Can You Traducir This? Machine Translation for Code-Switched Input
Jitao Xu, François Yvon

To cite this version:
Jitao Xu, François Yvon. Can You Traducir This? Machine Translation for Code-Switched Input. Workshop on Computational Approaches to Linguistic Code Switching, Association for Computational Linguistics, Jun 2021, Online, United States. hal-03218889

HAL Id: hal-03218889
https://hal.science/hal-03218889
Submitted on 10 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Can You Traducir This? Machine Translation for Code-Switched Input

Jitao Xu
Univ. Paris-Saclay, & CNRS, LISN
Orsay, France
jitao.xu@limsi.fr

François Yvon
Univ. Paris-Saclay, & CNRS, LISN
Orsay, France
francois.yvon@limsi.fr

Abstract

Code-Switching (CSW) is a common phenomenon that occurs in multilingual geographic or social contexts, which raises challenging problems for natural language processing tools. We focus here on Machine Translation (MT) of CSW texts, where we aim to simultaneously disentangle and translate the two mixed languages. Due to the lack of actual translated CSW data, we generate artificial training data from regular parallel texts. Experiments show this training strategy yields MT systems that surpass multilingual systems for code-switched texts. These results are confirmed in an alternative task aimed at providing contextual translations for a L2 writing assistant.

1 Introduction

Code-Switching (CSW) denotes the alternation of two languages within a single utterance (Poplack, 1980; Sitaram et al., 2019). It is a common communicative phenomenon that occurs in multilingual communities during spoken and written interactions. CSW is a well studied phenomenon in linguistic circles and has given rise to a number of theories regarding the structure of mixed language fragments (Poplack, 1978; Pfaff, 1979; Poplack, 1980; Belazi et al., 1994; Myers-Scotton, 1997). The Matrix Language Frame (MLF) theory (Myers-Scotton, 1997) defines the concept of matrix and embedded languages where the matrix language is the main language that the sentence structure should conform to and notably provides the syntactic morphemes, while the influence of the embedded language is lesser and is mostly manifested in the insertion of content morphemes.

The rise of social media and user-generated content has made written instances of code-switched language more visible. It is estimated that as much as 17% of Indian Facebook posts (Bali et al., 2014) and 3.5% of all tweets (Rijhwani et al., 2017) are code-switched. This phenomenon is also becoming more pervasive in short text messages, chats, blogs, and the like (Samih et al., 2016). Code-switching however remains understudied in natural language processing (NLP) (Aguilar and Solorio, 2020), and most work to date has focused on token-level language identification (LID) (Samih et al., 2016) and on language models for Automatic Speech Recognition (Winata et al., 2019). More tasks are being considered lately, such as Named Entity Recognition (Aguilar et al., 2018), Part-of-Speech tagging (Ball and Garrette, 2018) or Sentiment Analysis (Patwa et al., 2020).

We focus here on another task for CSW texts: Machine Translation (MT). The advent of Neural Machine Translation (NMT) technologies (Bahdanau et al., 2015; Vaswani et al., 2017) has made it possible to design multilingual models capable of translating from multiple source languages into multiple target languages (Firat et al., 2016; Johnson et al., 2017), where however both the input and output are monolingual. We study here the ability of such architectures to translate fragments freely mixing a “matrix” and an “embedded” language into monolingual utterances.

Our main contribution is to show that for the two pairs of languages considered (French-English and Spanish-English): (a) translation of CSW texts is almost as good as the translation of monolingual texts – a performance that bilingual systems are unable to match; (b) such results can be obtained by training solely with artificial data; (c) CSW translation systems achieve a near deterministic ability to recopy in the output target words found in the input, suggesting that they are endowed with some language identification abilities. Using these models, we are also able to obtain competitive results on the SemEval 2014 Task 5: L2 Writing Assistant, which we see as one potential application area of CSW translation.
2 Building translation systems for code-switched data

2.1 Code-switched data generation

Parallel corpora with natural CSW data are very scarce (Menacer et al., 2019) and, similar to Song et al. (2019a), we generate artificial CSW parallel sentences from regular translation data.

We first compute word alignments between parallel sentences using fast_align (Dyer et al., 2013). We then extract so-called minimal alignment units following the approach of Crego et al. (2005): these correspond to small bilingual phrase words, which the theory would generally consider ungrammatical.

unconstrained fragments containing both content and function that our artificial CSW sentences will contain insertions of arbitrary insertions from the “embedded” language. This means that the actual number of replacements does not exceed half of either the source or target sentences length, adjusting the actual number of replacements as:

\[ n = \min\left(\frac{S}{2}, \frac{T}{2}, r\right) \]  (2)

where \( S \) and \( T \) are respectively the length of the source and target sentences. We finally choose uniformly at random \( r \) alignment units and replace these fragments in the matrix language by their counterpart in the embedded language. Figure 1 displays examples of generated CSW sentences.

2.2 Machine translation for CSW data

2.2.1 Data preparation

We use WMT data for CSW data generation and for training MT systems. We discard sentences which do not possess the correct language by using the fasttext LID model (Bojanowski et al., 2017). We use Moses tools (Koehn et al., 2007) to normalize punctuations, remove non-printing characters and discard sentence pairs with a source / target ratio higher than 1.5, with a maximum sentence length of 250. We tokenize all WMT data using Moses tokenizer. Our procedure for artificial CSW data generation uses WMT13 En-Es parallel data with 14.5M sentences. For En-Fr, we use all WMT14 parallel data, for a grand total of 33.9M sentences. Our development sets are respectively newstest2011 and newstest2012 for En-Es, and newstest2012 and newstest2013 as development sets for En-Fr; the corresponding test sets are newstest2013 (En-Es) and newstest2014 (En-Fr).

2.2.2 Machine Translation systems

We use the fairseq (Ott et al., 2019) implementation of Transformer base (Vaswani et al., 2017) for our models with a hidden size of 512 and a feed-forward size of 2048. We optimize with Adam, set up with an initial learning rate of 0.0007 and an inverse square root weight decay schedule, as well as 4000 warmup steps. All models were trained with mixed precision and a batch size of 8192 tokens for 300k iterations on 4 V100 GPUs. For each language pair, we use a shared source-target inventory built with Byte Pair Encoding (BPE) of 32K merge operations, using the implementation published by Sennrich et al. (2016). Note that we do not share the embedding matrices. Our experiments with sharing the decoder’s input and output embeddings or sharing all encoder-decoder embeddings did not yield further gains.

We compare three settings for Code-Switch models:

- the base-csw setting, where we train two separate systems, one translating CSW into English, and the other translating CSW into Spanish or French.
- the multi-csw setting, where we train one model able to generate either pure matrix or embedded language in the output. To this...
Matrix | In Oregon, planners are experimenting with giving drivers different choices.
---|---
r = 1 | Dans Oregon, planners are experimenting with giving drivers different choices.
r = 2 | Dans Oregon, les planificateurs are experimenting with giving drivers different choices.
r = 3 | Dans Oregon, les planificateurs sont en offrant aux drivers different choices.
Embedded | Dans l’Oregon, les planificateurs tentent l’expérience en offrant aux automobilistes différents choix.

Figure 1: Examples of generated CSW sentences when taking English as the matrix language and varying the number \( r \) of replacements of embedded French segments (in boldface).

end, similar to a multilingual NMT model (Johnson et al., 2017), we add a tag at the beginning of each CSW sentence to specify the desired target language. Taking En-Fr as an example, we add a \(<\text{EN}>\) tag for CSW-En and a \(<\text{FR}>\) tag for CSW-Fr. We use the combination of CSW-En and CSW-Fr data for training, which implies that each source side (CSW sentence) is duplicated in the training data, once for each possible output.

- the joint-csw setting, which extends multi-csw by using one encoder and two separate decoders and training the two output languages simultaneously with a combined loss function: for each training (CSW) instance, the loss function sums the two prediction terms for the embedded and the matrix language. The training data remains the same.

Note that all our Code-Switch systems also have the ability to translate monolingual source data, in either direction.

For comparison purposes, we also use our parallel data to train two baselines: (a) regular NMT systems for the considered language pairs (base), similar to base-csw; (b) bilingual NMT systems, capable of translating from and into both two languages (bilingual). The selection of the desired target language relies on the same tagging mechanism as multi-csw, which means that both types of models see exactly the same examples. All resulting baseline Transformer models have the exact same hyperparameters and use the same training scheme as Code-Switch. Performance is computed with SacreBLEU (Post, 2018) and METEOR (Denkowski and Lavie, 2014).

3 Machine translation experiments

3.1 Results

We run tests using artificial CSW datasets, as mentioned in Section 2.2, as well as on the original test sets, in order to evaluate our models’ ability to translate both CSW and monolingual sentences. Results are in Table 1 where we also separately report scores for the ‘Matrix’ and ‘Embedded’ part of the test sets. As is obvious on the copy line, the ‘Embedded’ part contains mostly source language, and corresponds to an actual translation task whereas the ‘Matrix’ part mostly contains target words on the source side, and is much easier to translate.

On the left part of this table, we see that the baseline systems, either with two (base) or one single (bilingual) model(s), do better on monolingual test sets than their counterparts trained on CSW data (respectively base-csw and multi-csw). For both language pairs, the observed differences are in the range of 1-1.5 BLEU points. Conversely, when translating CSW sentences, *-csw models perform significantly better than the corresponding baselines models, which have never seen CSW in the source.

Moreover, we note the marked differences between BLEU scores obtained by these models when the matrix language for the CSW source is the target and when the embedded language is the target. In the former case, translation is near perfect; in the latter case they nonetheless use the little information available to improve over the monolingual scores (about 1-1.5 BLEU points), nearly matching the performance of the baseline systems. This is illustrated for Fr-En, for which joint-csw improved from 33.7 to 35.0; in the same condition, the bilingual system only improves by 0.1 point.

Among the three Code-Switch models, multi-csw is the weakest, while the other two achieve comparable performance. Interestingly, with joint training (joint-csw), we can recover with one single system the performance of the two separate systems used in the base-csw condition. On the monolingual tests, this system also matches the performance of the multilingual baseline (bilingual), which makes it overall our best contender of the lot.
### Table 1: Translating monolingual newstest data and artificial csw-newstest data for two language pairs where performance is measured via the BLEU (B) and METEOR (M) scores. We also report a trivial baseline that just recopies the source text. Small numbers contain BLEU scores computed separately when the target language is the embedded language (left) and the matrix language (right). For the monolingual tests (left part), these correspond to scores computed on the same sentences that are also included in the CSW tests.

<table>
<thead>
<tr>
<th>Testset</th>
<th>newstest2013</th>
<th>csw-newstest2013</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direction</strong></td>
<td>En-En</td>
<td>CSW-En</td>
</tr>
<tr>
<td><strong>Metrics</strong></td>
<td>B M B M</td>
<td>B M B M</td>
</tr>
<tr>
<td><strong>copy</strong></td>
<td>- - - -</td>
<td>50.3 57.8 46.8 31.7</td>
</tr>
<tr>
<td><strong>base</strong></td>
<td>33.2 58.3</td>
<td>33.8 36.4</td>
</tr>
<tr>
<td><strong>bilingual</strong></td>
<td>31.9 57.3</td>
<td>32.6 35.9</td>
</tr>
<tr>
<td><strong>base-csw</strong></td>
<td>32.0 57.4</td>
<td>32.7 36.0</td>
</tr>
<tr>
<td><strong>multi-csw</strong></td>
<td>31.1 56.7</td>
<td>31.5 35.4</td>
</tr>
<tr>
<td><strong>joint-csw</strong></td>
<td>31.9 57.3</td>
<td>32.6 36.0</td>
</tr>
</tbody>
</table>

| **Direction**    | En-Fr        | CSW-Fr           |
| **Metrics**      | B M B M      | B M B M          |
| **copy**         | - - - -      | 50.0 55.7 46.5 33.1 |
| **base**         | 37.9 60.9    | 35.4 37.9        |
| **bilingual**    | 36.3 59.6    | 34.5 37.6        |
| **base-csw**     | 36.7 59.9    | 34.3 37.5        |
| **multi-csw**    | 35.2 58.7    | 32.9 36.8        |
| **joint-csw**    | 36.2 59.5    | 34.0 37.3        |

3.2 Analysis

3.2.1 Code-Switching effect

In order to better study the effect of mixing languages, we modify the synthetic data generation method to keep one language as the matrix language, in which segments are incrementally replaced by translations of the embedded language. We relax the constraint on the maximum number of replacements and generate new test sets with an increasing number of replacements, ranging from 1 to 20, resulting in 20 versions of the CSW test sets (in each direction). In Figure 2, we plot the BLEU scores of both source CSW sentences and their translations for En-Fr language pair, using each language as the matrix language, to visualize the impact of progressively introducing more target fragments into the source.

The same behavior is observed for both language pairs and directions: on average, inserting random target fragments boosts the translation performance, with a larger payoff for the first few target segments. There exists an important gap for the output BLEU scores when CSW source sentences with different matrix languages reach the same (input) BLEU scores. Even though we generate a large number of replacements, the basic grammar structure of the matrix language is still maintained. Therefore, taking the target language as matrix gives the model a pre-translated sentence structure that is much easier to reproduce.

3.2.2 Implicit LID in translation

A second question concerns the ability of the translation system to identify target fragments in the source and to copy them in the target, even though these fragments are indistinguishable from genuine source segments. We use labels computed...
Figure 2: Evolution of the BLEU score of source CSW data and their target translation for En-Fr. (a) Direction CSW-En. The solid curve takes Fr as the matrix language, where we progressively inject more En segments; for the dash dot curve, En is the matrix language, with a growing number of Fr segments. (b) Direction CSW-Fr. Note that the target BLEU is always much higher than the source BLEU, with about a 20 points difference. The gap between the dash dot and solid curves is due to the basic sentence structure of the matrix language (see Section 3.2.1). As dash dot curves represent insertion in the reference target sentence, the corresponding BLEU score is always higher than the solid curve and actually reaches 100 (in the absence of any embedded language).

Refining the analysis, we also study whether the relative order of target words changes, or is preserved, during the translation. Table 3 reports the percentage of exact and switched-order copies. We observe again large differences with respect to the position of the matrix language. When the matrix language is the target language, the model always preserves the observed token order since it indicates a correct sentence structure for the hypothesis. When translating into the embedded language, we observe a larger number of word order changes: in this case, inserted target segments may not appear in their correct order in the CSW sentence, an issue that the model tries to fix. An example of this is in Figure 3, where we observe a swap between the input (“différent choix”) and output (“choix différent”) word orders.

Conversely, it is also interesting to look at the proportion of mixed language generated on the target side. Recall that in our training, the source is mixed-language, while the target is always monolingual. We use an in-house token-level language
identification (LID) model to identify the language of output tokens and to detect the CSW rate on the target side. As indicated in Table 2, our models generate almost pure monolingual translations, with a very low rate of CSW text. CSW-translation models thus seem to perform some language identification, as they almost perfectly sort out target language tokens (which are almost always copied) from the source language tokens (which are always translated).

A last issue concerns morphological errors: when inserting foreign words into a matrix source, one cannot expect to always also introduce the right inflection marks, some of which can only be determined once the target context is known. Another interesting phenomenon, that we do not simulate here, is when the embedded (target) lemma is adapted bears a morphological mark that only exist in the matrix language, which means that two linguistic systems are mixed within the same word, thereby posing more extreme difficulties for MT (Manandise and Gdaniec, 2011).

To illustrate the ability to correct grammar errors in input fragments, we manually noise a CSW sentence and display its translation in Figure 3. Where the input just contains the lemma of the French word “tenter” (to try), the model inserts a modal “doivent” to fix the context. Another illustration is for the adjective “différent” which is moved into post-nominal position, and for which an article (“un”) is inserted. This indicates that the model not only copies what already exists but also tends to adjust translations whenever necessary.

4 Computing translations in context

In this section, we evaluate CSW translation for the SemEval 2014 Task 5: L2 Writing Assistant (van Gompel et al., 2014), which can be handled as an MT task from mixed data.

4.1 Method

This task consists in translating L1 fragments in an L2 context, where the test set design is such that there is exactly one L1 insert in each utterance. We evaluated on two L1-L2 pairs: English-Spanish and French-English, and list below example test segments provided by the organizers for these pairs of languages (the insert and reference segments are in boldface):

- Input (L1=English,L2=Spanish): “Todo ello, in accordance con los principios que siempre hemos apoyado.”
  Output: “Todo ello, de conformidad con los principios que siempre hemos apoyado.”
- Input (L1=French,L2=English): “Je rentre à la maison because I am tired.”
  Output: “I return home because I am tired.”

The official metric for the SemEval evaluation is a word-based accuracy of the translations of the L1 fragment, which means that the L2 context of each sentence is not taken into account in scoring. Since our systems are full-fledged NMT systems, their output may not contain the reference L2 prefix and suffix. Therefore, two options are explored to compute these scores. The first is to post-process the output HYP and align it with the L2 reference context in REF. This alignment allows us to only score the relevant fragment in HYP. We refer to this option as free-dec.

The second option is to ensure that the L2 context will be present in the output translation. To this end, we use the force decoding mode of fairseq, implementing the methods of Post and Vilar (2018); Hu et al. (2019). We explored two different ways to express the L2 context as decoding constraints. The first turns every token in the L2 context as a separate constraint (token-cst). Continuing the previous example, “I, because, I, am, tired.” yield 5 constraints. The second uses the prefix and suffix of the L2 context as two multi-word constraints (presuf-cst). In this case, “I” and “because I am tired.” yield just 2 constraints. In both cases, constraints are required to be present in the prescribed order in the output.

4.2 Results

Scores are computed with the SemEval evaluation tool, which enables a comparison with other submissions for this task. Results are in Table 4 and 5. For En-Es, our CSW translator outperforms the best system in the official evaluation (van Gompel et al., 2014). Note that this model is not specifically designed nor tuned in any way for the SemEval task. For Fr-En, our system achieves better performance than the forth best participating system, with a clear gap with respect to the top results. In both cases, constraint decoding hurts performance: given that the automatic copy of target segments is already nearly perfect, introducing more constraints during
In Oregon, planners are experimenting with giving drivers different choices.

Dans l’Oregon, les planificateurs tentent l’expérience en offrant aux automobilistes différents choix.

In l’Oregon, planners tentent l’expérience with giving automobilistes différents choix.

Dans l’Oregon, les planificateurs doivent tenter l’expérience de donner à l’automobiliste un choix différent.

Figure 3: A noisy Code-Switched sentence with French as both the matrix and target language.

The search has here a clear detrimental effect for this task.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>B</th>
<th>M</th>
<th>B</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fr-En dev</td>
<td>97.3</td>
<td>75.6</td>
<td>97.6</td>
<td>76.4</td>
</tr>
<tr>
<td>Fr-En test</td>
<td>90.1</td>
<td>64.1</td>
<td>91.0</td>
<td>66.1</td>
</tr>
<tr>
<td>En-Es dev</td>
<td>97.4</td>
<td>98.8</td>
<td>97.6</td>
<td>99.0</td>
</tr>
<tr>
<td>En-Es test</td>
<td>89.9</td>
<td>95.3</td>
<td>90.4</td>
<td>95.5</td>
</tr>
</tbody>
</table>

Table 6: Results of other metrics on SemEval data.

METEOR scores for the Fr-En SemEval test are much worse than for En-Es. This is mostly due to the high “fragmentation penalty” computed by METEOR for English; the corresponding average $F_{mean}$ is about 0.99, showing that translations are mostly correct.

To better study the performance gap between these language pairs, we additionally score the development and test data with BLEU and METEOR. Results in Table 6 show that for these metrics, we achieve performance that are in that same ballpark for the two language pairs, suggesting that the observed difference in the SemEval metric is likely due to a mismatch between references and system outputs. The official metric is a word accuracy which may exclude acceptable translations by exact token match.

5 Related work

Research in the area of NLP for CSW has mostly focused on CSW Language Modeling, especially for Automatic Speech Recognition (Pratapa et al., 2018; Garg et al., 2018; Gonen and Goldberg, 2019; Winata et al., 2019; Lee and Li, 2020). Evaluation tasks, benchmarks have also been prepared for LID in user generated CSW content (Zubiaga et al., 2016; Molina et al., 2016), Named Entity Recognition (Aguilar et al., 2018), Part-of-Speech tagging (Ball and Garrette, 2018; Aguilar et al., 2020; Khanuja et al., 2020) and Sentiment Analysis (Patwa et al., 2020). CSW was also found useful in foreign language teaching: Renduchintala et al. (2019a,b) showed that replacing words by their counterparts in foreign language helps to learn foreign language vocabulary.

Regarding MT, most past work has focused on using artificial CSW data to help conventional translation systems. Huang and Yates (2014) used CSW corpus to improve word alignment and statistical MT. Dinu et al. (2019) experienced replacing and concatenating source terminology constraints by the corresponding translation(s) to boost the accuracy of term translations. Song et al. (2019a) shared the same idea by replacing phrases with pre-specified translation to perform “soft” constraint decoding. A different line of research is in (Bulte and Tezcan, 2019; Xu et al., 2020; Pham et al., 2020), who explore ways to combine a source sentence with similar translations extracted from translation memories. Yang et al. (2020) also pre-trained translation models by predicting original source segments from generated CSW sentences and claimed better results compared to other pre-
training methods (Conneau and Lample, 2019; Song et al., 2019b). Nevertheless, there barely exists work aimed at translating CSW sentences. Johnson et al. (2017) mentioned using a multilingual NMT system to translate CSW sentence to a third target language by showing only one example. To the best of our knowledge, only one parallel Arabic-English CSW corpus was specifically released for MT applications (Menacer et al., 2019). This CSW data was extracted from the UN data with Arabic as the matrix language: while translations into English were readily available, the purely Arabic side of the corpus was obtained using Google Translate to fill the missing Arabic bits.

6 Conclusion and outlook

In this study, we present a data augmentation method to generate artificial CSW data. We have shown that artificial data generated could be used to train NMT systems to translate both monolingual and CSW sentences (in one or even two different languages). With joint training of the two languages, we were able to build systems that were as good as a baseline bilingual system on monolingual texts, and much better for CSW texts. Our system does not need any explicit language identification and almost perfectly sorts out source tokens from target tokens in a CSW utterance. Another interesting feature of our system is that it always output monolingual translations. We finally report state-of-the-art results for the SemEval L2 Writing Assistant task for Es-En, while the related results for Fr-En are still somewhat lagging behind the best scores.

In the future, we would like to generate more realistic CSW data from monolingual sentences using a translation model. We also plan to explore ways to translate CSW texts simultaneously into both languages, so that the two decoding processes can mutually influence one another: in a first step in that direction, we have shown that training with a joint loss was actually beneficial for the translation into the two languages. Another line of research would be to continue experimenting with realistic language data, also containing other phenomena such as morphological binding. Finally, we also intend to study the somewhat more realistic condition where a mixture of languages A and B is translated into language C; we believe that the artificial CSW generation methods developed in our work would also be effective for this task.

7 Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2021-[AD011011580R1] made by GENCI. The authors wish to thank Josep Crego for his comments of an earlier version of this work. We also would like to thank the anonymous reviewers for their valuable suggestions. The first author is partly funded by Systran and by a grant from Région Ile-de-France.

References


Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019a. Code-switching for enhancing NMT with pre-specified translation. In...


