AN EVALUATION BENCHMARK FOR AUTOMATIC SPEECH RECOGNITION OF GERMAN-ENGLISH CODE-SWITCHING

Abbas Khosravani[1], *Philip N. Garner*[1], *Alexandros Lazaridis*[2]

[1]Idiap Research Institute, Switzerland
[2]Data, Analytics & AI Group — Swisscom AG, Switzerland

ABSTRACT

Code-switching arises when a (typically multilingual) speaker changes language during an utterance. This linguistic phenomenon causes problems for automatic speech recognition as the models are typically monolingual. In this work, we present a code-switching evaluation scenario for German-English that is created by resegmenting the German Spoken Wikipedia Corpus. Since these articles span a wide variety of (often technical) topics, they include a lot of borrowing and code-switching phenomena. The resulting corpus consists of around 34 hours of intra-sentential switches. We investigate end-to-end approaches using both monolingual and multilingual automatic speech recognition as well as language modeling to address the code-switching scenario. Results suggest that multilingual sequence-to-sequence approaches are to be preferred for code-switching thanks to the power of the attention mechanism. The segments are made available to the community as a benchmark.

Index Terms— Automatic Speech Recognition, Code-Switching, German, Multilingual, Benchmark

1. INTRODUCTION

The number of multilingual people in the world continues to grow [1]. *Code-switching* (CS, sometimes also referred to as code-mixing) is the use of elements from more than one language in the same utterance, such as in the German-English *'Für HEAVEN'S Willen!'* ('For HEAVEN'S sake!') and is a vital and widespread form of language use in bilingual speakers. The production of code-switches can be influenced by the properties of the words and the spoken context, typically more technical or international, and by the speakers' relative proficiency in both languages. This phenomenon poses a significant challenge for automatic speech recognition (ASR) systems.

We are interested in general in ASR for German and specifically Swiss German. Switzerland is a multilingual nation with four recognized national languages. German is the most widely spoken language, spoken by 5 million people or 63.5% of the population. Switzerland's linguistic communities are connected to specific territories; there exist towns with one official language where the neighboring town has a different one. Proficiency in the non-native national languages varies, and in recent years English is becoming the main foreign language taught in schools in many cantons. This is also due to globalization and the presence of more foreign companies in the big cities. The concomitant impact of English on German gives rise to the study of English code-switching in German.

On a syntactic level, code-switching is divided into intra-sentential and inter-sentential units. Typical examples of intra-sentential switches are phrasal elements from the embedded language (English) that occur in a matrix language (German) sentence as in *Berlin sei eben 'the place to be', erklärt ein Banker* ('Berlin is just the place to be, says a banker'). Inter-sentential code-switching, on the other hand, can be defined as grammatically complete English sentences which are added as non-obligatory clauses to a German sentence or occur outside the textual space of a German sentence, as in *'When in Rome, do as the Romans do.' Dieses englische Sprichwort drückt eine Binsenweisheit aus* (' "When in Rome, do as the Romans do". This English proverb expresses a truism'). Due to the larger acoustical variations of mixed languages within utterances, intra-sentential code-switching is much more difficult for an ASR system [2]. In this study, we report on the development of a German-English code-switching corpus based on the German Spoken Wikipedia Corpus (SWC) [3]. SWC is a large collection of speech data read by volunteers covering a broad variety of Wikipedia topics. Owing to the encyclopedic nature of the articles and the diverse range of technical and scientific topics, they include a large amount of borrowing and code-switching phenomena. The word-level alignment provided in this corpus allows us to extract segments with intra-sentential English code-switching and develop an evaluation scenario that can be used as a benchmark for research on code-switching speech recognition.

The rest of the paper is organized as follows. Sec. 2, introduces related work on code-switched ASR systems. Sec. 3, describes our German-English code-switching corpus and the data used for system development. The benchmark experiments are presented in Sec. 5 with conclusion in Sec. 6.

2. RELATED WORK

Unfortunately, due to the lack of available resources for English code-switching in German, there are very few studies in the literature. In [4], a German-English code-switching speech dataset was collected and studied; however, the domain and the quality of data is very limited (a digitized version of original audio recording collected from German-speaking Jewish refugees in 1993). There are several corpora with English as the embedding language and either Mandarin [5], Spanish [6], Hindi [7] and Cantonese [8] as the matrix language, but to the best of our knowledge, there is no reported corpus for German.

Automatic speech recognition (ASR) of an intra-sentential code-switched utterance is challenging and there is very little work on end-to-end (E2E) ASR [2]. Prior work mainly uses hybrid ASR systems such as [9, 10]. However, by the recent progress in E2E automatic speech recognition, they are becoming increasingly popular while achieving promising results on various ASR benchmarks [11, 12]. Since E2E ASR enables lexicon-free recognition, it has an advantage over the traditional hybrid system, especially for German which is characteristically highly inflected with a large vocabulary [11]. The main approach, in this case, is to train a single bilingual acoustic-language model for both languages. However, if the two languages do not have much in common, it may not be the best option for achieving the best performance in each language. It has been shown in [2, 13] that language identification (LID) can be beneficial in identifying the English code-switching in Chinese utterances. In [2] they separately trained a LID to directly adjust the posteriors of the multilingual E2E Connectionist Temporal Classification (CTC) model with the posteriors of the LID. Similarly, in [14] the authors highlight the importance of using language identification in the Mandarin-English code-switching scenario. In [2] the authors argue that an encoder-decoder based model cannot work well for code-switching scenarios as the output of the decoder depends on the previous outputs and that a CTC model is more desirable owing to the output independence assumption. In more recent work, [15] proposes a multi-encoder-decoder (MED) transformer architecture with two language-specific symmetric encoder branches and corresponding attention modules in the decoder. The authors showed that this framework can exploit the discrimination between the mixed languages, and alleviates the code-switching training data scarcity problem.

3. DATA

3.1. Evaluation

We develop a corpus for English code-switching in German, a 34h transcribed speech corpus of read Wikipedia articles which can be used as a benchmark for research on code-switching[1]. The articles are read by a large and diverse group of people. The code-switching speech segments are extracted from the German SWC [16, 3], perhaps the largest corpus of freely-available aligned speech for German. It contains 1014 spoken articles read by more than 350 identified speakers comprising 386h of speech. In SWC, since most of the articles are long, the recordings submitted by the volunteers are also long ($\sim$54min) on average. These audio files are manually annotated at word-level and also segment level in XML format. We use a language identification tool [17] to detect code-switching in the transcription of the audio files with consecutive indices[2]. To extract intra-sentential code-switching segments, we ensure that the detected code-switching is preceded and followed by German words or sentences. The final set consists of 34h of speech data and 12,511 code-switching segments. Apart from this, we also use a different set by extracting segments from SWC with no code-switching. This set contains 77h of speech data and 50,069 segments and is useful for comparison and benchmarking. Moreover, we report system performance on the German evaluation set of the *Common Voice* (CV) corpus [18] which is a multilingual collection of transcribed speech data collected and validated using crowdsourcing. The test set contains 25h of speech data from 4,378 individuals as 15,341 utterances.

3.2. Acoustic Model

For system development, we used different speech corpora. Compared to, say, English, there are relatively few speech corpora available for German. Fortunately, some efforts have been made recently to collect and contribute such resources for sustainable research [3, 19, 18, 20]. The *M-AILABS* resource was distributed by Munich Artificial Intelligence Laboratories[3] under a non-restrictive license and comprises hundreds of hours of speech audio in nine different languages taken from non-professional audio-books of the LibriVox project[4]. We also use the LibriSpeech corpus [21] which is part of the LibriVox project [19] with 460h of clean English speech data. VoxForge[5] is also another open speech dataset in various languages that was set up to collect transcribed speech from participant under uncontrolled conditions. Table 1 provides more information on the corpora used for our system development.

3.3. Language Model

Unlike a large amount of monolingual data available to train a language model, transcribed code-switching data necessary

[1] https://www.idiap.ch/en/dataset/code-switching

[2] We observed that the word-level alignment is not provided for every word

[3] https://www.caito.de/2019/01/the-m-ailabs-speech-dataset

[4] https://librivox.org

[5] http://www.voxforge.org

Table 1. *Statistics on the training speech data used for system development.*

Corpus	Lang	Dur(h)	Segments	Speakers
Common Voice	EN	695	435,909	15,883
	DE	314.9	196,404	2,818
VoxForge	EN	91.9	68,701	2,707
	DE	56.8	41,371	111
M-AILABS	EN	142	69,505	4
	DE	233.6	118,385	–
LibriSpeech	EN	460	132,553	2,484

for training is hard to come by. In this work, we aim at analyzing the effect of the language model on ASR in a CS scenario. To achieve this, we use different text corpora with CS sentences and without it. Table 2 provides some statistics on these corpora. Due to the diverse topics in the German Wikipedia corpus[6], it includes lots of code-switching and borrowing words from English. We used a language identification tool [17] to extract CS sentences from the Wikipedia corpus that results in 2.8M German sentences. We subtract this from the Wikipedia corpus to analyze the effect of CS text on the performance of ASR. To avoid overlap with the evaluation scenario, all the articles used in the SWC have been removed from the Wikipedia text corpus. In addition to Wikipedia, we also used news articles from the monolingual language model training data used in WMT18[7]. It includes German text crawled from online news in 2017, with the markup stripped out and sentences shuffled.

Table 2. *Statistics for the amount of training data used in language modeling. The number of sentences, words and distinct words (case-insensitive) is reported.*

Corpus	Sentences	Words	Distinct words
Wikipedia without CS	12.1M	499M	5.6M
Wikipedia with CS	2.8M	115M	1.3M
News	39.0M	603M	4.1M

4. SPEECH RECOGNITION SYSTEM

We use wav2letter++, an ASR framework designed from the outset to support end-to-end paradigms [22] which is now has consolidated into *Flashlight* [8]. It supports several end-to-end approaches including sequence-to-sequence models with attention (Seq2Seq) [12] and Connectionist Temporal Classification (CTC) [23]. Unlike the Seq2Seq approach, CTC does not use any specific decoder network and makes a conditional independence assumption. The model assumes that every output is conditionally independent of the other outputs given the input. Although this seems to be a shortcoming, it makes it more desirable for CS scenarios as the current output step does not explicitly rely on previous outputs. To be able to test this hypothesis, we conduct experiments using both Seq2Seq which has an encoder-decoder architecture, and CTC using the same encoder architecture. The network architecture is based on time-depth separable (TDS) convolution blocks [11]. In [24], it was shown that this TDS convolution block generalizes much better than other deep convolutional architectures and requires fewer parameters to train. This generalization is mainly due to some form of regularization, including dropout, label smoothing [25] and subword regularization [26, 11]. Subword-level ASR systems outperform both the character and word-level ones in the absence of a lexicon. We incorporate the unigram language model [26], which is a probabilistic approach to generate multiple subword segmentation. This in turn is essential for subword regularization as to improve the generalization and robustness of the ASR system to segmentation error [27]. For the encoder network, we use 12 TDS blocks with dropout and kernel size of 21×1 in three groups and set the number of channels in each group to (10, 14, 18) resulting in 39M parameters. We use a key-value attention [11] mechanism and an encoder of dimension 512. The model is trained using both CTC and Seq2Seq criteria using gradient descent. We also use 80-dimensional log-mel features, computed with a 25ms window and 10ms frameshift.

We select the best transcription by leveraging both the posteriors of an acoustic model (AM) and the probability of a language model (LM). We train multiple n-gram subword language models on different text corpora as described in Section 3.3 using the KenLM toolkit [28]. We use a lexicon-free beam-search decoder which utilizes a word separator which is predicted as a normal token and can also be part of a token to split the sequence of tokens into words. Therefore during training, there is no notion of words. The decoder uses a 6-gram subword LM to provide LM log-probability scores accumulated together with AM scores for a one-pass beam search decoding. We tune the language model weight on a validation set for each evaluation scenario. The validation set is designed as a small subset (10%) of the evaluation set.

5. EXPERIMENTS

5.1. Acoustic Model

The first experiment is designed to test the hypothesis that CTC can perform better than Seq2Seq in handling CS scenarios due to the conditional independence assumption on the output. We train a monolingual as well as a multilingual German-English system using each technique. For this experiment, we use the German news dataset to train a 6-gram subword LM. The results are given for different evaluation

[6] https://dumps.wikimedia.org

[7] http://www.statmt.org/wmt17/translation-task.html

[8] https://github.com/flashlight/flashlight

sets in Table 3.

Table 3. *Comparison of CTC and Seq2Seq models trained on both monolingual and multilingual speech data in terms of WER(%) on different evaluation sets. The LM is trained on German news articles and the perplexity for each set is also reported.*

		Monolingual		Multilingual	
Evaluation	LM_{ppl}	CTC	Seq2Seq	CTC	Seq2Seq
CS	105	33.2	32.8	30.6	**28.5**
SWC	72.1	23.9	22.7	23.5	**21.6**
CV	52.1	18.7	17.9	17.9	**15.1**

The results indicate a higher LM perplexity in the CS scenario compares to the others. This is mainly due to the lack of CS sentences in news articles and a different context with that of Wikipedia articles read in SWC. A multilingual model not only does not hurt the recognition of German utterances but also results in performance improvement in all scenarios. Knowledge transfer from English can be a good explanation for this. The Seq2Seq model obtained better performance compared to CTC. This is not expected due to the conditional independence assumption on the output [2]. However, this can be described by the power of the attention mechanism in the Seq2Seq model and maybe the high similarity of English as an embedding language to German rather than Chinese as reported in [2]. We observe a relative improvement of 13% in the CS scenario and more improvement of 15.6% in CV using the Seq2Seq technique.

5.2. Language Model

Language modeling in a CS scenario mainly suffers from the lack of adequate training material, as CS rarely occurs. In [29], it has been shown that incorporating CS text in the training of LM is beneficial in reducing WER in the CS scenario. CS text could either be generated automatically using a recurrent neural network (RNN) or by translating text from the embedded language to the matrix language. In this experiment, we want to test this hypothesis, but by using natural CS text from the Wikipedia corpus. We use the Wikipedia corpus both with CS sentences and with no CS sentences. Table 4 presents the results. From the results, it is clear that by incorporating CS text data, we can enrich the LM by lowering the perplexity in the CS set from 55.9 down to 42.1 and, as a result, improve the ASR performance by 3.4% relative. We also observe that using CS sentences in language modeling does not hurt the German speech recognition but also results in some improvement in CV set; perhaps as a result of data augmentation.

Table 4. *ASR performance on different evaluation scenarios using multilingual ASR models with the Seq2Seq criterion. We report results using LMs trained on Wikipedia with and without CS. The language model perplexity is also reported.*

	No LM	Without CS		With CS	
Evaluation	WER(%)	LM_{ppl}	WER(%)	LM_{ppl}	WER(%)
CS	32.7	55.9	26.4	42.1	**25.5**
SWC	25.0	41.5	**19.8**	40.6	20.0
CV	22.9	36.2	13.7	32.6	**13.1**

6. CONCLUSION AND FUTURE WORK

In this work, we introduce a benchmark dataset for English code-switching in German based on the Spoken Wikipedia Corpus. We also describe several techniques to improve the acoustic and language modeling of a code-switching ASR system. The sequence-to-sequence criterion provides superior performance to CTC in our benchmark experiments. Incorporation of code-switching text in language modeling provides a significant gain in ASR performance in a code-switching scenario. Exploring various methods to improve the LM either by generating code-switching sentences or translating English text to German as data augmentation techniques would be among future work to improve ASR in code-switching scenarios.

7. REFERENCES

[1] Tej K Bhatia and William C Ritchie, *The handbook of bilingualism*, John Wiley & Sons, 2008.

[2] Ke Li, Jinyu Li, Guoli Ye, Rui Zhao, and Yifan Gong, "Towards code-switching ASR for end-to-end CTC models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6076–6080.

[3] Timo Baumann, Arne Köhn, and Felix Hennig, "The spoken wikipedia corpus collection: Harvesting, alignment and an application to hyperlistening," *Language Resources and Evaluation*, vol. 53, no. 2, pp. 303–329, 2019.

[4] Eva Maria Eppler, *The syntax of German-English code-switching.*, Ph.D. thesis, University of London, 2005.

[5] Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li, "Seame: a Mandarin-English code-switching speech corpus in south-east asia," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[6] Alfredo Ardila, "Spanglish: an anglicized Spanish dialect," *Hispanic Journal of Behavioral Sciences*, vol. 27, no. 1, pp. 60–81, 2005.

[7] Anik Dey and Pascale Fung, "A Hindi-English code-switching corpus.," in *LREC*, 2014, pp. 2410–2413.

[8] Joyce YC Chan, PC Ching, and Tan Lee, "Development of a Cantonese-English code-mixing speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[9] Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li, "An analysis of a Mandarin-English code-switching speech corpus: SEAME," *Age*, vol. 21, pp. 25–8, 2010.

[10] Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang, and Chun-Nan Hsu, "Speech recognition on code-switching among the Chinese dialects," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 2006, vol. 1, pp. I–I.

[11] Awni Hannun, Ann Lee, Qiantong Xu, and Ronan Collobert, "Sequence-to-sequence speech recognition with time-depth separable convolutions," *Proc. Interspeech 2019*, pp. 3785–3789, 2019.

[12] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, 2015, pp. 577–585.

[13] Ne Luo, Dongwei Jiang, Shuaijiang Zhao, Caixia Gong, Wei Zou, and Xiangang Li, "Towards end-to-end code-switching speech recognition," *arXiv preprint arXiv:1810.13091*, 2018.

[14] Xian Shi, Qiangze Feng, and Lei Xie, "The ASRU 2019 Mandarin-English code-switching speech recognition challenge: Open datasets, tracks, methods and results," *arXiv:2007.05916*, 2020.

[15] Xinyuan Zhou, Emre Yılmaz, Yanhua Long, Yijie Li, and Haizhou Li, "Multi-encoder-decoder transformer for code-switching speech recognition," *arXiv preprint arXiv:2006.10414*, 2020.

[16] Arne Köhn, Florian Stegen, and Timo Baumann, "Mining the spoken wikipedia for speech data and beyond," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4644–4647.

[17] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov, "Learning word vectors for 157 languages," in *Language Resources and Evaluation Conference*, 2018, number CONF.

[18] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[19] "LibriVox: Free public domain audiobooks," Jan. 2014.

[20] Stephan Radeck-Arneth, Benjamin Milde, Arvid Lange, Evandro Gouvêa, Stefan Radomski, Max Mühlhäuser, and Chris Biemann, "Open source German distant speech recognition: Corpus and acoustic model," in *International Conference on Text, Speech, and Dialogue*. Springer, 2015, pp. 480–488.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[22] Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert, "Wav2letter++: A fast open-source speech recognition system," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6460–6464.

[23] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[24] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1243–1252.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[26] Taku Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 66–75.

[27] Jennifer Drexler and James Glass, "Subword regularization and beam search decoding for end-to-end automatic speech recognition," in *ICASSP 2019-2019 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 6266–6270.

[28] Kenneth Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*. Association for Computational Linguistics, 2011, pp. 187–197.

[29] E Yilmaz, H Heuvel, and DA van Leeuwen, "Acoustic and textual data augmentation for improved ASR of code-switching speech," in *Proceedings of Interspeech*. Hyderabad, India: ISCA, 2018, pp. 1933–1937.