# TRUE BILINGUAL NMT

**Mohamed Anwar, Lekan Raheem & Maab Elrashid**
African Institute for Mathematical Sciences (AIMS)
{mghanem,rwaliyu,mnimir}@aimsammi.org

**Melvin Johnson & Julia Kreutzer**
Google Research
{melvinp,jkreutzer}@google.com

## ABSTRACT

Bilingual machine translation permits training a single model that translates monolingual sentences from one language to another. However, a model is not truly bilingual unless it can translate back and forth in both language directions it was trained on, along with translating code-switched sentences to either language. We propose a true bilingual model trained on WMT14 English-French (En-Fr) dataset. For better use of parallel data, we generated synthetic code-switched (CSW) data along with an alignment loss on the encoder to align representations across languages. Our model strongly outperforms bilingual baselines on CSW translation while maintaining quality for non-code switched data.

## 1 INTRODUCTION

Neural Machine Translation (NMT) systems can be divided into two categories: bilingual models and multilingual models. Bilingual models are able to translate from one language to another; while multilingual models are able to translate between multiple languages (Firat et al. (2016); Johnson et al. (2017)). We argue that bilingual models in that sense aren't actually bilingual since they can't translate in the opposite direction, and can't translate code-switched sentences either. Code-Switching (CSW) denotes the alternation of two languages within a single utterance (Poplack (1980); Sitaram et al. (2020)).

According to Vilhanova (2018), Africa is the most multilingual continent in the world, and this requires rethinking what we expect from our NMT models when they're deployed. In this paper, we aim to build a true bilingual NMT model. It is a single model that is able to translate in both directions and also translate code-switched sentences. This model was trained using only parallel data accompanied with synthetic code-switched data. To make the encoder create language-agnostic representations, we propose an alignment loss function applied only on the encoder.

## 2 DATA

Since English and French are among the most high-resource spoken languages in Africa, we used WMT14 English-French benchmark for training, newstest2008-2013 for validation, and newstest2014 for test. Parallel corpora for code-switched data is very scarce (Menacer et al. (2019)), however, there have been works on generating synthetic code-switched data. Similar to Song et al. (2019) and Xu & Yvon (2021), we created code-switched data, by first extracting word alignment using *fast-align* toolkit (Dyer et al. (2013)), and then extracting minimal alignment units following the approach of (Crego (2005)). We chose the matrix language — defined by the Matrix Language Frame (MLF) theory (Poulisse (1998)) — randomly (50%). Similar to MLM pre-training used by BERT (Devlin et al. (2019)), we randomly replaced 15% of the sentence length with its aligned segments in the embedded language. We combined the code-switched data generated with the parallel bidirectional data after prepending a target language token (Johnson et al. (2017)). For more details, check Appendix A.

## 3 METHODOLOGY

Following Arivazhagan et al. (2019) steps, we use parallel data, and enforce the encoder to make language-agnostic representations about the input sequences by minimizing the max-pooled cosine distance of the encoder representations of the parallel data as shown in the following equation:

$$\Omega = \mathop{\mathbb{E}}_{x_{src}, x_{tgt} \sim D_{(en,fr)}} [1 - sim(Enc(x_{src}), Enc(x_{tgt}))] \tag{1}$$

Where $\Omega$ is the encoder loss, $D_{(en,fr)}$ is our data containing parallel language pairs combined with code-switched ones, $x_{src}$ is the source sentence which could be monolingual or code-switched while $x_{tgt}$ is the target which is always monolingual, $Enc(x)$ is the max-pooled encoder representation of sentence $x$ similar to Gouws et al. (2016) and Coulmance et al. (2016), and $sim$ is the cosine-similarity. Unlike Arivazhagan et al. (2019) where the whole model's parameters were updated, we updated the encoder parameters only, as shown in Figure 1.

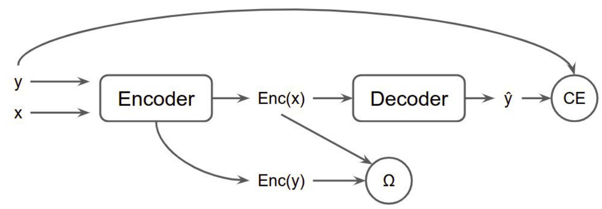


Figure 1: shows the loss functions used where $CE$ is the cross entropy, $\Omega$ is the encoder loss.

## 4 Experiments and Results

In all of our experiments, we used Transformer-Base (Vaswani et al. (2017)) configuration with the Fairseq (Ott et al. (2019)) framework. All models were trained on four Tesla T400 GPUs using WMT-14 data for training, with a shared vocabulary of 40K BPE Sennrich et al. (2016) sub-words. The model's hyperparameters can be found in Appendix B.

We created two baselines using the same training details as mentioned above: 1) **Bidirectional**: $En \leftrightarrow Fr$ model trained only on parallel data in both directions. 2) **csw**: $En \leftrightarrow Fr$ model trained only on synthetic code-switched data. Our model **bi+csw+cosine** was trained on bidirectional and code-switched data along with the encoder criterion mentioned in eq. 1. To check the effect of the encoder alignment loss, we trained another model without the encoder criterion **bi+csw**. Table 1 shows the results of all models trained using the same parameters seen in Table 5.

| Model | Steps | Unidirectional | | CSW | |
|---|---|---|---|---|---|
| | | to Fr | to En | to Fr | to En |
| Bidirectional | 642K | **39.57** | **36.17** | 57.86 | 60.77 |
| csw | 594K | 8.38 | 13.66 | 68.49 | 66.65 |
| bi+csw | 420k | 38.57 | 34.75 | 68.38 | 66.31 |
| bi+csw+cosine | 612k | 39.19 | 35.43 | **68.69** | **66.96** |

Table 1: Case-sensitive detokenized 4-gram BLEU Score on unidirectional and CSW data from newstest2014 using SacreBLEU (Post (2018)) with beam-size of 5.

From Table 1 we see that the **Bidirectional** baseline performs well for bi-directional translation. However, **csw** baseline performs well only on CSW translation. Our combined models **bi+csw** and **bi+csw+cosine** work well across the board where **bi+csw+cosine** has the best performance on CSW data while achieving competitive results on bi-directional translation compared to the bidirectional baseline.

## 5 Conclusion

In this paper, we introduced two ways to make best-use of parallel data that can improve model's performance on both unidirectional and code-switched data: 1) A statistical way to generate code-switched data that can be aggregated with unidirectional data for training. 2) A loss function that trains the encoder to generate language-independent representations. We show that these two techniques boosted our model's performance on both unidirectional and code-switched data. This is still a work-in-progress, and we are exploring new ways to improve our model even more; and more importantly experimenting with Arabic, one of the most spoken languages in Africa.

## REFERENCES

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. The missing ingredient in zero-shot neural machine translation, 2019.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. Trans-gram, fast cross-lingual word-embeddings, 2016.

Josep Crego. Reordered search and tuple unfolding for ngram-based smt. 01 2005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `https://aclanthology.org/N13-1073`.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism, 2016.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments, 2016.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation, 2017.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pp. 177–180. Association for Computational Linguistics, 2007. URL `http://dl.acm.org/citation.cfm?id=1557769.1557821`.

Mohamed Menacer, David Langlois, Denis Jouvet, Dominique Fohr, Odile Mella, and Kamel Smaïli. Machine Translation on a parallel Code-Switched Corpus. In *Canadian AI 2019 - 32nd Conference on Canadian Artificial Intelligence*, Lecture Notes in Artificial Intelligence, Ontario, Canada, May 2019. URL `https://hal.archives-ouvertes.fr/hal-02106010`.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling, 2019.

Shana Poplack. Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching 1. *Linguistics*, 18:581–618, 01 1980. doi: 10.1515/ling.1980.18.7-8.581.

Matt Post. A call for clarity in reporting bleu scores, 2018.

Nanda Poulisse. Duelling languages: Grammatical structure in codeswitching. *International Journal of Bilingualism*, 2(3):377–380, 1998. doi: 10.1177/136700699800200308. URL `https://doi.org/10.1177/136700699800200308`.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016.

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

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. Code-switching for enhancing NMT with pre-specified translation. *CoRR*, abs/1904.09107, 2019. URL `http://arxiv.org/abs/1904.09107`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Viera Vilhanova. *Multilingualism in Africa. Challenges and Solutions*. 02 2018.

Jitao Xu and François Yvon. Can you traducir this? machine translation for code-switched input. *CoRR*, abs/2105.04846, 2021. URL `https://arxiv.org/abs/2105.04846`.

# A APPENDIX

## A.1 DATASET

We used WMT-14 benchmark for training, with a shared vocabulary of 40K BPE Sennrich et al. (2016) sub-words. We used 4096 tokens per batch, and we removed any sentences whose length are less than 2 tokens and more than 250 tokens. Also we prepended a target-language tag to the sentences as shown in Table 2. All data was tokenized and normalized using Moses SMT (Koehn et al. (2007)). The stats of the data after cleaning can be found in Table 3.

| Source | Target |
|---|---|
| <2fr> The weather today is nice . | **Il fait beau aujourd'hui .** |
| <2en> **Il fait beau aujourd'hui .** | The weather today is nice . |
| <2en> **Il fait** nice today . | The weather today is nice . |
| <2fr> **Il fait** nice today . | **Il fait beau aujourd'hui .** |

Table 2: Example of the formulation of our data, where the bold words are in French and the normal ones are in English. The first two examples are monolingual source sentences; the first is from English to French, and the second is from French to English. The last two examples are code-switched on the source and monolingual on the target.

| Dataset | Train Size | Valid Size | Test Size | Total |
|---|---|---|---|---|
| Fr-En | 35789717 | 15827 | 3003 | 35808547 |
| Bidirectional | 71579434 | 31654 | 6006 | 71617094 |
| CSW | 77198306 | 32581 | 6006 | 77236893 |
| Bi+CSW | 148777740 | 64235 | 12012 | 148853987 |

Table 3: WMT-14 French-English data stats used for training, validation and test.

## A.2 CODE-SWITCHED DATA GENERATION METHOD

The following are the steps that we followed to generate the code-switched (CSW) data; a sample can be found in Table 4. These steps were adapted from (Xu & Yvon (2021) which can be summarized into the following:

1. **Data preprocessing**: data tokenization and normalization was done using this perl script: *clean-corpus-n.perl* from moses SMT (Koehn et al. (2007)).
2. **Alignment**: Fast-Align (Dyer et al. (2013)) tool with *gdfa* (grow-diag-final-and).
3. **Random replacement**: Aligned segments were replaced by considering the following:
   (a) Matrix Language is chosen randomly (50-50)%.
   (b) Replace around 15% of the input sequence.
   (c) Short sequences (less than 7 tokens) have just one replacement.
   (d) Positions of aligned segments are chosen uniformly.

| *CSW Sentence* | *English Translation* | *French Translation* | *Matrix Language* |
|---|---|---|---|
| Difficult Year **pour les** Pharmacists . | Difficult Year for Pharmacists . | Année difficile pour les pharmaciens . | English |
| **Il ne** believe **pas que l'Ontario emboîtera le pas** . | He does not believe that Ontario will follow suit . | Il ne croit pas que l'Ontario emboîtera le pas . | French |
| Asked how he had developed his character , the **acteur** and singer Justin Timberlake **avait rappelé** how he " **grandi dans** Tennessee , bathed in the blues and country music " . | When asked how he came up with his character, actor and singer Justin Timberlake recalled that he " grew up in Tennessee, surrounded by blues and country music ". | Interrogé sur la façon dont il a composé son personnage , l'acteur et chanteur Justin Timberlake avait rappelé avoir " grandi dans le Tennessee, baigné par le blues et la country " . | English |
| **Mes camarades** cried with **joie et mes parents ont conservé** every journaux qu'ils ont trouvés . | My classmates cried with joy , and my parents saved every newspaper they could find . | Mes camarades de classe ont pleuré de joie, et mes parents ont gardé tous les journaux qu'ils ont pu trouver . | French |

Table 4: A sample of the code-switched data generated form newstest2014 dataset. Bold words are French while normal ones are English.

## B Appendix

Table 5 holds all the hyper-parameters we used for training all models. All models were trained till convergence with patience = 10.

| Hyper-parameter | Value |
|---|---|
| Number of Layers | 6 |
| Hidden size | 512 |
| FFN inner hidden size | 2048 |
| Attention heads | 8 |
| Attention head size | 64 |
| Dropout | 0.1 |
| Attention Dropout | 0.0 |
| Warmup Steps | 4000 |
| Learning Rate | 5e-4 |
| Learning Rate Decay | inverse_sqrt |
| Batch Size | 4096 tokens |
| Label Smoothing | 0.1 |
| Weight Decay | 0.0001 |
| Adam $\epsilon$ | $10^{-9}$ |
| Adam $\beta_1$ | 0.9 |
| Adam $\beta_2$ | 0.98 |
| Encoder Criterion Weight | 10 |

Table 5: The hyperparameter values setting for training.