Play música alegre: A Large-Scale Empirical Analysis of Cross-Lingual Phenomena in Voice Assistant Interactions

Donato Crisostomi
Alexa AI, Amazon
Sapienza, University of Rome
doncris@amazon.com

Alessandro Manzotti
Alexa AI, Amazon
manzotti@amazon.com

Enrico Palumbo
Alexa AI, Amazon
palumboe@amazon.com

Davide Bernardi
Alexa AI, Amazon
dvdbe@amazon.com

Sarah Campbell
Alexa AI, Amazon
srh@amazon.com

Shubham Garg
Alexa AI, Amazon
gargshu@amazon.com

Abstract

Cross-lingual phenomena are quite common in informal contexts like social media, where users are likely to mix their native language with English or other languages. However, few studies have focused so far on analyzing cross-lingual interactions in voice-assistant data, which present peculiar features in terms of sentence length, named entities, and use of spoken language. Also, little attention has been posed to European countries, where English is frequently used as a second language. In this paper, we present a large-scale empirical analysis of cross-lingual phenomena (code-mixing, linguistic borrowing, foreign named entities) in the interactions with Alexa in European countries. To do this, we first introduce a general, highly-scalable technique to generate synthetic mixed training data annotated with token-level language labels and we train two neural network models to predict them. We evaluate the models both on the synthetic dataset and on a real dataset of code-switched utterances, showing that the best performance is obtained by a character convolution based model. The results of the analysis highlight different behaviors between countries, having Italy with the highest ratio of cross-lingual utterances and Spain with a marked preference in keeping Spanish words. Our research, paired to the increase of the cross-lingual phenomena in time, motivates further research in developing multilingual Natural Language Understanding (NLU) models, which can naturally deal with cross-lingual interactions.

1 Introduction

The interaction of different languages produces a variety of linguistic phenomena, the most prominent examples being code-switching and lexical borrowing. Code-swiching (CS), or code-mixing[1], refers to the alternation of languages within an utterance or a conversation (Poplack, 2004), while linguistic borrowing occurs when a word is adopted from a language and integrated into another without translation. Examples of these are: (i) "Play música alegre" (ii) "Bravo, that was a great performance", with the former being a case of code-switching and the latter exhibiting lexical borrowing. These phenomena are particularly frequent in bilingual countries, where the local language, called frame-language, is influenced by a second language, which is instead called the mixing-language. This phenomenon is abstracted by the *Matrix Language Frame* model (Poulisse, 1998) in code-switching literature. Common pairs of frame-mixing languages are for example *Spanglish* (Spanish-English) and *Hinglish* (Hindi-English).

Countries for which these phenomena happen usually undergo a broader influence which also permeates their culture, as it happens for example with American artistic production of cinema and music. As a side effect, utterances originated in the frame language are rich in foreign named entities, which contribute to their linguistic heterogeneity. Voice assistants operating in these locales have to face a significant amount of foreign words while being in most cases trained on monolingual corpora, hence posing a severe threat to their performance.

Indeed, the growing interest in multi-lingual models (Devlin et al., 2019; Alexis and Lample, 2019; Conneau et al., 2020) and datasets (FitzGerald et al., 2022; Xu et al., 2020) may help mitigate the problem. We will use in the rest of the paper

[1]We will use the terms code-switching and code-mixing interchangeably, despite they are sometimes used in linguistic literature to denote different phenomena.

Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22), pages 42 - 52
December 7, 2022 ©2022 Association for Computational Linguistics

the term *cross-lingual* to denote utterances which contain one or more words from a mixing language while belonging to a frame language. These may be caused by any of the mentioned phenomena, *i.e.* code-switching, lexical borrowing and foreign named entities.

A major challenge, both in improving the performances of multilingual models on cross-lingual data and in their overall evaluation, is the scarcity of cross-lingual datasets. Nevertheless, while human annotation is already costly and time-consuming in general, annotating cross-lingual data is made harder by the fact that bilingual annotators are needed for each pair of languages of interest; these may be especially hard to find for less common languages. In particular, while there has been some interest for different kinds of data (*e.g.* social media), voice assistant data, which is the focus of this paper, has been mostly ignored. Although such datasets may be obtained by crowdsourcing, the process would be expensive and time-consuming. This reason leads to the necessity of a procedure to generate synthetic data over several language pairs while providing large-scale datasets. These can be used to train a learning model to infer cross-lingual utterances. The trained model can finally be employed on voice assistants data to detect real cross-lingual utterances.

Our contribution is three-fold: (i) We propose in section 3 a scalable synthetic data generation technique to obtain challenging benchmarks which exhibit a significant ratio of cross-lingual influences. The method is language agnostic and here we employ it on four common European languages (German, French, Italian, Spanish) with English as mixing language. (ii) We compare the performance of different baselines in detecting cross-lingual utterances by solving the more fine-grained task of word level language identification. To validate the generation procedure, we test the models trained on the synthetic distribution over a benchmark dataset obtained through an extremely precise heuristic. (iii) Finally, we analyze in section 6 the phenomenon of cross-lingual influence in a large set of cross-lingual utterances detected using our method on Alexa user queries.

2 Related work

Code-switching has received significant interest both in the linguistic literature (Poplack, 2004, 1980; Lipski, 2005; Bhatt and Bolonyai, 2011) and in Natural Language Processing (NLP); (Sitaram et al., 2019) provide a survey of code-switching in NLP. From a linguistic point of view, the two phenomena differ in the fact that the latter occurs in the lexicon, while code-switching mostly regards the utterance-construction level (Muysken, 1995). Despite the apparently different definitions, the two are not always clearly distinct from one another, and may be thought of as lying on a continuum (Sitaram et al., 2019; Bali et al., 2014).

	de	fr	it	es
code switched	31359	5391	6139	4256
non code switched	63944	18491	20100	23744

Table 1: Size of the four benchmark datasets.

Various efforts have been made to collect code-switched annotated data over which to perform core NLP tasks, such as NER (Aguilar et al., 2018; Singh et al., 2018), POS (Vyas et al., 2014; Barman et al., 2016) and ASR (Lyu et al., 2015; Deuchar et al., 2014). Nevertheless, most of the available resources have been gathered from Twitter, and therefore do not resemble the distribution of data encountered by a voice assistant. Few works exist on generating synthetic CS data: in (Pratapa et al., 2018), a synthetic dataset is obtained by applying linguistic theory-based rules, while in (Gupta et al., 2020) an encoder-decoder architecture is used for the generation. These approaches, however, focus on strict code-switching, while we aim to also encompass lexical borrowing and foreign named entities.

The de-facto standard way to infer code-switched utterances is to train models on the task of word-level language identification. Again, existing datasets of code-switched text annotated with word-level language labels have been collected from Twitter (Patro et al., 2017; Maharjan et al., 2015) or Facebook (Barman et al., 2014), leaving conversational data out of the scope. Provided a word-level annotated dataset, any sequence-labeling algorithm can be employed to solve the task. Approaches include conditional random fields (Sikdar and Gambäck, 2016; Shrestha, 2016), recurrent neural networks (Chang and Lin, 2014; Samih et al., 2016) and transfer learning (Aguilar and Solorio, 2020).

3 Data

3.1 Synthetic data generation

As anticipated in section 1, the cost in time and resources of annotating large-scale datasets by crowdsourcing makes synthetic generation the only viable alternative. However, these phenomena show a significant degree of mutability both in time and space (Sitaram et al., 2019), making them elusive to be addressed in a unified manner which is theoretically sound. While some have tried to generate linguistically-correct code-switched data (Pratapa et al., 2018), we trade off a rigorous formulation with a simpler one to deal with all the considered phenomena in a unified manner. Requiring no real cross-lingual (CL) samples, our scalable approach generalizes among any pair of languages. We show that this relaxation does not undermine the effectiveness of the approach by benchmarking a model trained on such generated data over a high-precision CL dataset ("benchmark dataset"). Indeed, our objective is to generate a dataset rich of cross-linguality which can be used to train a model able to detect any CL utterance (code-switching, language borrowing etc.).

The generation follows (Gella et al., 2014), where each utterance can have at most two languages and at most one switching point. While these may not be true in general, they mostly hold in voice assistant data, where utterances are usually short.

Slot switching Our procedure leverages *slot resolution artifacts* which are typically available to conversational agents[2]: these, in fact, need to map entities to actionable items, *e.g.* both '*chapter*', '*section*' and '*paragraph*' are mapped to a coarser entity type which denotes more generally a part of a book. Slot resolution artifacts are usually implemented as human-authored many-to-one maps, where the fine-grained entities are language-specific and the coarser entity type is language-agnostic. The latter can be used as a syntactically safe switching point to obtain cross-lingual utterances. A cross-lingual dataset can be obtained from a chosen monolingual dataset in the frame language by matching instantiations of entity types in the frame-locale utterances and replacing them with random instantiations of the same entity type in the mixing locale. Then, to obtain the token-level language annotations, it is sufficient to assign each switched token to the mixing language. For example, for

(1) "A_{IT} che_{IT} $\text{capitolo}_{\text{IT}}$ sono_{IT} $\text{arrivato}_{\text{IT}}$"

we use the map {capitolo, sezione, paragrafo $\rightarrow$ BOOKSECTION} to obtain the language-agnostic entity 'BOOKSECTION' which contains a set of its instantiations in English (or any other language) {chapter, section, paragraph $\rightarrow$ BOOKSECTION}, allowing us to pick one to produce

(2) "A_{IT} che_{IT} $\text{chapter}_{\text{EN}}$ sono_{IT} $\text{arrivato}_{\text{IT}}$".

We empirically set the mixing probability to 70% after inspecting a subset of utterances. As the mapping from the language-agnostic entities to their instantiations in a chosen language is not univocal, we choose one of the latter at random.

Named-entities switching Nevertheless, slot resolution artifacts only cover specific slots. Another common phenomenon is the use of English words in named entities, such as song names, video names or app names. To obtain a reliable language annotation for named entities, we use a high-precision and low-recall heuristic that checks that each token of the named entity is part of only a specific language dictionary. For instance, when using IT as a frame language and EN as a mixing language, given a song such as "*nel blu dipinto di blu*" we check if '*nel*', '*blu*', '*dipinto*', '*di*', '*blu*' are all part of the IT dictionary and none of them is part of the EN dictionary. Only in that case they are placed in the IT catalog; if the converse happens, they are placed in the EN catalog. Entities for which none of these events happens are not switched. We populate the language-specific catalogs from the data and replace the named entities sampling from either the frame or mixing catalogs of the same entity type (e.g. "Song" $\rightarrow$ sample using the song names catalogs) with a probability proportional to the catalog size. This method creates fairly representative utterances in the context of personal assistants, since we mainly have short sentences with cross-linguality concentrated on named entities and loanwords. While the framework is general and can be used for any pair of languages, we used English as mixing language for the four considered

[2] As an alternative, publicly available resources may also be used: a slot can replaced with a word in the same WordNet synset (Fellbaum, 1998). WordNet has been translated and adapted to many languages, like German, French, Italian, and Spanish (Hamp and Feldweg, 1997; Sagot and Fišer, 2008; Toral et al., 2010; Gonzalez-Agirre et al., 2012).

European languages to mimic the real linguistic phenomenon. We applied our method to manually annotated, de-identified and anonymized Alexa utterances. These span more than two years of data for all the languages considered. Starting from these data we create our cross-lingual data set. We generated four datasets of $\approx 100k$ utterances for the four corresponding locales, each split in training, validation and test with a 80-10-10 ratio. This size was chosen to keep a fairly high variance of the English words present in the utterances.

It is worth to note that we do not require, and hence do not expect, the generated utterances to faithfully resemble the cross-lingual phenomena that we aim to capture. In fact, adhering to the definition of cross-linguality that we outlined in section 1, we more simply aim to generate utterances in a frame language containing one or more words from a mixing language, possibly preserving the original syntax and semantics. If we now consider the set of natural cross-lingual utterances to be a subset of all the possible cross-lingual utterances, we have that a model capable of detecting samples from the former should also be able to detect those from the latter. Given that the set of natural cross-lingual utterances is constrained by the linguistic patterns of the considered phenomena, the subset assumption makes intuitive sense but is not assumed to hold for all distributions. We show, however, that this assumption is valid enough to capture most of the cross-linguality in conversational data, assessing the effectiveness of models trained on the synthetic distribution on a benchmark of real cross-lingual utterances.

3.2 Benchmark dataset

To validate our data generation technique, we need a ground-truth dataset over which to evaluate the proposed models after they have been trained on the generated distribution. Provided that no such dataset exists for conversational data, we take inspiration from (Mendels et al., 2018) to obtain a high-precision set of utterances from de-identified and anonymized live traffic. The approach leverages the idea of anchor words, *i.e.* words belonging specifically to one language among a large pool of languages. Provided anchor words for both the frame and mixing languages, an utterance is code-switched if it contains both an anchor from the frame and one from the mixing language. Analogously to (Mendels et al., 2018), we relax the definition of anchor word by restricting the pool of languages to contain only the mixing language, yielding what are called *weak* anchor words. This is motivated by the fact that most foreign words in the considered frame languages are English, so this relaxation significantly improves the recall while keeping its false positive rate minimum. The set of weak anchor words for the frame language L can be computed as the set difference between its word lexicon V_L and the lexicon of the mixing language $V_{L'}$

$$\text{AnchorSet}(L) = V_L \setminus V_{L'}. \quad (1)$$

The set of weak anchor words for the mixing language can be computed in the symmetric way.

While this procedure has limitations in terms of recall, the obtained set of utterances exhibits almost no false positives. Nevertheless, to obtain a benchmark dataset over which to evaluate both False Positive Rate (FPR) and recall of the trained models, negative samples are also needed. For this we use the set of utterances for which all the words are anchor words of the frame language. As before, although many not code-switched utterances will be this way ignored, the resulting ones will be negative samples with extremely high confidence.

To avoid making assumptions on the ratio of code-switched utterances, the two datasets are kept separated. The one consisting of only code-switched utterances is used to compute the recall, while the one containing only non-code-switched utterances is used to compute the FPR. Table 1 shows the dimensions of the four datasets.

4 Models

We describe in this section the proposed baselines, namely an ad hoc deterministic heuristic and two neural models. These will be trained over the synthetic datasets generated according to section 3 and used to infer real code-switched Alexa utterances.

We consider as baseline a dictionary-based heuristic parameterized by two thresholds t_1 and t_2. The latter deterministically classifies an utterance as code mixed if at least $t_1\%$ of the lemmatized words do not appear in the frame language vocabulary while appearing in the mixing language vocabulary and no more than $t_2\%$ appear in the mixing vocabulary while not belonging to the frame vocabulary. Despite its simplicity, the heuristic allows to arbitrarily trade-off recall and precision by manually tweaking the two parameters.

We then propose two neural models, one character based and the other transformer based. The intuition behind the former is that character-level convolutions (Sitaram et al., 2019) should be able to capture the distinguishing morphological features of the considered languages which are key to the task. In particular, given an input utterance, each word is split in characters and embedded via a trainable embedding layer to obtain $\mathbf{w} \in \mathbb{R}^{l \times d}$, where l is the maximum word length encountered in the data and $d = 50$ is the chosen embedding dimension. The embedded word is then passed through a set of $m = 256$ 1-D convolutional filters with kernel size $k = 3$, yielding a tensor $\in \mathbb{R}^{m \times o}$, where o is given by $(l - k + 1)$. At this point, the maximum is taken along the axis on which the resulting feature maps are stacked, so to have a new word embedding tensor $\mathbf{e} \in \mathbb{R}^{o}$. Three different sets of filters of different kernel sizes are then passed over $\mathbf{e}$, having sizes $3, 4$ and 5 in our implementation. Max pooling over time allows to obtain a fixed-dimension digest for each of the resulting maps, which can be concatenated to form a single tensor to be fed to a bidirectional LSTM along with the rest of the utterance. The latter returns a dynamic representation of the word and its context, which is then mapped to the label space by a standard fully-connected layer. A visual overview of the architecture is given in fig. 1. We will refer to this model as '`CharBased`'. The second proposed neural model leverages multilingual BERT (Devlin et al., 2019) to obtain contextualized embeddings which are then fed to a standard sequence classification pipeline, as can be seen in fig. 2. In details, each word is first tokenized and encoded by the mBERT tokenizer and fed to a pretrained mBERT model along with the whole utterance. The embedding is then provided by the last hidden state of the pretrained model. Since the tokenizer is based on the Wordpiece model (Schuster and Nakajima, 2012), words are often split in subwords: the word '*microfono*' for example would be split in '*micro*' and '*##fono*'. To still obtain word-level predictions, the resulting embeddings are averaged. Utterances are finally fed to a bidirectional LSTM whose output is mapped to the label space again by a fully-connected layer. We will refer to this model as '`BertBased`' in the rest of the paper.

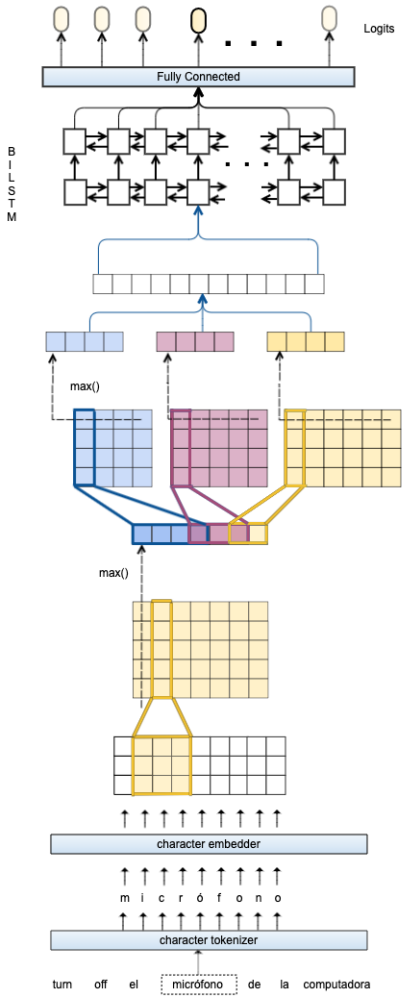

Figure 1: Diagram of the character-convolution-based model.

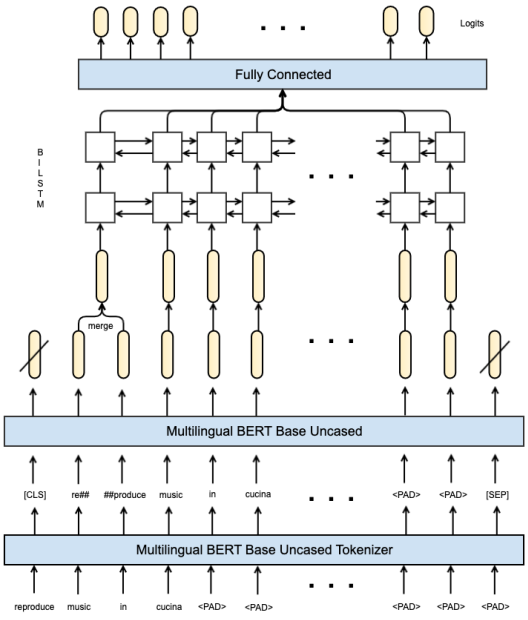

Figure 2: Diagram of the contextual model.

Code-Switching Detection on synthetic data												
	DE			FR			IT			ES		
	F1	prec	recall	F1	prec	recall	F1	prec	recall	F1	prec	recall
`Baseline`	+0%	+0%	+0%	+0%	+0%	+0%	+0%	+0%	+0%	+0%	+0%	+0%
`CharBased`	+25.1%	+35.9%	+9%	+23.9%	+35.7%	+8%	+29.4%	+40.7%	+12.7%	+29.7%	+40.0%	+13.3%
`BertBased`	+26.4%	+37.7%	+10.5%	+25.8%	+36.2%	+11.2%	+30.6%	+40.9%	+15.1%	+31.2%	+42.2%	+14.1%

Table 2: Evaluation results for the task of code-switched utterance detection of the two neural models expressed as relative improvement over the threshold based `Baseline` presented in section 4, performed over a held-out artificially generated test set.

Code-Switching Detection on benchmark data								
	DE		FR		IT		ES	
	recall	FPR	recall	FPR	recall	FPR	recall	FPR
`Baseline`	+0%	+0%	+0%	+0%	+0%	+0%	+0%	+0%
`BertBased`	+7.7%	−22.4%	−1.9%	−48.4%	+2.5%	−46.6%	−0.2%	−35.5%
`CharBased`	+18.4%	−21.3%	+2.9%	−48.7%	+7.9%	−46.2%	+3.4%	−35.4%

Table 3: Evaluation results for the task of code-switched utterance detection of the two neural models expressed as relative improvement over the threshold based `Baseline` presented in section 4, performed over the benchmark dataset obtained as in section 3.

	DE	FR	IT	ES
DE		+16%	−1%	+53%
FR			−15%	+31%
IT				+54%

Table 4: Relative difference in % of utterances containing cross-lingual phenomena by country. Cell ij contains the difference in the ratio of cross-lingual utterances between language i and language j.

5 Evaluation

As can be seen in table 2, the two neural models obtain similar results on a held-out test set generated according to the same procedure presented in section 3, with `BertBased` slightly outperforming the character based model. On the other hand, table 3 shows that the latter obtains the best results on the benchmark dataset, yielding much higher recall while maintaining a low False Positive Rate (FPR). The results are expressed as relative improvements of the two models over the deterministic heuristic introduced in section 4. Precision and recall are given in table 3 because they are computed on two separate datasets to avoid having to pick an arbitrary ratio between code-switched and non-code-switched utterances.

6 Results

Object of this analysis are code-switched utterances detected from real Alexa queries by a model trained on an artificial dataset generated according to section 3. A separate model was trained for each locale versus English, and the inference was made on real data coming from the corresponding locale. As can be seen in table 4, German, French and Italian exhibit similar ratios of cross-lingual utterances, with Italy being the country where they are most common. On the other hand, Spanish shows a remarkably different situation. As shown in fig. 4, this difference is mostly attributable to English words which do not represent named entities: in Spain, people for example do not use '*timer*', '*computer*' or '*film*', as they prefer their Spanish correspondants '*temporizadora*', '*computatora*' and '*pelicula*'. This phenomenon is confirmed in fig. 3, where we see the most common words causing cross-linguality. Figure 3 also shows that the distribution is extremely skewed: for instance, '*timer*' in Italian causes almost the 10% of all the cross-lingual utterances. This phenomenon reflects the underlying distribution of voice assistant utterances, where a set of frequent queries make up for a large part of all of utterances. Finally, we can see in fig. 5 the way cross-lingual utterances are distributed in different domains is common to the different locales. Coherently with the large amount of foreign named entities causing cross-linguality, we can see that most utterances belong to 'Media & Entertainment', which is expected to contain many international artists and song names. '*DeviceControl*' also accounts for a significant part of the utterances; these usually contain commands, like for example '*play*', '*next*', '*stop*' etc., which are traditionally expressed in English even in non-English speaking countries.

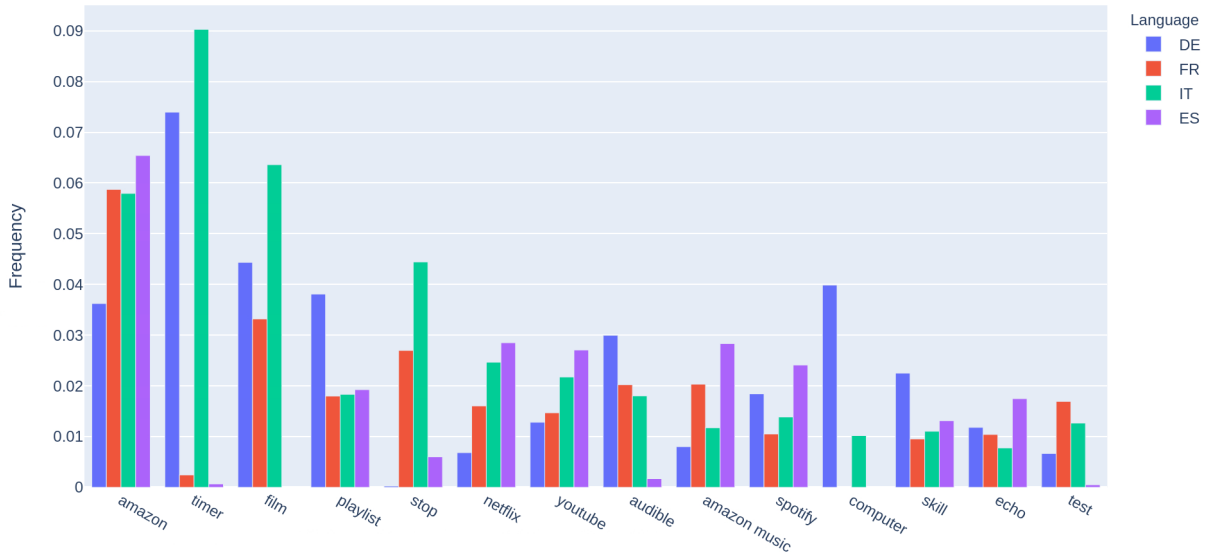

Figure 3: Most common English words used when interacting with Alexa in the four considered locales.

7 Conclusions

In this paper, we have presented a large-scale analysis of the cross-lingual phenomena encountered by voice assistants. We first have proposed an artificial data generation technique, then we have presented two neural models that can be trained on the synthetic data to infer real cross-lingual utterances. Finally, we have employed the top-performing model to infer such utterances from real data. The fact that loanwords and foreign named entities cover most of the found cross-lingual utterances may indicate that code-switching is rare in voice assistants in the considered locales. This may be explained by the fact that users code-switch the most in colloquial situations, while their way of speaking when querying a voice assistant is constrained by its understanding capacity. Nonetheless, multilingual models still have a great opportunity of transfer learning on the large amount of foreign named entities and loanwords that are present in the data. The results show that the use of English words in DE, FR, IT, ES is strongly skewed on popular entities such as '*Amazon*', '*Netflix*', and '*YouTube*', and on specific loanwords such as '*timer*', '*computer*' and '*stop*'. The use of these popular named entities is consistent across locales and the ratio of cross-lingual interactions is similar, except for ES, where users tend to prefer Spanish words to English loanwords. The analysis also shows that most of the mixing words are contained in the 'Media & Entertainment' domains and on named entities such as Service Names, Media names, Item names and Dish names.

As we have explained in section 3, the current generation technique does not aim to model the complex phenomenon of code-switching in a theoretically correct manner. The simplicity of the procedure nevertheless allows it to be repurposed to focus on the latter. An interesting future direction could be to limit the attention to code-switching in the data generation, so that a model trained on that data could be used to collect a code-switched dataset of voice assistant queries. Given the low FPR exhibited by the model, the collected utterances represent an high-quality resource which could in future be used to train generative models to produce better synthetic data, which in turn can be used to train detectors in an iterative manner.

From an architectural prospective, models tackling word-level language identification expressly designed to solve the task of cross-lingual or code-switched detection could benefit from the utterance-level information about their distribution in the dataset. This could encourage the design of a multi-headed model tackling both tasks in an end-to-end approach.

Finally, we aim to expand the set of considered languages to encompass other frame and mixing languages, for example considering Hinglish in India. It might be particularly interesting to compare the obtained results for Spanish with ones obtained over Spanish spoken in the United States and in Mexico, as they may involve more code-switching.

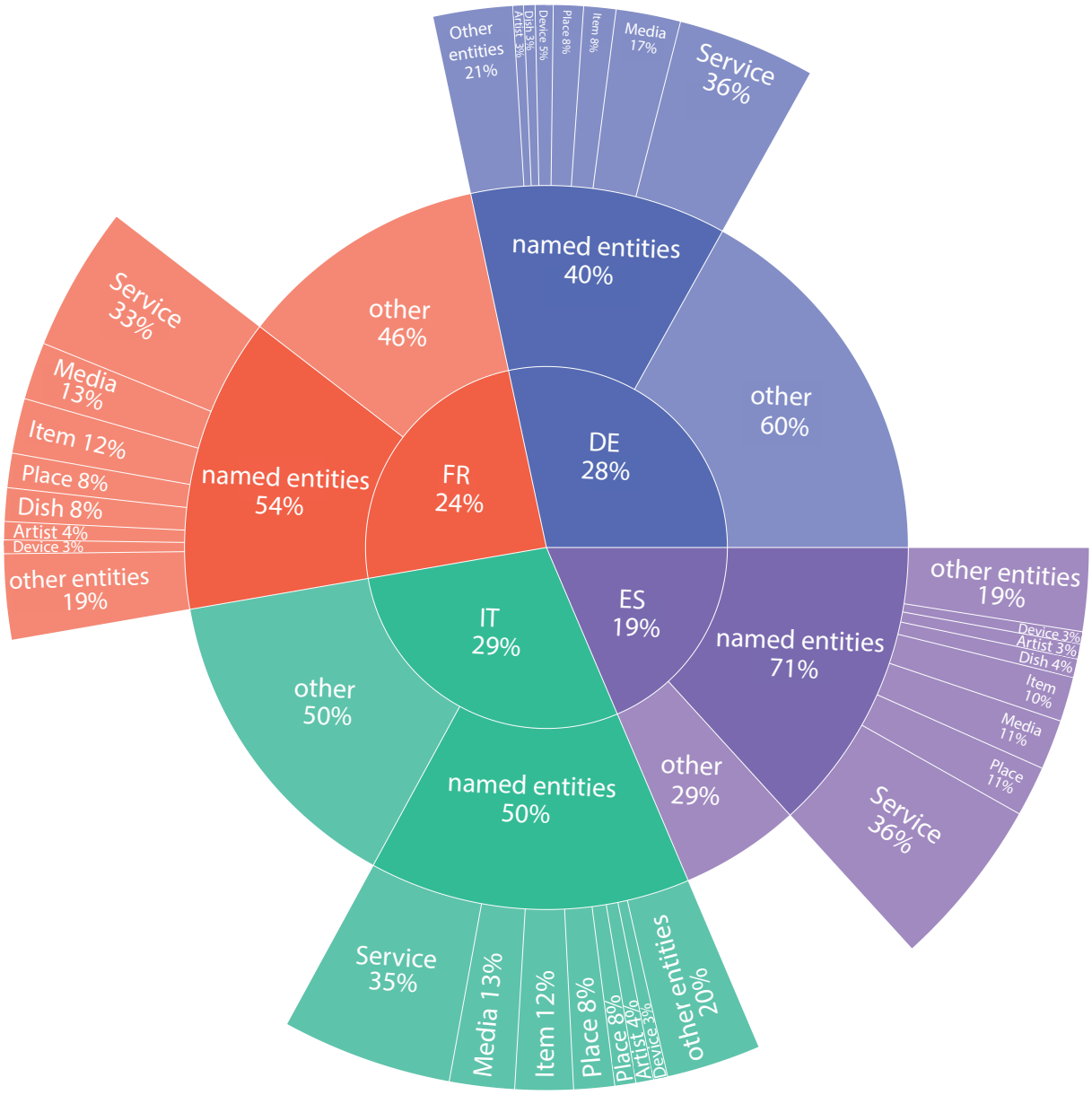

Figure 4: Distribution of a set of ≈ 60k cross-lingual utterances.

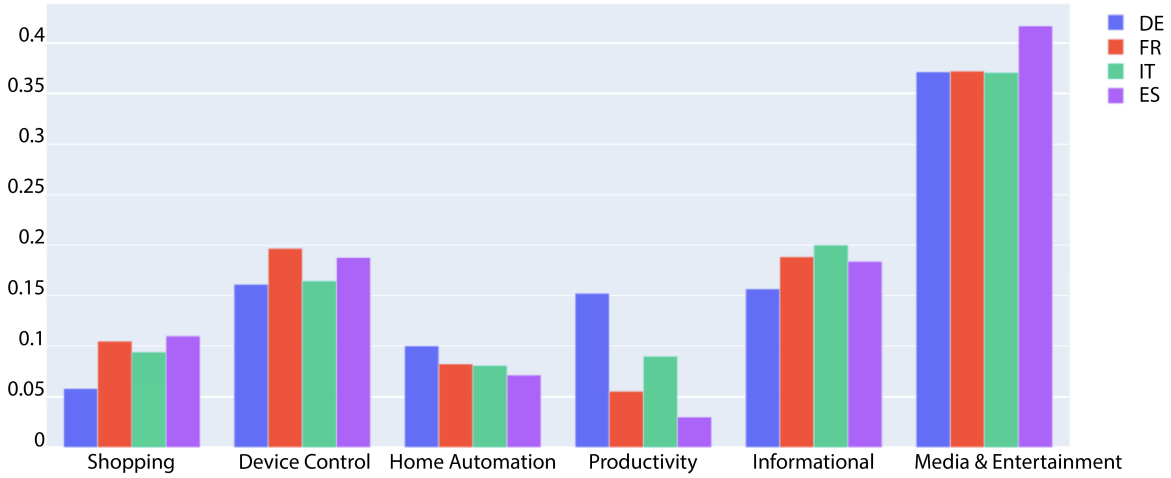

Figure 5: Distribution of domains in cross-linguistic utterances.

8 Limitations

An overall limitation of the work stands from the lack of absolute results, as the latter can only be disclosed as relative improvements over a baseline due to internal policy. As stated in sections 1 and 3, the analysis only regards four European languages (German, French, Italian and Spanish) with English as mixing language. Therefore, while the same approach can be used with different languages, the reported findings only regard the mentioned ones. Moreover, the quality of the generated synthetic data heavily depends on the quality of the slot resolution artifacts presented in section 3. In this work, these artifacts are human-curated according to the highest industry standards, but are subject to IP and hence not publicly accessible. Unfortunately, this also makes the code non disclosable. Finally, as discussed in section 3, the data generation technique may not fully capture the complex linguistic patterns involved in code-switching. We argue that it is however enough to encompass a large quantity of cross-lingual utterances encountered by vocal assistants, and prove it by showing the efficacy of the models trained over synth data in dealing with a high-precision benchmark dataset of real cross-lingual utterances.

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Gustavo Aguilar and Thamar Solorio. 2020. From English to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8033–8044, Online. Association for Computational Linguistics.

Conneau Alexis and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar. Association for Computational Linguistics.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Utsab Barman, Joachim Wagner, and Jennifer Foster. 2016. Part-of-speech tagging of code-mixed social media content: Pipeline, stacking and joint modelling. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 30–39, Austin, Texas. Association for Computational Linguistics.

Rakesh M. Bhatt and Agnes Bolonyai. 2011. Code-switching and the optimal grammar of bilingual language use. *Bilingualism: Language and Cognition*, 14(4):522–546.

Joseph Chee Chang and Chu-Cheng Lin. 2014. Recurrent-neural-network for language detection on twitter code-switching corpus. *CoRR*, abs/1412.4314.

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.

Margaret Deuchar, Peredur Davies, Jon Russell Herring, M. Carmen Parafita Couto, and Diana Carter. 2014. *5. Building Bilingual Corpora*, pages 93–110. Multilingual Matters.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Spandana Gella, Kalika Bali, and Monojit Choudhury. 2014. "ye word kis lang ka hai bhai?" testing the limits of word level language identification. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 368–377, Goa, India. NLP Association of India.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280, Online. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

John M. Lipski. 2005. Code-switching or borrowing? no sé so no puedo decir, you know.

Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li. 2015. Mandarin–english code-switching speech corpus in south-east asia: Seame. *Language Resources and Evaluation*, 49(3):581–600.

Suraj Maharjan, Elizabeth Blair, Steven Bethard, and Thamar Solorio. 2015. Developing language-tagged corpora for code-switching tweets. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84, Denver, Colorado, USA. Association for Computational Linguistics.

Gideon Mendels, Victor Soto, Aaron Jaech, and Julia Hirschberg. 2018. Collecting code-switched data from social media. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pieter Muysken. 1995. *Code-switching and grammatical theory*, page 177–198. Cambridge University Press.

Jasabanta Patro, Bidisha Samanta, Saurabh Singh, Abhipsa Basu, Prithwish Mukherjee, Monojit Choudhury, and Animesh Mukherjee. 2017. All that is English may be Hindi: Enhancing language identification through automatic ranking of the likeliness of word borrowing in social media. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2264–2274, Copenhagen, Denmark. Association for Computational Linguistics.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching1. 18(7-8):581–618.

Shana Poplack. 2004. *Code-Switching*, pages 589–596.

Nanda Poulisse. 1998. Duelling languages: Grammatical structure in codeswitching. *International Journal of Bilingualism*, 2(3):377–380.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. 2016. Multilingual code-switching identification via LSTM recurrent neural networks. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 50–59, Austin, Texas. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Prajwol Shrestha. 2016. Codeswitching detection via lexical features in conditional random fields. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 121–126, Austin, Texas. Association for Computational Linguistics.

Utpal Kumar Sikdar and Björn Gambäck. 2016. Language identification in code-switched text using conditional random fields and babelnet. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 127–131, Austin, Texas. Association for Computational Linguistics.

Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Named entity recognition for Hindi-English code-mixed social media text. In *Proceedings of the Seventh Named Entities Workshop*, pages 27–35, Melbourne, Australia. Association for Computational Linguistics.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W. Black. 2019. A survey of code-switched speech and language processing. *CoRR*, abs/1904.00784.

Antonio Toral, Stefania Bracale, Monica Monachini, and Claudia Soria. 2010. Rejuvenating the italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*, Mumbai.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979, Doha, Qatar. Association for Computational Linguistics.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.