How effective is machine translation on low-resource code-switching? A case study comparing human and automatic metrics

Li Nguyen[1,2], **Christopher Bryant**[1], **Oliver Mayeux**[3], **Zheng Yuan**[1,4]
[1]ALTA Institute, University of Cambridge, UK
[2]Linguistics and Language Technology Lab, FPT University, Vietnam
[3]Trinity College, University of Cambridge, UK
[4]Department of Informatics, King's College London, UK
{nhbn2, cjb255, ofm23}@cam.ac.uk, zheng.yuan@kcl.ac.uk

Abstract

This paper presents an investigation into the differences between processing monolingual input and code-switching (CSW) input in the context of machine translation (MT). Specifically, we compare the performance of three MT systems (Google, mBART-50 and M2M-100_{big}) in terms of their ability to translate monolingual Vietnamese, a low-resource language, and Vietnamese-English CSW respectively. To our knowledge, this is the first study to systematically analyse what might happen when multilingual MT systems are exposed to CSW data using both automatic and human metrics. We find that state-of-the-art neural translation systems not only achieve higher scores on automatic metrics when processing CSW input (compared to monolingual input), but also produce translations that are consistently rated as more semantically faithful by humans. We further suggest that automatic evaluation alone is insufficient for evaluating the translation of CSW input. Our findings establish a new benchmark that offers insights into the relationship between MT and CSW.

1 Introduction

Code-switching (CSW) is the linguistic phenomenon where two or more languages are mixed within a discourse or utterance. This is illustrated in the following example which mixes English and Vietnamese.

(1) and *mỗi* group *phải có* a different focus
each must have
'and each group must have a different focus'
(from CanVEC, Nguyen and Bryant, 2020)

Code-switching occurs frequently and naturally among bilingual speakers and has recently become increasingly visible in social media data (Doğruöz et al., 2021; Winata et al., 2022). Despite its prevalence however, Natural Language Processing (NLP) applications are typically designed to process monolingual data and so often struggle with CSW input (Solorio et al., 2021; Sitaram et al., 2020; Doğruöz et al., 2021; Nguyen et al., 2021, 2022). For machine translation (MT), no current system is designed to support code-switched text (Çetinoğlu et al., 2016; Menacer et al., 2019); and despite increasing research attention in recent years (see e.g. Chen et al., 2022 for an overview), work in this area remains sparse.

In this paper, we explore the limits of three off-the-shelf state-of-the-art machine translation systems in terms of their ability to translate Vietnamese/English CSW data, using both automatic and human evaluation metrics. As far as we are aware, this is the first study to investigate the efficacy of machine translation on CSW data involving a low-resource language which is also structurally vastly different from English. In fact, existing work has mainly focused on comparatively better resourced and/or typologically similar languages, such as Spanish/English (Xu and Yvon, 2021), French/English (Xu and Yvon, 2021; Weller et al., 2022) or Hindi/English (Appicharla et al., 2021). Vietnamese/English, or Vietnamese in particular, remains severely under-represented in NLP.

We conduct our analysis using a variety of both automatic and human metrics in order to i) better understand the strengths and weaknesses of different systems, and ii) gain some insight into the relationship between automatic and human metrics with respect to CSW input. We find that systems not only achieve higher scores on CSW input (compared to monolingual input) according to automatic metrics, but also produce translations that are considered more semantically faithful by humans. Automatic metrics furthermore fail to correlate with human judgements, which suggests that automatic evaluation alone is not enough for evaluating MT output of CSW input. We release our annotations to facilitate future research.

Findings of the Association for Computational Linguistics: ACL 2023, pages 14186–14195
July 9-14, 2023

2 Experimental Setup

2.1 Data

We conduct our experiments using the Canberra Vietnamese English natural speech corpus[1] (CanVEC), which consists of 23 self-recorded conversations among 45 Vietnamese immigrants living in Canberra, Australia (Nguyen and Bryant, 2020). One advantage of CanVEC is that it contains transcribed CSW produced by bilingual speakers in an informal speech setting – an environment that has been found to be most conducive to natural CSW behaviour (Poplack, 1980, 1993; Labov, 2004; Torres Cacoullos and Travis, 2018; Nguyen, 2018, 2020). This differs to other NLP work in this domain, which has explored either scripted CSW speech corpora (Chan et al., 2005; Shen et al., 2011; Modipa et al., 2013; Yilmaz et al., 2017) or social media text (Doğruöz and Skantze, 2021; Winata et al., 2022).

The full CanVEC corpus consists of 14,047 clauses,[2] of which 3,313 contain CSW.[3] From these 3,313, we then selected a random sample of 100 clauses, which i) contained at least 5 tokens, and ii) represented the maximum number of unique speakers. The first condition was set to ensure clauses were of a minimum length to aid contextual translation, while the second condition was set to ensure the data was diverse and did not overly represent individual speakers. Various statistics about CanVEC and our test set are shown in Table 1.

Having selected 100 CSW clauses, we gave them to two bilingual annotators with complementary language competencies; i.e. L1 English/L2 Vietnamese and L1 Vietnamese/L2 English. Each annotator then translated the CSW clauses into monolingual English and monolingual Vietnamese respectively. The monolingual English translations were used as references, while the monolingual Vietnamese translations were used as source text, which allowed us to compare CSW translations against a more typical monolingual baseline.

[1] https://github.com/Bak3rLi/CanVEC

[2] The corpus was originally segmented into finite clauses to test a specific theoretical model of code-switching, the Matrix Language Framework (Myers-Scotton, 1997), but many of these clauses are equivalent to short sentences (Nguyen, 2020).

[3] The original CanVEC paper reports 2,721 mixed clauses because it redistributed non-clause utterances (e.g. noun phrases and interjections) to the language neutral tag.

	Clauses	Avg. Len.	Tokens	Avg. EN Toks (%)
CanVEC (Mixed)	3,313	7.5	24,807	30.90%
CanVEC (Sample)	100	8.6	862	29.13%

Table 1: Descriptive statistics of the Mixed part of CanVEC in relation to our random sample. Avg. EN Toks is the average proportion of English tokens in a mixed clause; i.e. most CSW clauses are majority Vietnamese with English mixed in.

2.2 MT systems

We employ three widely used multilingual NMT models, which support both English and Vietnamese, and represent the cutting edge in both commercial and academic research.

Google Translate[4] is one of the world's most popular translation services that supports 133 languages. We access it using the *translatepy*[5] v2.3 Python API.[6]

mBART-50 is an extension of pre-trained multilingual BART (Liu et al., 2020) that has been fine-tuned on 50 languages (Tang et al., 2021). We use the *mbart-50-many-to-many* model.[7]

M2M-100 is another multilingual model that has been trained to translate between any pair of 100 different languages (Fan et al., 2021). It has been noted to perform better on non-English translations than other models and produce fluent translations with high semantic accuracy. We use the large 1.2B parameter model.[8]

3 Evaluation

3.1 Automatic evaluation

Robust evaluation is still an unsolved problem in machine translation and many metrics have been proposed (Chatzikoumi, 2020). In our experiments, we compare five different automatic metrics, which evaluate translation output quality in different ways.

BLEU is the most widely used metric for automatic MT evaluation. It estimates similarity between system output and human reference translations in terms of precision of word n-gram overlap,

[4] https://translate.google.com/

[5] https://pypi.org/project/translatepy/

[6] This free API may not be as good as the paid API.

[7] https://github.com/facebookresearch/fairseq/tree/main/examples/multilingual

[8] https://github.com/facebookresearch/fairseq/tree/main/examples/m2m_100

weighted by a brevity penalty to punish overly short translations (Papineni et al., 2002).

chrF computes an F-score using character n-grams (Popović, 2015). This helps reduce penalties when matching morphological variants of words. In our experiments, we used the default chrF_2 which weights recall twice as much as precision.

TER evaluates a system in terms of the number of edit operations (i.e. insertions, deletions, shifts and substitutions) required to change a hypothesis sentence into a reference sentence (Snover et al., 2006).

METEOR is a token-based metric that additionally rewards semantic similarity in terms of exact string match, stem match and synonym match (Denkowski and Lavie, 2014).

COMET is a trained metric that is designed to output a score that correlates with the human perception of translation quality (Rei et al., 2020). It uses a cross-lingual encoder, XLM-R (Conneau et al., 2020), and pooling operations to obtain sentence-level representations of the source, hypothesis, and reference. These sentence embeddings are combined and then passed through a feed-forward network to produce a score.

We use the implementation in *sacrebleu*[9] for the first three metrics (case agnostic, ignoring punctuation), and the pre-trained *wmt20-comet-da* model for COMET.[10] METEOR is available separately.[11]

3.2 Human evaluation

In addition to automatic metrics, we also manually rated system output according to three human metrics: *Fluency*, *Grammaticality*, and *Semantic Faithfulness* (Koehn, 2009; Dorr et al., 2011). These metrics are defined as follows.

- *Fluency*: does the translation sound natural/idiomatic in the target language?
- *Grammaticality*: is the translation grammatical, independent of the source?[12]
- *Semantic Faithfulness*: does the translation retain the intended meaning of the source?

We trained two bilingual, domain-expert annotators to assign judgements for each metric on a binary scale (0: bad, 1: good). We used a binary scale because the input clauses in our experiments are short and there were unlikely to be a lot of translation errors that would warrant a more granular scale (Koehn, 2009, p.218). It is nevertheless worth mentioning that robust human evaluation of machine translation output is still an active area of research and alternate methodologies exist (van der Lee et al., 2019; Freitag et al., 2021; Licht et al., 2022; Saldías Fuentes et al., 2022).

4 Experiments

We evaluated our three chosen MT systems in two settings: code-switching to English (*csw-en*)[13] and monolingual Vietnamese to English (*vi-en*). Recall that the sentences in the *vi-en* setting are the same as the *csw-en* setting except all English words and phrases were manually translated to Vietnamese by a human translator (Section 2.1). This enabled us to directly compare the effect of CSW against a highly controlled baseline.

Altogether, we obtained 200 translations from each system (100 clauses x 2 settings) and 600 translations in total (3 systems). We then asked our bilingual annotators to manually assign binary judgements to each translation based on the three human metrics (1800 judgements). Specifically, after training,[14] we asked the L1 English annotator to assign judgements for *Fluency* and *Grammaticality*, and the L1 Vietnamese annotator to assign judgements for *Semantic Faithfulness*. We believe judgements for *Fluency* and *Grammaticality* require native assessment of the English translation irrespective of the source, while *Semantic Faithfulness* also requires native assessment of the Vietnamese source. In all cases, a positive judgement was only awarded if the translation fully met the criteria of the given metric; this conservative approach ensured greater confidence that positive judgements truly reflected a more competent translation.

5 Results and discussion

Results from all experiments are shown in Table 2. In terms of automatic metrics, we can see that

[9] https://github.com/mjpost/sacrebleu

[10] https://github.com/Unbabel/COMET

[11] https://www.cs.cmu.edu/~alavie/METEOR/

[12] For an example of how we distinguish Fluency and Grammaticality in particular, see Appendix A.

[13] We also explored code-switching to Vietnamese (*csw-vi*) but found the output was very noisy so ultimately discounted these results. This was because the *csw-vi* setting is operationalised as an *en-vi* model even though the majority of input tokens are not English.

[14] Both annotators doubly annotated 10% of the sample on all metrics. The average inter-annotator agreement rate across all metrics and settings was 91.7%.

		Automatic Metrics					Human Metrics (%)		
Setting	**System**	**BLEU** ↑	**chrF$_2$** ↑	**TER** ↓	**METEOR** ↑	**COMET** ↑	**Fluency**	**Gram.**	**Sem.**
Baseline	Source (CSW)	12.345	30.759	87.354	0.137	-0.303	-	-	-
csw-en	Google	**27.159**	**52.861**	**51.171**	**0.315**	0.098	38	62	73
	mBART-50	25.935	49.916	56.792	0.292	0.182	**92**	**98**	**75**
	M2M-100$_{\text{big}}$	24.362	48.101	54.333	0.287	**0.271**	82	97	69
vi-en	Google	**15.639**	**41.346**	**64.169**	**0.257**	**-0.011**	57	78	42
	mBART-50	10.658	36.433	73.185	0.214	-0.158	**92**	**100**	**57**
	M2M-100$_{\text{big}}$	12.216	35.862	68.150	0.218	-0.115	75	95	55

Table 2: Performance of all systems translating code-switching to English (*csw-en*) and monolingual Vietnamese to English (*vi-en*) in terms of automatic metrics and human metrics (*Fluency*, *Grammaticality*, *Semantic Faithfulness*) compared to a do-nothing code-switching baseline. The best scores are highlighted in bold.

CSW-Input	hồi mà con *made eye contact* với Jimmy
Google	When I was *made Eye Contact* with Jimmy
mBART-50	That's when I *made eye contact* with Jimmy.
M2M-100$_{\text{big}}$	I *made eye contact* with Jimmy.
VI-Input	hồi mà con bắt gặp ánh mắt của Jimmy
Google	When I caught Jimmy's gaze
mBART-50	That's when I met Jimmy's eyes.
M2M-100$_{\text{big}}$	When I saw Jimmy's eyes.
Reference	When I *made eye contact* with Jimmy

Table 3: System output for an example clause showing how CSW input may be more favourably constrained towards a reference compared to monolingual input.

Google outperforms **mBART-50** and **M2M-100$_{\text{big}}$** on all metrics except for COMET in *csw-en*; this suggests that **Google** is the best of the three MT systems on our CSW/monolingual test sets. It is furthermore noteworthy that performance on *csw-en* translation for all three MT systems is consistently and significantly higher than monolingual *vi-en* translation. In fact, a do-nothing CSW baseline seems to outperform **mBART-50** and **M2M-100$_{\text{big}}$** at monolingual *vi-en* translation in terms of BLEU.

We hypothesise that this is because the translation might be considered 'easier' when CSW fragments only need to be copied to the output. For example, Table 3 shows that all systems generate output containing the phrase "made eye contact" when that same phrase is present in the CSW input, but generate synonymous output "caught Jimmy's gaze", "met Jimmy's eyes" and "saw Jimmy's eyes" from the monolingual Vietnamese input. BLEU thus benefits more from this exact word match compared to other automatic metrics. Consequently, CSW translation is more constrained than monolingual translation, which might make it 'easier' to achieve higher scores.

In contrast, system performance on human metrics is more varied, and different systems performed better and worse on different metrics. For example, **mBART-50** achieved near-perfect scores for both *Fluency* and *Grammaticality* regardless of whether the input was CSW or monolingual, while **Google** achieved higher scores in the monolingual setting and **M2M-100$_{\text{big}}$** achieved higher scores in the CSW setting on the same metrics. Holistically, this suggests that **mBART-50** may be the most stable and effective of the three systems in terms of processing CSW input in relation to these metrics. **Google**, in contrast, appears to be the weakest system, which contradicts our findings from the automatic evaluation. This lack of agreement is not entirely surprising however, given that it is already challenging to develop automatic metrics that correlate with human judgements in monolingual settings (Fomicheva and Specia, 2019), let alone CSW settings where languages are mixed.

Among the three human metrics, we also observe that the scores for *Semantic Faithfulness* were consistently higher given CSW input compared to monolingual input. While this is again likely due to the constraining nature of CSW input, this result potentially suggests a specific aspect of MT where CSW input can contribute to enhancing system output. We direct readers to Appendix B for some detailed examples. Ultimately, we consider this finding worthy of further investigation, especially in relation to the development of models involving the understanding and/or generation of code-switching texts.

6 Conclusion

In this work, we compared the performance of three state-of-the-art MT systems on CSW input, using both automatic and human metrics. We found that systems not only achieved higher scores on automatic metrics when processing CSW input (com-

pared to monolingual input), but also produced translations that were consistently rated as more semantically faithful by humans. We furthermore observed that automatic and human metrics do not agree, which again highlights the need for more sophisticated, robust metrics, especially in non-monolingual tasks. Our findings establish a new benchmark in the relationship between MT and CSW, and motivate further research into how CSW might be used to improve future systems.

Limitations

The main limitation of our work is that 100 clauses is a small test test, but this was necessary to keep our human evaluation experiments manageable. We furthermore believe this was sufficient to be able to draw meaningful conclusions about the capabilities of different systems.

Another limitation is that we were only able to evaluate low-resource CSW in the context of Vietnamese and English. Future work might explore whether the same observations hold with CSW involving other low-resource languages, but this would require access to more suitable corpora and annotators.

Ethics Statement

We made every effort to make sure the work described in this paper adheres to the ACL Code of Ethics.

Acknowledgements

We would like to thank Professor Paula Buttery (Director, ALTA Institute, University of Cambridge) for her continuing support.

References

Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2021. IITP-MT at WAT2021: Indic-English multilingual neural machine translation using Romanized vocabulary. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 238–243, Online. Association for Computational Linguistics.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11, Austin, Texas. Association for Computational Linguistics.

Joyce Y. C. Chan, P. C. Ching, and Tan Lee. 2005. Development of a Cantonese-English Code-mixing Speech Corpus. In *Interspeech*.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Shuguang Chen, Gustavo Aguilar, Anirudh Srinivasan, Mona Diab, and Thamar Solorio. 2022. Calcs 2021 shared task: Machine translation for code-switched data.

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.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

A. Seza Doğruöz and Gabriel Skantze. 2021. How "open" are the conversations with open-domain chatbots? a proposal for speech event based evaluation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 392–402, Singapore and Online. Association for Computational Linguistics.

Bonnie J. Dorr, Matt Snover, and Nitin Madnani. 2011. Machine translation evaluation and optimization. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global autonomous language exploitation*. Springer, New York.

Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and

Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Marina Fomicheva and Lucia Specia. 2019. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press, Cambridge.

William Labov. 2004. Field methods of the project on linguistic change and variation. In J. Baugh and J. Sherzer, editors, *Language in use: Readings in sociolinguistics*, page 29. Prentice Hall, Englewood Cliffs, NJ.

Daniel Licht, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, and Philipp Koehn. 2022. Consistent human evaluation of machine translation across language pairs. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 309–321, Orlando, USA. Association for Machine Translation in the Americas.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

M. A. Menacer, D. Langlois, D. Jouvet, D. Fohr, O. Mella, and K. Smaïli. 2019. Machine translation on a parallel code-switched corpus. In *Advances in Artificial Intelligence*, pages 426–432, Cham. Springer International Publishing.

Thipe I Modipa, Marelie H Davel, and Febe de Wet. 2013. Implications of Sepedi/English code switching for ASR systems. In *Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching (revised with a new afterword)*. Clarendon Press, Oxford.

Li Nguyen. 2018. Borrowing or code-switching? Traces of community norms in Vietnamese-English speech. *The Australian Journal of Linguistics*, 38(4):443–466.

Li Nguyen. 2020. *Cross-generational linguistic variation in the Canberra Vietnamese heritage language community: A corpus-centred investigation*. PhD dissertation, University of Cambridge.

Li Nguyen and Christopher Bryant. 2020. CanVEC - the canberra Vietnamese-English code-switching natural speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4121–4129, Marseille, France. European Language Resources Association.

Li Nguyen, Christopher Bryant, Sana Kidwai, and Theresa Biberauer. 2021. Automatic language identification in code-switched hindi-english social media text. *Journal of Open Humanities Data*.

Li Nguyen, Zheng Yuan, and Graham Seed. 2022. Building educational technologies for code-switching: Current practices, difficulties and future directions. *Languages*, 7(3).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shana Poplack. 1980. "Sometimes I'll start a sentence in Spanish y termino en español": toward a typology of code-switching. *Linguistics*, 18.

Shana Poplack. 1993. Variation theory and language contact. In D. Preston, editor, *American dialect research: An anthology celebrating the 100th anniversary of the American Dialect Society*, page 252. Benjamins, Amsterdam.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Belén Saldías Fuentes, George Foster, Markus Freitag, and Qijun Tan. 2022. Toward more effective human evaluation for machine translation. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 76–89, Dublin, Ireland. Association for Computational Linguistics.

Han Ping Shen, Chung Hsien Wu, Yan Ting Yang, and Chun Shan Hsu. 2011. CECOS: A Chinese-English code-switching speech database. In *2011 International Conference on Speech Database and Assessments, Oriental COCOSDA 2011 - Proceedings*.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2020. A survey of code-switched speech and language processing.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Thamar Solorio, Shuguang Chen, Alan W. Black, Mona Diab, Sunayana Sitaram, Victor Soto, Emre Yilmaz, and Anirudh Srinivasan, editors. 2021. *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Online.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Rena Torres Cacoullos and Catherine E. Travis. 2018. *Bilingualism in the community: Code-switching and grammars in contact*. Cambridge University Press, Cambridge.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. 2022. End-to-end speech translation for code switched speech. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1435–1448, Dublin, Ireland. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2022. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges.

Jitao Xu and François Yvon. 2021. Can you traducir this? machine translation for code-switched input. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 84–94, Online. Association for Computational Linguistics.

Emre Yilmaz, Jelske Dijkstra, Hans Van De Velde, Frederik Kampstra, Jouke Algra, Henk Van Den Heuvel, and David Van Leeuwen. 2017. Longitudinal speaker clustering and verification corpus with code-switching Frisian-Dutch speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*.

GLOSSARY

1	First person
2	Second person
CLF	Classifier
DET	Determiner
PL	Plural
POSS	Possessive
SG	Singular
Q	Question marker

A Distinguishing Fluency and Grammaticality

We specified in Section 3.2 the three metrics that we used for human judgement in this work, namely *Fluency*, *Grammaticality* and *Semantic Faithfulness*. We consider the distinction between *Grammaticality* and *Fluency* an especially important aspect of languages in contact as it is likely to involve non-standard or hybrid features that may not be easily translated into the target language. Despite some overlap, there are cases in the dataset where these two criteria are clearly separated. Example (2) illustrates.

(2) *chính* *vì* *dùng* **mirror** **[M2M]**
main because use

nó *mới* *có* *điểm* *chết* *đấy* *mày*
3SG then have point death PRT 2SG

Translation: 'Because in the mirror, you have a point of death'

Intended meaning: 'The blind spot is precisely because of using the mirror, you$_{[\text{-VOCATIVE}]}$'

Here, the machine translation is grammatically correct, but not fluent to a native's ear. An expected fluent output in this case would be 'You have a blind spot precisely because of the mirror.' The use of the non-idiomatic expression 'point of death' and the topicalisation of the prepositional phrase 'in the mirror', therefore, while not wrong, could not be marked as fluent.

B Analysis of Semantic Faithfulness

We reported near the end of the Discussion (section 5) that the scores for *Semantic Faithfulness* are always higher in the code-switching data (*csw-en*) compared to monolingual data (*vi-en*). This difference is confirmed as statistically significant ($p < 0.05$) using a bootstrap resampling test (Efron and Tibshirani, 1993). Here, we provide some qualitative examples.

(3) **Google**

a. *con có muốn trở-thành* **[vi-en]**
2SG have want become

bạn-gái *của mấy người đó không*
girlfriend POSS PL person DET Q

Translation: 'Do you want to be your girlfriend?'

b. *con có muốn trở-thành* **[csw-en]**
2SG have want become

girlfriend *của mấy người đó không*
POSS PL person DET Q

Translation: 'Do you want to become the girlfriend of those people?'

Intended meaning: 'Do you want to become their girlfriend?'

(4) **mBart-50**

a. ***có-thể*** *con mượn thêm* **[vi-en]**
maybe 1SG borrow extra

cái máy của Sarah *nữa*
CLF machine POSS more

Translation: 'I can spin Sarah's machine too.'

b. **maybe** *con mượn thêm* **[csw-en]**
1SG borrow extra

cái máy của Sarah *nữa*
CLF machine POSS more

Translation: 'Maybe I can borrow Sarah's machine.'

Intended meaning: 'Maybe I can borrow Sarah's machine too.'

(5) **M2M-100$_{\text{big}}$**

a. ***đoạn băng*** *nó quay thế-nào* **[vi-en]**
clip video DET record how

Translation: 'How did that tape go?'

b. **clip** *nó quay thế-nào* **[csw-en]**
DET record how

Translation: 'How is the clip recorded?'

Intended meaning: 'How was the [video] clip recorded?'

As we can see, even when the code-switching part of the source only comprises a single word (highlighted in purple), the translation output is noticeably enriched. In (3) for example, while the **Google** system could not capture either the correct possessor ('of those people') or the precise meaning of the infinitive ('become'), it was able to do so on both occasions in a *csw-en* setting. This is particularly striking considering that the source sentence is long (which should give sufficient context) *and* that the only difference between (3a) and (3b) is the language of one lexical item ('*bạn-gái*' vs 'girlfriend'). Similarly, examples (4) and (5) show comparable behaviour for **mBart-50** and **M2M-100$_{\text{big}}$**, where a single code-switch noticeably adds to the output's semantics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7: Mandatory limitations section

☐ A2. Did you discuss any potential risks of your work?
Not applicable. No known risks

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2.1: New human translations Section 4: New human judgments

☑ B1. Did you cite the creators of artifacts you used?
Section 2.1: Data Section 2.2: Machine translation models Section 3.1: Machine translation metrics

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Is it necessary to include a license for, e.g., the BLEU score in the paper?

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Anonymization of the original corpus was described in the original corpus paper. We added new annotations to an already anonymized, ethically-created corpus.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2.1: Corpus description

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Table 1

C ☒ Did you run computational experiments?

We evaluated existing models on new data and carried out an automatic and human evaluation.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 2.1: Human translators Section 4: Human judgments of machine translation output

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3.2: Definitions of metrics on a binary scale.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Annotators are co-authors

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Annotators are co-authors

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 2.1: L1/L2 information about annotators