Improving code-switched ASR with linguistic information

Jie Chi and **Peter Bell**
Centre for Speech Technology Research, University of Edinburgh, UK

Abstract

This paper seeks to improve the performance of automatic speech recognition (ASR) systems operating on code-switched speech. Code-switching refers to the alternation of languages within a conversation, a phenomenon that is of increasing importance considering the rapid rise in the number of bilingual speakers in the world. It is particularly challenging for ASR owing to the relative scarcity of code-switching speech and text data, even when the individual languages are themselves well-resourced. This paper proposes to overcome this challenge by applying linguistic theories in order to generate more realistic code-switching text, necessary for language modelling in ASR. Working with English-Spanish code-switching, we find that Equivalence Constraint theory and part-of-speech labelling are particularly helpful for text generation, and bring 2% improvement to ASR performance.

1 Introduction

Although accurate statistics on the number of worldwide bilingual speakers are hard to determine, it has been generally believed that more than half of the population can communicate in more than one language (Ansaldo et al., 2008; Bialystok et al., 2012; Grosjean, 2010). With the rising popularity of voice assistant and translation applications, automatic speech recognition (ASR) systems have been increasingly integrated into people's lives. Considering the abundance of bilingual countries [1], and code-switching [2] is common in everyday conversations, there has been great interest into developing such system in the corresponding setting.

The most widely accepted definition of code-switching is the phenomenon whereby a speaker shifts from one language to another within a single utterance, especially in an environment in which both languages are being used (Heredia and Altarriba, 2001). Some previous work make a distinction between *code-switching* and *code-mixing*, where the former occurs at sentence-level while the latter occurs at word-level (Myers-Scotton, 1997; Gumperz, 1982). However, in recent years, this distinction has becomes unclear (Bali et al., 2014). To avoid confusion, we will only use the term code-switching in this work, and denote the differences in switching position as sub-types of it (Myers-Scotton, 1989; Muysken et al., 2000; Major, 2002; Winford, 2003). Although different language pairs may present varying extents or types of code-switching, they can generally be categorised as inter-sentential, intra-sentential, and tag switching, where respectively the language switches at sentence or clause boundary; within the sentence or clause; or by inserting a tag phrase [3]. In this paper, we focus only intra-sentential switching, which is a much harder task compared with other types, because the acoustic variance of mixed languages within the sentence can be larger than across sentence (Li et al., 2019).

The challenge of developing a code-switched ASR system comes from both linguistic and computational perspectives. On the one hand, code-switching can be driven by multiple factors, which makes it flexible but hard to predict. Linguists have studied the phenomenon for decades, and the main views held are that people tend to code-switch to compensate for lack of language proficiency, express solidarity or certain feelings, discuss some particular topic, and distinguish themselves from other classes to imply a certain social status (Grosjean, 1982; Holmes and Wilson, 2017; Leung, 2006). ASR systems generally require a

[1] https://www.uottawa.ca/clmc/55-bilingual-countries-world

[2] Code-switching can also be found in text, such as social media or news paper headlines, but in this paper we are only focus only on the spoken form

[3] Intra-word switching can be arguably considered as a type of code-switching (Myers-Scotton, 1989), but a complete discussion is out of the scope of this paper.

Proceedings of the 29th International Conference on Computational Linguistics, pages 7171–7176
October 12–17, 2022.

large amount of transcribed data in a monolingual setting. However, a relatively small number of the approximately 7000 existing languages have large readily corpora, and the data scarcity is more severe for code-switching problems, which – to make matters worse – usually involve one or more low-resourced languages (Austin and Sallabank, 2011).

Motivated by the these considerations, we propose a novel code-switched ASR framework with the aid of established linguistic theories. We demonstrate the effectiveness of the approach on Spanish-English conversational code-switching data from the Bangor-Miami corpus (Deuchar, 2011). In doing so, we prove both phonological and syntactic information can improve the performance of language modelling and ASR.

2 Related work

There have been many attempts to approach the problem of modelling code-switched speech from acoustic, pronunciation and language modeling perspectives for conventional hybrid ASR systems. In early work a language identification (LID) system was combined to determine which hypothesis from multiple monolingual decoders should be chosen (Lyu et al., 2006; Wu et al., 2006; Bhuvanagiri and Kopparapu, 2010). However, the drawback of this approach is a strong assumption that the speech segments are independent from each other and a heavy dependency on the accuracy of LID module (Weiner et al., 2012). The choice of phone inventories is important, and many studies have been conducted to merge phone sets of different language pairs manually or automatically, which improves the performance of ASR systems by effectively yielding more training data for each phone, and enabling the pronunciation influence between languages to be learnt (Kohler, 1998; Lyudovyk and Pylypenko, 2014; Sivasankaran et al., 2018). Considering that there is a much larger amount of monolingual text than code-switched text, code-switched text generation by imposing language theories to parallel monolingual texts has been a popular research direction (Li and Fung, 2014; Winata et al., 2019; Pratapa et al., 2018). From this, language models can be improved by training them on the generated text or can also be achieved by apply the theories directly to restrict the search paths in a WFST framework(Li and Fung, 2013).

In recent years, end-to-end (E2E) models have also been explored to handle the problem. To address the issue that E2E models usually require a large amount of data, transfer learning from monolingual models has been used (Luo et al., 2018; Shan et al., 2019; Mary N J et al., 2020; Winata et al., 2020). However, when the models are fine-tuned on code-switched data, the performance on monolingual speech is degraded. To improve the robustness of the model, techniques such as Learning Without Forgetting and adversarial training have been proposed (Shah et al., 2020; Madhumani et al., 2020).

3 Methodology

3.1 Phoneme mapping

We use the standard International Phonetic Alphabet (IPA) as the basis for our acoustic modelling units. As English and Spanish have partly different inventories (Edwards, 1992; Goldstein, 2000), instead of treating them as completely two phone sets, we merge them according to their phonological features (Mortensen et al., 2016). The features are a set of global attributes, which describes the characteristic of sound, such as whether it is produced with nasal airflow. After representing the feature with vectors, where each attribute can be negative or positive, we compute the hamming edit distance between each pair of phonemes. In this way, we map each Spanish phoneme to its nearest English equivalent.

3.2 CS text generation

3.2.1 Parallel text generation

We use the Google translate API to generate parallel English and Spanish text. The API is not only capable of translating from one to the other, but also can translate code-switched sentences to one of the language while keeping the segment in the target language unchanged. We receive one translated sentence for each monolingual text in the corpus, while for each code-switched sentence, we obtain translations in both languages. As the translation quality varies across the sentences under manual inspection, we use Pseudo Fuzzy-match Score (PFS) shown in Equation 1 to filter any translation pairs whose PFS is less than 0.6 (He et al., 2010; Pratapa et al., 2018). s here is the monolingual source sentence, we forward translate s to target t, then reverse translate the target t into pseudo source s'.

$$PFS = \frac{EditDistance(s, s')}{max(|s|, |s'|)} \quad (1)$$

3.2.2 Constituency parse generation

To generate word level alignments between parallel sentences we use *fast_align* toolkit, which is a unsupervised aligner (Dyer et al., 2013). Then following (Pratapa et al., 2018), we generate the parse tree for English text with Berkeley neural parser (Kitaev and Klein, 2018) and then use the alignments to generate the equivalent parse for the Spanish sentence.

3.2.3 Equivalence Constraint theory

The Equivalence Constraint (EC) theory claims that, in general, *"Codes will switch at points where the surface structures of the languages map onto each other"* (Sankoff and Poplack, 1981). For example, in English and Spanish, code-switching cannot happen within possessive phrases or noun/adjective clauses because the grammatical structures are different and thus reject the transfer.

EC has been successfully applied to code-switched text data and approved crucial for language modelling task (Pratapa et al., 2018), which is our implementation inspired from. However, it should be noted that even though text data can be in the form of informal conversation on Twitter or other Internet platforms, they may still not follow the same patterns as speech (Sitaram et al., 2019).

We apply EC to generate CS text over permutation. To improve the naturalness, we rank generated texts for each pair by following metrics and select at most top 10 for each sentence pair.

Switch-points are points within a sentence where the languages of the words on the two sides are different (Pratapa et al., 2018). The metric SP Fraction (SPF) is defined as the number of SP in a sentence divided by the total number of word boundaries in the sentence. We set it to 0.1 for all experiments.

CMI counts the number of switches in the utterance (Gambäck and Das). It can be computed at the utterance level by finding the most frequent language in the utterance and then counting the frequency of the words belonging to all other languages present. The computation is shown in Equation 2, where W denotes the utterance and $N(W)$ denotes the number of words in W. l_max means the count of words in the most frequent language in that utterance, and $P(W)$ is the number of SP. We set it to 0.3 for all experiments.

$$CMI(W) = \frac{N(W) - l\_max + P(W)}{N(W)} \quad (2)$$

POS tags As we have translate code-switched text to parallel sentences, with monolingual words remained, we can find the POS tags of the switched words in both languages. Therefore, by calculating the frequency of the POS tags of the preceding/current/following words, we give the sentence higher ranking if proper nouns, nouns, determiners or interjections precede switched words, and nouns or subordinating conjunctions are switched, which is consistent with (Soto et al., 2018)

4 Experimental setup

4.1 Data

Although the corpus is public, there has been no standard preprocessing procedures for it. In this paper, we first classified all utterances into Spanish, English and Code-switched sets. If there are both English and Spanish exclusive words in the utterance, we consider it as code-switched case, but if there are words which exist both in the English and Spanish lexicon, the category depends on the rest of the sentence [4]. After cleaning, we retain 44971 utterances, splitting them into training, development and test by 7:1:2. To better illustrate the distribution of the dataset, the statistics are presented in Table 1.

4.2 Training

4.2.1 Acoustic models

We used the Kaldi TDNN recipe [5] to develop the hybrid systems. 40 dimensional MFCC features are extracted first to train a GMM-HMM model. Before training the neural network, we apply speed perturbation to augment data. The TDNN-HMM training is identical for all systems: We use 40 dimensional high resolution MFCC with 100 dimensional i-vector features as input. The network consists of 7 TDNN hidden layers which contain 758 hidden units per layer. The start and final learning rates are 0.00015 and 0.000015 respectively and we train the model for 4 epochs with a mini-batch size of 128.

4.2.2 Language models

We use SRILM toolkit to train n-gram models for comparison. For each setting, we trained a 3-gram model for decoding and a 4-gram for rescoring. The lexicon is identical to all experiments. We

[4] The processing script will be released after acceptance.

[5] https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5

Table 1: The statistics of the processed Miami corpus, where the duration unit is hour.

	English		Spanish		CS	
	Number	Duration	Number	Duration	Number	Duration
Training	20813	10.9	7789	4.3	1879	1.6
Dev	3000	1.6	1250	0.6	250	0.2
Test	6000	3.1	2500	1.2	500	0.5

Table 2: WER and PPL on testset, where the top block shares the same language model which is trained only on the original transcript, and the bottom block shares the acoustic model with phoneme mapping.

	Test WER				Test PPL		
	English	Spanish	CS	total	English	Spanish	CS
baseline	44.0	56.8	49.3	47.8	109.7	144.8	152.8
mapping base	43.6	56.4	49.1	47.5	109.7	144.8	152.8
translation	43.4	56.0	49.0	47.2	90.3	126.4	134.9
+ external	43.4	56.0	49.0	47.3	92.3	128.4	149.6
+ random	43.4	55.9	49.1	47.2	89.1	127.8	145.2
+ POS	43.3	55.6	48.7	47.0	88.7	126.0	130.1
+ EC	43.3	55.6	48.6	47.0	87.2	122.8	125.7
+ EC + POS	43.2	55.3	48.4	46.9	87.1	120.2	123.8

used the CMUDict for English and Santiago Spanish Lexicon for Spanish. There are in total 206500 English words and 91121 Spanish words, any uncovered words are treated as UNK.

5 Results and discussion

Table 2 presents the word error rate (WER) and perplexity (PPL) on the test set for all experiments. Our *baseline* model uses the union of English and Spanish phoneme sets while *mapping base* maps the Spanish phoneme set to English phoneme set. Their language models are identical which are trained only on the transcripts of Bangor-Miami corpus. It can be observed from the result that phoneme mapping can improve the performance by 0.3 absolute WER, so we fix it as our acoustic model and the only difference among the experiments on the bottom block is that they have different synthetic texts for language modelling. *translation* model denotes that the language model is trained on the transcripts as well as the translation of them and + symbol describes with what techniques, code-switched text have been generated and added to the training text. We interpolate the LM with models trained on external text data [6] to show that simply using larger but out-of-domain text data doesn't help improve the performance. After obtaining the word alignments of parallel sentences, we compare the results of generating the code-switched text by simply random replacement of the aligned words, or ranking the possible replacements with the POS tags of the preceding/current/following words or the acceptance under EC theory.

We can find that POS and EC have similar improvement on WER, while the combination of them shows their advantages cannot directly add up. One possible explanation can be that the implementation of EC is based on the constituency parse, which is also heavily related to the POS tags. Therefore, the linguistic information implied by them are overlapped with each other and only little improvement can be achieved when combined. Our model with the best performance uses all of linguistic information we discussed before, with approximately 2% improvement on both WER and PPL.

6 Conclusions

In this paper, we present a framework for code-switched ASR task. By using phonological features for phoneme mapping, and POS tags and EC theory for more more natural code-switched text generation, we eventually achieve 2% improvement on PPL as well as WER. It should be noted that although the experiments are conducted on Bangor-Miami corpus, there are no language specific constraints with this approach, which shows

[6] Here, we use TED talk subtitles to train the LMs for English and Spanish (Tiedemann, 2012).

a potential to be extended to cover more language pairs. As a future work, we would like to compare the performances of different linguistic theories in our proposed framework, which can serve as an indirect validation of their influence on different language pairs. Also, motivated by natural distributions of linguistic structures, exploring different sampling techniques can also be promising.

7 Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh.

References

Ana Inés Ansaldo, Karine Marcotte, Lilian Scherer, and Gaelle Raboyeau. 2008. Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *Journal of Neurolinguistics*, 21(6):539–557.

Peter K Austin and Julia Sallabank. 2011. *The Cambridge handbook of endangered languages*. Cambridge University Press.

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.

K Bhuvanagiri and Sunil Kopparapu. 2010. An approach to mixed language automatic speech recognition. *Oriental COCOSDA, Kathmandu, Nepal*.

Ellen Bialystok, Fergus Craik, and Gigi Luk. 2012. Bilingualism: Consequences for mind and brain. *Trends in cognitive sciences*, 16:240–50.

Margaret Deuchar. 2011. The miami corpus: Documentation file.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*.

Harold T. Edwards. 1992. *Applied Phonetics: The Sounds of American English*. Singular Publishing Group.

Björn Gambäck and Amitava Das. On Measuring the Complexity of Code-Mixing.

Brian Goldstein. 2000. *Resource guide on cultural and linguistic diversity*. Singular resource guide series. Singular Pub. Group, San Diego, Calif.

F Grosjean. 2010. *Bilingual: life and reality*. Harvard University Press.

François Grosjean. 1982. *Life with two languages: An introduction to bilingualism*. Harvard University Press.

John J. Gumperz. 1982. *Discourse Strategies*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Yifan He, Yanjun Ma, Andy Way, and Josef Genabith. 2010. Integrating n-best smt outputs into a tm system. *He, Yifan and Ma, Yanjun and Way, Andy and van Genabith, Josef (2010) Integrating N-best SMT outputs into a TM system. In: COLING 2010 - 23rd International Conference on Computational Linguistics, 23-27 August 2010, Beijing, China.*

Roberto Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science - CURR DIRECTIONS PSYCHOL SCI*, 10:164–168.

Janet Holmes and Nick Wilson. 2017. *An Introduction to Sociolinguistics*, 5 edition. Routledge, London.

Nikita Kitaev and Dan Klein. 2018. Multilingual constituency parsing with self-attention and pre-training. *CoRR*, abs/1812.11760.

J. Kohler. 1998. Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 1, pages 417–420 vol.1.

Carrie Leung. 2006. Codeswitching in print advertisements in hong kong and sweden.

Ke Li, Jinyu Li, Guoli Ye, Rui Zhao, and Yifan Gong. 2019. Towards Code-switching ASR for End-to-end CTC Models. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6076–6080. ISSN: 2379-190X.

Ying Li and Pascale Fung. 2013. Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7368–7372.

Ying Li and Pascale Fung. 2014. Code switch language modeling with functional head constraint. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4913–4917.

Ne Luo, Dongwei Jiang, Shuaijiang Zhao, Caixia Gong, Wei Zou, and Xiangang Li. 2018. Towards end-to-end code-switching speech recognition. *CoRR*, abs/1810.13091.

Dau-Cheng Lyu, Ren-yuan Lyu, Yuang-chin Chiang, and Chun-nan Hsu. 2006. Speech recognition on code-switching among the chinese dialects. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

Tetyana Lyudovyk and Valeriy Pylypenko. 2014. Code-switching speech recognition for closely related languages. In *Spoken Language Technologies for Under-Resourced Languages*.

Gurunath Reddy Madhumani, Sanket Shah, Basil Abraham, Vikas Joshi, and Sunayana Sitaram. 2020. Learning not to discriminate: Task agnostic learning for improving monolingual and code-switched speech recognition.

Roy C. Major. 2002. The bilingualism reader. li wei (ed.). london: Routledge, 2000. *Studies in Second Language Acquisition*, 24(3):491–493.

Metilda Sagaya Mary N J, Vishwas M. Shetty, and S. Umesh. 2020. Investigation of methods to improve the recognition performance of tamil-english code-switched data in transformer framework. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7889–7893.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Pieter Muysken, Pieter Cornelis Muysken, et al. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

C. Myers-Scotton. 1997. *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press.

Carol Myers-Scotton. 1989. Codeswitching with english: types of switching, types of communities. *World Englishes*, 8(3):333–346.

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.

David Sankoff and Shana Poplack. 1981. A formal grammar for code-switching. *Paper in Linguistics*, 14(1):3–45.

Sanket Shah, Basil Abraham, Gurunath Reddy M, Sunayana Sitaram, and Vikas Joshi. 2020. Learning to recognize code-switched speech without forgetting monolingual speech recognition.

Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie. 2019. Investigating end-to-end speech recognition for mandarin-english code-switching. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6056–6060.

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.

Sunit Sivasankaran, Brij Mohan Lal Srivastava, Sunayana Sitaram, Kalika Bali, and Monojit Choudhury. 2018. Phone merging for code-switched speech recognition. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics.

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The Role of Cognate Words, POS Tags and Entrainment in Code-Switching. In *Interspeech 2018*, pages 1938–1942. ISCA.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Jochen Weiner, Ngoc Thang Vu, Dominic Telaar, Florian Metze, Tanja Schultz, Dau-Cheng Lyu, Eng-Siong Chng, and Haizhou Li. 2012. Integration of language identification into a recognition system for spoken conversations containing code-switches. In *Spoken Language Technologies for Under-Resourced Languages*.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung. 2020. Meta-transfer learning for code-switched speech recognition. *CoRR*, abs/2004.14228.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. *CoRR*, abs/1909.08582.

Donald Winford. 2003. *An introduction to contact linguistics / Donald Winford.* Language in society (Oxford, England) ; v. 33. Blackwell Pub.

Chung Hsien Wu, Yu Hsien Chiu, Chi Jiun Shia, and Chun Yu Lin. 2006. Automatic segmentation and identification of mixed-language speech using delta-bic and lsa-based gmms. *IEEE Transactions on Speech and Audio Processing*, 14(1):266–275.