593	0.4420	**0.5933**

Table 3: Loss and F1 metrics of the three multilingual models: (i) fine-tuned on the Basque corpus, (ii) fine-tuned on the Spanish corpus, and (iii) fine-tuned on the Bilingual corpus, in their ability to understand multiple intents and entities. The results obtained at the end of training (epoch 50) are shown when evaluated over the BaSCo test set.



(a) Model trained on the Basque corpus.

(b) Model trained on the Spanish corpus.

(c) Model trained on the Bilingual corpus.

Figure 2: t-SNE representation of the sentence embeddings of the validation set of the Basque corpus (green), the validation set of the Spanish corpus (yellow), and the BaSCo test set (purple) of the models trained on the Basque, Spanish, and Bilingual corpus.

As it can be appreciated, the models trained on the monolingual Basque and Spanish corpora (Figures 2a and 2b) present a major dispersion of the points in the plane, having greater difficulty in determining clear groupings. In contrast, in Figure 2c there is a very noticeable overlapping of dots of different colours (i.e. of different languages) yielding clearly delimited groupings. These clusters have a clear semantic charge because they concentrate sentences that share the same intent. This property can be easily ascertained in the interactive web version of the figure[3], where sentences with, for example, the intent label "greeting" are grouped around the same point on the map, regardless of whether they are in Spanish, Basque or Basque-Spanish CS.

These results are very positive, as they show that state-of-the-art NLU models have the ability to understand intra-sentential CS chatbot interactions when they are fine-tuned using in-domain monolingual corpora in both CS languages.

6 Conclusions and Future Work

The main contribution of this work is proving that fine-tuning a mBERT NLU model with in-domain bilingual data enables it to detect intents and en-

[3]https://clusters-mbert.dialogue.vicomtech.org

tities of intra-sentential CS chatbot interactions with industrial grade robustness. It is essential for the corpora employed in the fine-tuning process to share the same intent and entity labels; with this proviso, the model learns to relate and project text entries from the different languages onto a common label space. As a result, the model represents utterance embeddings of the same meaning but different language in the same area of the vector space. Hence, unlike the original mBERT model that groups embeddings into clusters depending on their language, the NLU model fine-tuned on the bilingual in-domain corpus of chatbot interactions happens to be language agnostic, classifying utterances by their meaning regardless of language. This property endows the model with the ability to correctly classify utterances with intra-sentential CS. Given the scarcity of annotated CS data, this outcome is very promising. Considering the results achieved, we strongly recommend exploiting in-domain bilingual corpora to fine-tune mBERT NLU models of real-world chatbot applications in CS regions. Such type of corpora can be easily compiled from the collection of annotated text samples used to train monolingual NLU models without CS, as it has been done in this work.

Overall, it can be stated that progress has been made towards building multilingual conversational assistants that incorporate CS strategies and that can therefore better understand multilingual users. The results obtained for CS between Basque and Spanish, languages belonging to different families, should in principle also be extrapolated to other language pairs.

A tentative line of future research would be to try to further improve performance using data augmentation techniques. To this end, methods that recombine the sentences of the Basque and Spanish corpora while maintaining their semantic labels could be explored. An alternative line of research would be to explore whether the performance of mBERT models fine-tuned on multiple languages sharing utterance labels improves proportionally to the number of languages they are trained with. This would require translating the training corpora to other languages and revising their labelling a posteriori.

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. Lince: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1803–1813.

Maia Aguirre, Laura García-Sardiña, Manex Serras, Ariane Méndez, and Jacobo López. 2022. Basco: An annotated basque-spanish code-switching corpus for natural language understanding. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3158–3163, Marseille, France. European Language Resources Association.

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan W Black. 2020. What code-switching strategies are effective in dialog systems? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 213–222.

Noor Al-Qaysi and Mostafa Al-Emran. Code-switching usage in social media: A case study from oman. *International Journal of Information Technology and Language Studies*, 1(1).

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3766–3780.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Fengyu Cai, Wanhao Zhou, Fei Mi, and Boi Faltings. 2022. Slim: Explicit slot-intent mapping with bert for joint multi-intent detection and slot filling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7607–7611. IEEE.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multi-lingual intent detection and slot filling in a joint bert-based model.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855.

John J Gumperz. 1982. *Discourse strategies*. Cambridge University Press.

Yaoshiang Ho and Samuel Wookey. 2019. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*, 8:4806–4813.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1663–1674.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Petr Lorenc. 2021. Joint model for intent and entity recognition.

Meryem M'hamdi, Doo Soon Kim, Franck Dernoncourt, Trung Bui, Xiang Ren, and Jonathan May. 2021. X-metra-ada: Cross-lingual meta-transfer learning adaptation to natural language understanding and question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3617–3632.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. Agif: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1807–1816.

Anirudh Srinivasan. 2020. Msr india at semeval-2020 task 9: Multilingual models can do code-mixing too. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 951–956.

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. On the language-specificity of multilingual bert and the impact of fine-tuning.

G Richard Tucker. 2001. A global perspective on bilingualism and bilingual education. *GEORGETOWN UNIVERSITY ROUND TABLE ON LANGUAGES AND LINGUISTICS 1999*, page 332.

Gokhan Tur, Dilek Hakkani-Tür, and Larry Heck. 2010. What is left to be understood in atis? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24. IEEE.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *NAACL 2021*, page 142.