Back-translation Approach for Code-switching Machine Translation: A Case Study

Maraim Masoud[0000−0003−0503−7577], Daniel Torregrosa, Paul Buitelaar[0000−0001−7238−9842], and Mihael Arčan[0000−0002−3116−621X]

Insight Centre for Data Analytics
Data Science Institute
National University of Ireland Galway
name.surname@insight-centre.org

Abstract. Recently, machine translation has demonstrated significant progress in terms of translation quality. However, most of the research has focused on translating with pure monolingual texts in the source and the target side of the parallel corpora, when in fact code-switching is very common in communication nowadays. Despite the importance of handling code-switching in the translation task, existing machine translation systems fail to accommodate the code-switching content. In this paper, we examine the phenomenon of code-switching in machine translation for low-resource languages. Through different approaches, we evaluate the performance of our systems and make some observations about the role of code-mixing in the available corpora.

Keywords: Machine-translation · Code-switching · Back-translation

1 Introduction

The popularity of social media platforms creates an opportunity for multi-lingual speakers and language learners to alternate between one or many languages. This results in a new form of a hybrid language form called code-mixed language. Code-mixing[1] is defined as "the embedding of linguistic units such as phrases, words, and morphemes of one language into an utterance of another language" [25]. The phenomenon is commonly observed in multilingual communities [28] and usually employed for different communication purposes such as asking questions [9], swearing [21], expressing emotions [43], and content clarification [13]. An example of a code-mixing, as shown in [34], is presented in Table 1.

Studies have linked many triggers for the use of mixed-code in speech and writing such as metaphorical switching, situational switching and lexical borrowing [6]. The phenomenon presents itself prominently in user-generated contents, especially with low-resource languages. Consequently, there is a growing need for translating code-mixed hybrid language into standard languages. Thus, automatic machine translation has been an important task for this phenomenon.

[1] The terms "code-mixing" and "code-switching" are used interchangeably in the machine translation sub-field.

Source Sentence(ES):	I put the fork en la mesa
Translation Sentence(EN):	I put the fork on the table

Table 1. The Spanish sentence is code-mixed with the English phrase 'I put the fork' creating what is known as *Spanglish*. The second sentence represents the English translation of the the first sentence.

However, due to the lack of parallel data for low-resourced scenarios where code-switching is very common, existing machine translation systems fail to properly handle code-mixed text.

The current neural machine translation (NMT) using a sequence to sequence translation framework [40] has achieved impressive results in recent years [44]. One of the key innovations that led to this advancement is the introduction of the attention mechanism [3, 24]. While being able to approach human-level translation [31], NMT still requires a huge amount of parallel data. Such data might not always be available for low-resource languages. As monolingual data is easily available, one way to utilize them for the machine translation task is following back-translation [36]. This technique is used to leverage monolingual data during the training. It is an inverse target-to-source translation approach which generates synthetic source sentences by translating monolingual sentences of the target language into the source language with a pre-existing target-to-source translation model. These pseudo-source sentences together with the original target sentences are then concatenated to the original parallel corpus to train a new source-to-target MT system.

Current machine translation systems do not support code-mixed text, and are only designed to work with a monolingual language in both ends of the translation system [7]. This limitation makes it unsuitable to rely on current NMT systems for daily communications where code-mixed language is prevalent.

This paper presents code-mixed machine translation for Tamil-English language pair; however, this context is very common with other languages.

2 Related Work

The code-switching behaviour has been investigated from different perspectives [2] and for many languages [23, 1]. Early work in this domain focused on exploring the phenomenon from linguistics and sociolingusitic perspectives, and then move towards investigating it computationally for NLP applications [17, 8]. Recently, code-mixed languages have seen a lot of interest in downstream NLP tasks such as part of speech tagging [39, 14], named entity recognition [46, 1], dependency parsing [29]. Additionally, the phenomenon has also been considered for NLP applications such as sentiment analysis [22, 42, 16], machine translation [37, 15], and question answering [12, 9]. Despite all the research attempts, code-mixing still presents serious challenges for the natural language processing community

[7]. The noticeable lack of resources such as annotated corpora and NLP tools continue to pose a challenge and reduces the chances of improving [7, 12].

A lot of work has been done on machine translation for non-code-switching cases [35, 3, 24]. However, there is relatively little work focus on mixed language machine translation. Sinha et al. [37] performed cross morphological analysis to handle code-mixed translation task from Hinglish into both pure English and pure Hindi. The work of Johnson et al. [15] on Google's multilingual zero-shot translation handles code-switching phenomenon. In their work, they show that the model can represent multiple languages in the same space. However, the results are not as good as monolingual inputs. As opposed to monolingual inputs, the lack of gold standard parallel data has significantly contributed to the minimum research in code-switch translation.

Back-translation has been proposed as a corpus augmentation technique which has been widely used to expand parallel corpora for machine translation tasks [45, 36]. It is a way to leverage monolingual data without modifying the translation model. It requires a counter loop of training (target-to-source) to generate synthetic parallel data from the target data [11]. The idea dates back to statistical machine translation [5]. Recently, back-translation has been widely adopted for neural machine translation systems [36, 45] and shown to be beneficial when training data is scarce as in low-resource languages scenarios [18, 30, 41]. Currey et al. [10] and Karakanta et al. [18] show how synthetic data can improve low-resource language pairs. While the former applies a single round of back-translation, where the source is a copy of the monolingual target data, the latter tries multiple rounds of back-translation.

A comparative analysis on the effect of synthetic data on NMT is demonstrated by Park et al. [27] and Poncelas et al. [30]. In the work of Park et al. the models trained only with synthetic data. Then, the performance was evaluated with models trained with parallel corpora composed of: (i) synthetic data in the source-side only; (ii) synthetic data in the target side only; and (iii) a mixture of parallel sentences of which either the source-side or the target-side is synthetic. In the work of Poncelas et al., the NMT model was trained with three different parallel corpora: A synthetic (source side only), a synthetic (target side only), and a mixture in either source or target side.

3 Methodology

To tackle the code-switching issue in the translation task, we were inspired by the evaluation pipeline introduced by Poncelas et al. [30] on testing the impact of back-translation. We evaluate three different approaches. Each approach is deployed with a different NMT model and a different dataset variation; a dataset with original translation, a dataset with hybrid back-translated (synthetic) data, and only monolingual source and target (no-code-mixing) dataset. The three approaches are:

– **Baseline approach**: In this approach, the NMT model is trained using the original dataset in its base form, without any modification, with the

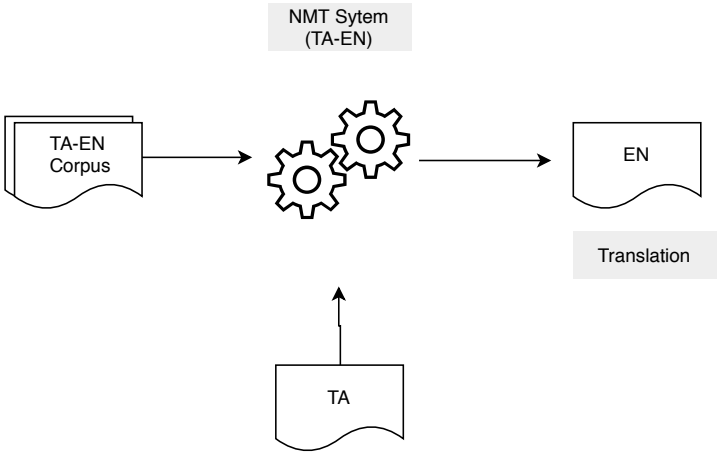

Fig. 1. Baseline approach: The illustration shows the translation from Tamil (TA) into English (EN).

exception to the standard pre-processing steps (tokenization, lowercasing and cleaning). In this setting, the code-mixed tokens are kept without any modification. The model in this approach serves as a baseline for comparison. Figure 1 shows a diagram for the baseline approach.

- **Hybrid approach**: The models in this approach are built with original sentence pairs combined with back-translated code-mixed tokens. In this setting, the corpus is modified with two variations of the back-translated data. Firstly, the English tokens (code-mixed tokens) are identified and extracted from Tamil sentences. These tokens are then translated using different translation models: (a) the baseline model, (b) Google Translate[2]. Upon generating two versions of Tamil translations of these English tokens, these tokens are injected back to the Tamil sentences to create Tamil only sentences on the target side of the corpus. Thus, the final resulted corpus containing original and synthetic (back-translated corpus) is then used to train our model. A visualization of the pipeline for this approach is illustrated in Figure 2.
- **Monolingual (no-code-mixing) approach**: In this approach, the NMT model is trained on a refined corpus in which the code-mixing tokens (English tokens in the Tamil side) are identified and removed, creating a monolingual data for the Tamil side. Figure 3 shows the pipeline for this approach.

[2] translate.google.com retrieved February 2019.

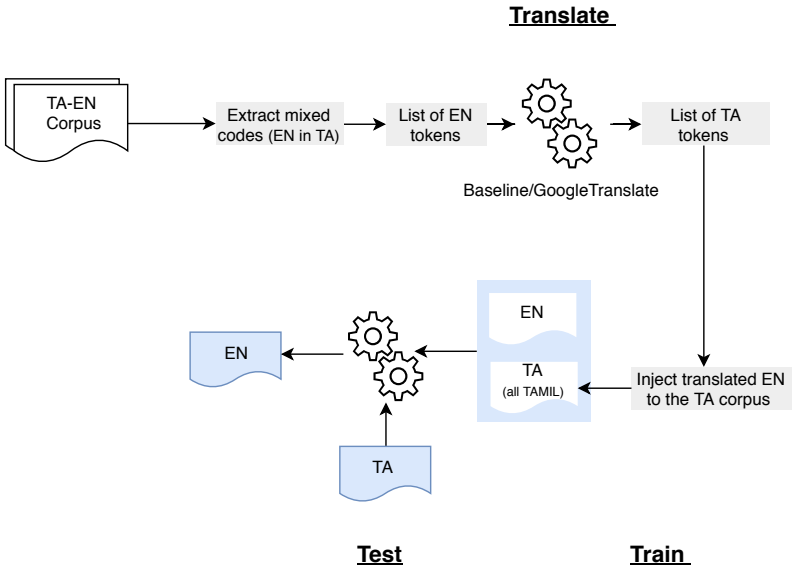

Fig. 2. Hybrid approach: Pipeline for back-translated code-mixed model.

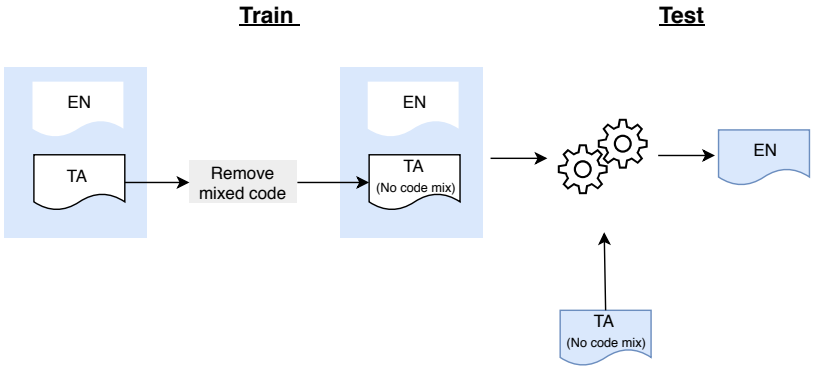

Fig. 3. Monolingual approach: Pipeline for No-code Mixing.

4 Experimental Setting

In this section, we describe the dataset as well as the framework used to train and evaluate the approaches.

4.1 Data Description

For the scope of this work, we use a parallel corpus of code-mixed English-Tamil and English. The choice of adding this particular dataset was influenced by the availability of a public dataset for this task. Additionally, as the code-mixing

	Tamil		English	
	Tokens	**Lines**	**Tokens**	**Lines**
Train	906,391	159,182	966,911	159,182
Validation	11,446	2,000	11,725	2,000
Evaluation	12,016	2,000	12,873	2,000

Table 2. Statistics of the Tamil-English Corpus.

phenomenon is well-noticed among Indian language speakers, the code-mixing dataset for English-Tamil was selected for this task. The dataset was combined from OPUS[3] and EnTam[4]. Although OPUS has a large dataset for Tamil ↔ English pairs, we excluded some data due to encoding issues. The final dataset is cleaned, shuffled, tokenized and lowercased using the OpenNMT toolkit[5]. In total, the dataset contains 163,182 sentences and around 129,710 English tokens in the Tamil side of the corpus. Table 2 shows a breakdown of the number of tokens and sentences in the English and the Tamil sides of the corpus.

To study the effects of code-mixing in the translation, different data settings have been used in the training and the evaluation of the NMT models. The dataset variations are original data, hybrid (original mixed with synthetic – back-translated) data, and monolingual data in both sides. These settings allow us to reuse the code-mixed tokens as well as observe their role in the demonstrated corpus.

An NMT model has been built for each dataset variation as explained in Section 3. These different configuration scenarios allow us to trace the quality of the code-switching translations as well as its role in conversation.

4.2 NMT Framework

The experiment was performed using the OpenNMT [19], which is a generic deep learning framework based on sequence to sequence models. The framework is used for a variety of NLP tasks including machine translation. We deployed the framework in its default setting: two hidden layers, 500 hidden LSTM (Long Short Term Memory) units per layer, 13 epochs, batch size of 64, and 0.3 dropout probability and word embeddings of 500 dimension. To compensate for the limited vocabulary issue, the Byte Pair Encoding (BPE) [35], which is a form of byte compression, was used with the following parameters: a maximum vocabulary size of 50,000 subwords and a maximum of 32,000 unique BPE merge operations. For each approach mentioned above, word and BPE translation models were trained.

[3] `http://opus.nlpl.eu/` retrieved February 2019

[4] `http://ufal.mff.cuni.cz/~ramasamy/parallel/html/#download` retrieved January 2019

[5] `http://opennmt.net/OpenNMT` used on February 2019

4.3 Evaluation Metrics

The performance of the different approaches was evaluated using different translation evaluation metrics: BLEU [26], TER [38], METEOR [4] and chrF [32]. BLEU (Bilingual Evaluation Understudy) is an automatic evaluation that boasts high correlation with human judgements, and METEOR (Metric for Evaluation of Translation with Explicit ORdering) is based on the harmonic mean of precision and recall. ChrF is a character n-gram metric, which has shown very good correlations with human judgements especially when translating to morphologically rich(er) languages. Finally, translation error rate (TER) [38] is a metric that represents the cost of editing the output of the MT systems to match the reference. High score of BLEU, METEOR, and Chrf means the system produces a highly fluent translation, but a high score of TER is a sign of more post-editing effort and thus the lower the score the better. Additionally, we used bootstrap resampling [20] with a sample size of 1,000 and 1,000 iterations, and reported statistical significance with $p < 0.05$.

5 Results and Discussion

This section describes the quantitative and qualitative results of the four models; the baseline (Baseline), the hybrid model with baseline back-translated tokens (Hybrid-Baseline), the hybrid model with Google back-translated tokens (Hybrid-Goggle), and monolingual model with no-code-mixing (Monolingual).

5.1 Quantitative Results

We report the performance of the different models using the following metrics: BLEU, Meteor, TER and ChrF. The quantitative evaluation for Tamil $\leftrightarrow$ English is presented in Table 3. All models slightly outperform the no-code-mixing models, which reports a decrease of ~1 BLEU point for the Tamil $\rightarrow$ English translation direction and ~5 BLEU points in English $\rightarrow$ Tamil direction. This suggests that by removing the code-switching tokens, the model gets confused due to the ordering and misalignment; thus the drop in the score. The decline in the model performance after removing the code-mixed tokens can be related to the high performance shown in the back-translated code-mixed models. These models report the best performance in both translation direction. The Hybrid Google back-translated model for English into Tamil translation reports 24.65 and 25.28 BLEU points for the word and BPE based translation models, respectively. Our back-translated models reports lower results of 21.96 for the word-based model and 22.53 for the BPE-based model.

In the case of translating Tamil text into English, where the code-mixing takes place, the BPE baseline performed best, followed by close results in terms of BLEU and METEOR for the back-translated models, whereas our model (21.93) performed similarly to Google Translate (21.35) for the BPE based model. From the results, we observed that the approaches with the back-translated models

	Tamil→English				English→Tamil			
Model	**BLEU**	**METEOR**	**ChrF**	**TER**	**BLEU**	**METEOR**	**ChrF**	**TER**
Word-Baseline	20.57	24.69	42.46	0.68	16.14	19.49	60.63	0.74
Word-Monolingual	19.46	23.99	42.28	0.69	16.65	24.10	70.39	0.74
Word-Hybrid-Baseline	21.05	24.60	42.92	0.68	21.96	21.51	70.16	0.86
Word-Hybrid-Google	21.85	24.48	42.73	0.68	24.65	23.35	72.48	0.69
BPE-Baseline	22.46	25.50	44.14	0.66	18.47	20.17	68.33	0.97
BPE-Monolingual	19.99	24.36	42.43	0.68	16.98	25.16	71.15	0.73
BPE-Hybrid-Baseline	21.93	25.39	44.55	0.66	22.53	23.34	71.63	0.82
BPE-Hybrid-Google	21.35	25.25	44.13	0.67	25.28	25.13	73.71	0.65

Table 3. Quantitative results for the evaluation of Tamil ↔ English. Results marked in blue are the best for BLEU metric, decayed in shade, darker blue is the best, the second best is lighter. The Average of all metrics are reported in this table.

outperform both baseline and the no-code-mixing models. From this experiment, we observed that the code-mixing tokens, as in our demonstrated corpus, play an important role in the meaning of the sentences. Thus, the drop in BLEU score is observed when code-mixing tokens are eliminated.

5.2 Qualitative Results

Table 4 analyses a sentence translated using all trained models from Tamil to English. The analysis showed that, all models did not correctly convey the message as in the reference sentence. This is due to small size of training data. Among all the results, the result from the baseline model seems to be the closest. Code-mixed tokens are incorrectly translated by the baseline and Google, thus injecting their translations creates a confusion and caused a failure in conveying the proper meaning of the message. Additionally, by removing the code-mixed tokens *stairwell* the model still provides a some context about meaning of the sentence.

Source	அவர் ஒரு மனிதன் stairwell இந்த புகைப்படத்தை எடுத்து பார்த்தேன் .
Reference	he saw a man take this photo of the stairwell .
Baseline	he saw a man take this photo into a man .
Hybrid-Baseline	and he took this photo **with** a man .
Hybrid-Google	he took this photo in a man **' s wall** .
Monolingual	he took this photo out of a man .

Table 4. Qualitative analysis of a sentence translated by all models for Tamil to English translation. Bold faced fragments are translating mistakes which are injected in the back-translation models.

6 Conclusion and Future Work

In this work, we explored the code-switching phenomenon in machine translation for a low-resourced scenario considering English-Tamil as our target language pair. We further investigated how back-translation can be used as a strategy to handle this phenomenon. The results show that the code-mixing part in this particular dataset potentially plays a supportive role. This can be observed by the little impact on the translated sentences when the code-mixing tokens are removed. This also explained by the slight improvement in the translation score (~1 BLEU points) when the code-mixing tokens (English tokens) in the Tamil side are correctly translated before training the models.

One future work will further investigate the role of code-switching in the available corpora. A second path will experiment with multilingual embedding as a preprocessing step for the translation of code-switched languages. This approach has already shown good performance in tasks such as sentiment analysis [33].

Acknowledgments

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under grant agreement number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund, and the Enterprise Ireland (EI) Innovation Partnership Programme under grant number IP20180729, NURS – Neural Machine Translation for Under-Resourced Scenarios.

References

1. Aguilar, G., AlGhamdi, F., Soto, V., Diab, M., Hirschberg, J., Solorio, T.: 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. pp. 138–147. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
2. Alex, B.: Automatic detection of english inclusions in mixed-lingual data with an application to parsing. The University of Edinburgh (2008)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2014)
4. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)

5. Bojar, O., Tamchyna, A.: Improving translation model by monolingual data. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 330–336. Association for Computational Linguistics, Edinburgh, Scotland (Jul 2011)
6. Boztepe, E.: Issues in code-switching: Competing theories and models. Working Papers in TESOL and Applied Linguistics **3**(2), 1–27 (2003)
7. Çetinoğlu, Ö., Schulz, S., Vu, N.T.: Challenges of computational processing of code-switching. In: Proceedings of the Second Workshop on Computational Approaches to Code Switching. pp. 1–11. Association for Computational Linguistics, Austin, Texas (Nov 2016)
8. Çetinoğlu, Ö., Schulz, S., Vu, N.T.: Challenges of computational processing of code-switching. In: Proceedings of the Second Workshop on Computational Approaches to Code Switching. pp. 1–11. Association for Computational Linguistics, Austin, Texas (Nov 2016)
9. Chandu, K.R., Chinnakotla, M., Black, A.W., Shrivastava, M.: Webshodh: A code mixed factoid question answering system for web. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 104–111. Springer (2017)
10. Currey, A., Miceli Barone, A.V., Heafield, K.: Copied monolingual data improves low-resource neural machine translation. In: Proceedings of the Second Conference on Machine Translation. pp. 148–156. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017)
11. Edunov, S., Ott, M., Auli, M., Grangier, D.: Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 489–500. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)
12. Gupta, V., Chinnakotla, M., Shrivastava, M.: Transliteration better than translation? answering code-mixed questions over a knowledge base. In: Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching. pp. 39–50 (2018)
13. Hidayat, T.: An analysis of code switching used by facebookers (a case study in a social network site) (2012)
14. Jamatia, A., Gambäck, B., Das, A.: Part-of-speech tagging for code-mixed English-Hindi twitter and Facebook chat messages. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. pp. 239–248. INCOMA Ltd. Shoumen, Bulgaria, Hissar, Bulgaria (Sep 2015)
15. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. Transactions of the Association for Computational Linguistics **5**, 339–351 (2017)
16. Joshi, A., Prabhu, A., Shrivastava, M., Varma, V.: Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 2482–2491. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016)
17. Joshi, A.K.: Processing of sentences with intra-sentential code-switching. In: Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics (1982)
18. Karakanta, A., Dehdari, J., Genabith, J.: Neural machine translation for low-resource languages without parallel corpora. Machine Translation **32**(1-2), 167–189 (Jun 2018)

19. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics **System Demonstrations**, 67–72 (2017)
20. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 conference on empirical methods in natural language processing (2004)
21. Lantto, H.: Code-switching, swearing and slang: The colloquial register of basque in greater bilbao. International Journal of Bilingualism **18**(6), 633–648 (2014)
22. Lee, S., Wang, Z.: Emotion in code-switching texts: Corpus construction and analysis. In: Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing. pp. 91–99. Association for Computational Linguistics, Beijing, China (Jul 2015)
23. Li, D.C.: Cantonese-English code-switching research in Hong Kong: A Y2K review. World Englishes **19**(3), 305–322 (2000)
24. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1412–1421. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015)
25. Myers-Scotton, C.: Common and uncommon ground: Social and structural factors in codeswitching. Language in society **22**(4), 475–503 (1993)
26. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
27. Park, J., Song, J., Yoon, S.: Building a neural machine translation system using only synthetic parallel data. CoRR (2017)
28. Parshad, R.D., Bhowmick, S., Chand, V., Kumari, N., Sinha, N.: What is India speaking? Exploring the "Hinglish" invasion . Physica A: Statistical Mechanics and its Applications **449**, 375 – 389 (2016)
29. Partanen, N., Lim, K., Rießler, M., Poibeau, T.: Dependency parsing of code-switching data with cross-lingual feature representations. In: Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages. pp. 1–17. Association for Computational Linguistics, Helsinki, Finland (Jan 2018)
30. Poncelas, A., Shterionov, D., Way, A., de Buy Wenniger, G.M., Passban, P.: Investigating backtranslation in neural machine translation. In: 21st Annual Conference of the European Association for Machine Translation. p. 249
31. Popel, M.: CUNI transformer neural MT system for WMT18. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 482–487. Association for Computational Linguistics, Belgium, Brussels (Oct 2018)
32. Popović, M.: ChrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015)
33. Pratapa, A., Choudhury, M., Sitaram, S.: Word embeddings for code-mixed language processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3067–3072. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)
34. Ramirez, A.: Bilingualism through Schooling: Cross-Cultural Education for Minority and Majority Students. State University of New York Press (1985)

35. Sennrich, R., Haddow, B.: Linguistic Input Features Improve Neural Machine Translation. In: Proceedings of the First Conference on Machine Translation. pp. 83–91. Association for Computational Linguistics, Berlin, Germany (August 2016)
36. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 86–96. Association for Computational Linguistics, Berlin, Germany (Aug 2016)
37. Sinha, R.M.K., Thakur, A.: Machine translation of bi-lingual Hindi-English (Hinglish) text. In: 10th Machine Translation summit (MT Summit X). pp. 149–156. Phuket, Thailand (2005)
38. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas. vol. 200 (2006)
39. Solorio, T., Liu, Y.: Part-of-Speech tagging for English-Spanish code-switched text. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. pp. 1051–1060. Association for Computational Linguistics, Honolulu, Hawaii (Oct 2008)
40. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. pp. 3104–3112 (2014)
41. Torregrosa, D., Pasricha, N., Masoud, M., Chakravarthi, B.R., Alonso, J., Casas, N., Arcan, M.: Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In: Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks. pp. 125–133. European Association for Machine Translation, Dublin, Ireland (19–23 Aug 2019)
42. Vilares, D., Alonso, M.A., Gómez-Rodríguez, C.: Sentiment analysis on monolingual, multilingual and code-switching twitter corpora. In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 2–8. Association for Computational Linguistics, Lisboa, Portugal (Sep 2015)
43. Wang, Z., Lee, S., Li, S., Zhou, G.: Emotion detection in code-switching texts via bilingual and sentimental information. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 763–768. Association for Computational Linguistics, Beijing, China (Jul 2015)
44. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR (2016)
45. Zhang, J., Matsumoto, T.: Corpus augmentation for neural machine translation with chinese-japanese parallel corpora. Applied Sciences **9**(10), 2036 (2019)
46. Zirikly, A., Diab, M.: Named entity recognition for Arabic social media. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. pp. 176–185. Association for Computational Linguistics, Denver, Colorado (Jun 2015)