BiSync: A Bilingual Editor for Synchronized Monolingual Texts
Josep Crego, Jitao Xu, François Yvon

To cite this version:

HAL Id: hal-04163029
https://hal.science/hal-04163029
Submitted on 17 Jul 2023

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.

Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License
BiSync: A Bilingual Editor for Synchronized Monolingual Texts

Josep Crego‡ Jitao Xu† François Yvon§
‡SYSTRAN, 5 rue Feydeau, 75002 Paris, France
†NetEase YouDao, Beijing, China
§Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
firstname.lastname@{‡systrangroup.com,§cnrs.fr} xujt01@rd.netease.com

Abstract

In our globalized world, a growing number of situations arise where people are required to communicate in one or several foreign languages. In the case of written communication, users with a good command of a foreign language may find assistance from computer-aided translation (CAT) technologies. These technologies often allow users to access external resources, such as dictionaries, terminologies or bilingual concordancers, thereby interrupting and considerably hindering the writing process. In addition, CAT systems assume that the source sentence is fixed and also restrict the possible changes on the target side. In order to make the writing process smoother, we present BiSync, a bilingual writing assistant that allows users to freely compose text in two languages, while maintaining the two monolingual texts synchronized. We also include additional functionalities, such as the display of alternative prefix translations and paraphrases, which are intended to facilitate the authoring of texts. We detail the model architecture used for synchronization and evaluate the resulting tool, showing that high accuracy can be attained with limited computational resources. The interface and models are publicly available at https://github.com/jmcrego/BiSync and a demonstration video can be watched on YouTube.

1 Introduction

In today’s globalized world, there is an ever-growing demand for multilingual communication. To give just a few examples, researchers from different countries often write articles in English, international companies with foreign subsidies need to produce documents in multiple languages, research institutions communicate in both English and the local language, etc. However, for many people, writing in a foreign language (L2) other than their native language (L1) is not an easy task. With the significant advances in machine translation (MT) in the recent years, in particular due to the tangible progress in neural machine translation (NMT, Bahdanau et al., 2015; Vaswani et al., 2017), MT systems are delivering usable translations in an increasing number of situations. However, it is not yet realistic to rely on NMT technologies to produce high quality documents, as current state-of-the-art systems have not reached the level where they could produce error-free translations. Also, fully automatic translation does not enable users to precisely control the output translations (e.g. with respect to style, formality, or term use). Therefore, users with a good command of L2, but not at a professional level, can find help from existing computer-assisted language learning tools or computer-assisted translation (CAT) systems. These tools typically provide access to external resources such as dictionaries, terminologies, or bilingual concordancers (Bourdaillet et al., 2011) to help with writing. However, consulting external resources causes an interruption in the writing process due to the initiation of another cognitive activity, even when writing in L1 (Leijten et al., 2014). Furthermore, L2 users tend to rely on L1 (Wolfsberger, 2003) to prevent a breakdown in the writing process (Cumming, 1989). To this end, several studies have focused on developing MT systems that ease the writing of texts in L2 (Koehn, 2010; Huang et al., 2012; Venkatapathy and Mirkin, 2012; Chen et al., 2012; Liu et al., 2016).

However, existing studies often assume that users can decide whether the provided L2 texts precisely convey what they want to express. Yet, for users who are not at a professional level, the evaluation of L2 texts may not be so easy. To mitigate this issue, researchers have also explored round-trip translation (RTT), which translates the MT output in L2 back to L1 in order to evaluate the quality of L2 translation (Moon et al., 2020). Such studies suggest that it is then helpful to augment L2 writing
with the display of the corresponding synchronized version of the L1 text, in order to help users verify their composition. In such settings, users can obtain synchronized texts in two languages, while only making an effort to only compose in one.

A bilingual writing assistant system should allow users to write freely in both languages and always provide synchronized monolingual texts in the two languages. However, existing systems do not support both functionalities simultaneously. The system proposed by Chen et al. (2012) enables free composition in two languages, but only displays the final texts in L2. Commercial MT systems like Google, DeepL, and SYSTRAN always display texts in both languages, but users can only modify the source side, while the target side is predicted by the system and is either locked or can only be modified with alternative translations proposed by the system. CAT tools, on the contrary, assume the source sentence is fixed and only allow edits on the target side.

In this paper, we present BiSync, a bilingual writing assistant aiming to extend commercial MT systems by letting users freely alternate between two languages, changing the input text box at their will, with the goal of authoring two equally good and semantically equivalent versions of the text.

2 BiSync Text Editor

In this work, we are interested in a writing scenario that broadens the existing commercial online translation systems. We assume that the user wants to edit or revise a text simultaneously in two languages. See Figure 1 for a snapshot of our BiSync assistant. Once the text is initially edited in one language, the other language is automatically synchronized so that the two entry boxes always contain mutual translations. In an iterative process, and until the user is fully satisfied with the content, texts are revised in either language, triggering automatic synchronizations to ensure that both texts remain mutual translations. The next paragraphs detail the most important features of our online BiSync text editor.

**Bidirectional Translations** The editor allows users to edit both text boxes at their will. This means that the underlying synchronization model has to perform translations in both directions, as the role of the source and target texts are not fixed and can change over time.

**Synchronization** This is the most important feature of the editor. It ensures that the two texts are always translations of each other. As soon as one text box is modified, BiSync synchronizes the other box. To enhance the user experience, the system waits a few seconds (delay) before the synchronization takes place. When a text box is modified, the system prevents the second box from being edited until the synchronization has been completed. Users can also disable the synchronization process, using the "freeze" button ( ). In this case, the frozen text will not be synchronized (modified). Changes are only allowed in the unfrozen text box. This is the standard modus operandi of most commercial translation systems that consider the input text as frozen, allowing only a limited number of edits in the translation box.

**Prefix Alternatives** The editor can also provide several translation alternatives for a given sentence prefix. When users click just before a word w in a synchronized sentence pair, the system displays the most likely alternatives that can complete the translation starting from the word w in a drop-down menu. Figure 2 (bottom) illustrates this functionality, where translation alternatives are displayed after the prefix "Je rentre", in the context of the English sentence "I'm going home because I'm tired". In the example in Figure 2 (bottom), the user clicked right before the French word "à".

**Paraphrase Alternatives** Another important feature of our BiSync editor is the ability to propose edits for sequences of words at arbitrary positions in both text boxes. Figure 2 (top) illustrates...
this scenario, where paraphrases for the English segment "going home" are displayed in the context "I'm ... because I'm tired" and given the French sentence "Je rentre à la maison parce que je suis fatigué". Such alternatives are triggered through the selection of word sequences in either text box.

Other Features Like most online translation systems, BiSync has a "listen" button that uses a text-to-speech synthesizer to read the content in the text box, a "copy" button that copies the content to the clipboard. It also displays the number of characters written in each text box. Figures 1 and 2 illustrate these features.

Settings The "gear" button (⚙️) pops up several system parameters that can be configured by users: The "IP" and "Port" fields identify the address where the BiSync model is launched and waits for translation requests. The "Languages" field indicates the pair of languages that the model understands and is able to translate. "Alternatives" denotes the number of hypotheses that the model generates and that are displayed by the system. "Delay" sets the number of seconds the system waits after an edit takes place before starting the synchronization. The countdown is reset each time a revision is made. Figure 3 displays BiSync settings with default parameters.

![Image of BiSync settings](image)

Figure 3: Default BiSync settings.

3 Under the Hood: the BiSync Model

Given the features that we would like to offer in the BiSync text editor, we are interested in an end-to-end model producing: (a) translations in both directions; (b) translations from scratch, with only one text box filled in; (c) updates of existing translations, so that small changes in one text box result in small updates in the other text box; (d) paraphrases and alternatives in context.

We consider a pair of parallel sentences \((f, e)\) and a sentence \(f'\) as an update of \(f\). The objective is to generate the sentence \(e'\) that is parallel to \(f'\) while remaining as close as possible to \(e\). Three types of update are distinguished. Figure 4 displays an example for each update type:

- **Insertion**: adding one or more consecutive words at any position in \(f\);
- **Deletion**: removing one or more consecutive words at any position in \(f\);
- **Substitution**: replacing one or more consecutive words at any position in \(f\) by one or more consecutive words.

Note that in practice, training such models requires triplets \((f', e, e')\), as sentences \(f\) are not used by the models studied in this work.

Inspired by Xiao et al. (2022), we integrate several control tokens into the source-side of training examples of a standard NMT model to obtain the desired results. This approach is straightforward and does not require to modify NMT architectures or decoding algorithms. Therefore, our integration of control tokens is model-agnostic and can be applied to any NMT architecture. Several tokens are used to indicate the target language (\(<en>\), \(<fr>\))
and the update type (<ins>, <del>, <sub>). A special token <gap> indicates a sequence of masked tokens used to generate paraphrases.

Training examples are built following the next three patterns:

\[
\begin{array}{c|c|c}
\text{f'} & \text{update} & \text{e'} \\
\end{array}
\]

The first pattern refers to a regular translation task (TRN), and is used when translating from scratch, without any initial sentence pair (f, e), as in the following example:

*The white cat* <fr> *Le chat blanc* where only the target language tag is appended to the end of the source sentence.

\[
\begin{array}{c|c|c|c}
\text{f'} & \text{<lang>} & \text{e} & \text{<update>} & \text{e'} \\
\end{array}
\]

The second pattern corresponds to an update task (INS, DEL or SUB) to be used for resynchronizing an initial sentence pair (f, e) after changing f into f', as shown in the following examples:

\[
\begin{array}{c|c|c|c|c|c|c|c}
\text{The white cat} & \text{<fr>} & \text{Le chat} & \text{<ins>} & \text{Le chat blanc} \\
\text{The white cat} & \text{<fr>} & \text{Le chat est blanc} & \text{<del>} & \text{Le chat blanc} \\
\text{The white cat} & \text{<fr>} & \text{Le chat noir} & \text{<sub>} & \text{Le chat blanc} \\
\end{array}
\]

where the edited source sentence f' = [The white cat] is followed by the target language tag <fr>, the initial target sentence e, and a tag indicating the edit type that updates the initial source sentence f.

\[
\begin{array}{c|c|c|c|c}
\text{f} & \text{<lang>} & \text{e}_g & \text{e}_G & \text{e'} \\
\end{array}
\]

The third pattern corresponds to a bilingual text infilling task (BTI, Xiao et al., 2022). The model is trained to predict the tokens masked in a target sentence e_g in the context of the source sentence f:

\[
\begin{array}{c|c|c|c|c}
\text{The white cat} & \text{<fr>} & \text{Le <gap> blanc} & \text{chat} & \text{e'} \\
\end{array}
\]

where \(e_g = [\text{Le <gap> blanc}]\) is the target sentence with missing tokens to be predicted. The model only generates the masked tokens \(e_G = [\text{chat}]\).

### 3.1 Synthetic Data Generation

While large amounts of parallel bilingual data \((f', e')\) exist for many language pairs, the triplets required to train our model are hardly available. We therefore study ways to generate synthetic triplets of example \((f', e, e')\) from parallel data \((f', e')\) for each type of task (INS, DEL, SUB and BTI) introduced above.

**Insertion** We build examples of initial translations e for INS by randomly dropping a segment from the updated target e'. The length of the removed segment is also randomly sampled with a maximum length of 5 tokens. We also impose that the overall ratio of removed segment does not exceed 0.5 of the length of e'.

**Deletion** Simulating deletions requires the initial translation e to be an extension of the updated target e'. To obtain extensions, we employ a NMT model enhanced to fill in gaps (fill-in-gaps). This model is a regular encoder-decoder Transformer model trained with a balanced number of regular parallel examples (TRN) and paraphrase examples (BTI) as detailed in the previous section. We extend training examples \((f', e')\) with a <gap> token inserted in a random position in e' and use fill-in-gaps to decode these training sentences, as proposed in (Xu et al., 2022). In response, the model predicts tokens that best fill the gap. For instance:

*The white cat* <fr> *Le chat* <gap> blanc \(\sim\) est

the target extension is therefore e = [Le chat est blanc].

**Substitution** Similar to deletion, substitution examples are obtained using the same fill-in-gaps model. A random segment is masked from e', which is then filled by the model. In inference, the model computes an n-best list of substitutions for the mask, and we select the most likely sequence that is not identical to the masked segment. For instance:

*The white cat* <fr> *Le chat* <gap> \(\sim\) [blanc; bleu; clair; blanche; ...]

the target substitution is e = [Le chat bleu].
Note that extensions and substitutions generated by fill-in-gaps may be ungrammatical. For instance, the proposed substitution $e = [Le \; chat \; blanche]$ has a gender agreement error. The correct adjective should be "blanc" (masculine) instead of "blanche" (feminine). This way, the model always learns to produce grammatical $e'$ sentences parallel to $f'$.

**Paraphrase** Given sentence pairs $(f, e)$, we generate $e_c$ by masking a random segment from the initial target sentence $e$. The length of the masked segment is also randomly sampled with a maximum length of 5 tokens. The target side of these examples $(e_c)$ only contains the masked token(s).

## 4 Experiments

### 4.1 Datasets

To train our English-French (En-Fr) models we use the official WMT14 En-Fr corpora as well as the OpenSubtitles corpus (Lison and Tiedemann, 2016). A very light preprocessing step is performed to normalize punctuation and to discard examples exceeding a length ratio 1.5 and a limit of $[1, 250]$ measured in words. Statistics of each corpus is reported in Table 1.

<table>
<thead>
<tr>
<th>Corpora</th>
<th>#Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europarl v7</td>
<td>2,007,723</td>
</tr>
<tr>
<td>Commoncrawl</td>
<td>3,244,152</td>
</tr>
<tr>
<td>UN</td>
<td>12,886,831</td>
</tr>
<tr>
<td>News Commentary</td>
<td>183,251</td>
</tr>
<tr>
<td>Giga French-English</td>
<td>22,520,376</td>
</tr>
<tr>
<td>Open Subtitles v18</td>
<td>57,123,540</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>97,965,873</strong></td>
</tr>
</tbody>
</table>

Table 1: Statistics of training corpora.

For testing, we used the official newstest2014 En-Fr test set made available for the same WMT14 shared task containing 3,003 sentences. All our data is tokenized using OpenNMT tokenizer. We learn a joint Byte Pair Encoding (Sennrich et al., 2016) over English and French training data with 32k merge operations.

The training corpora used for learning our model consist of well-formed sentences. Most sentences start with a capital letter and end with a punctuation mark. However, our BiSync editor expects also incomplete sentences, when synchronization occurs before completing the text. To train our model to handle this type of sentences, we lowercase the first character of sentences and remove ending punctuation in both source and target examples with a probability set to 0.05.

### 4.2 Experimental Settings

Our BiSync model is built using the Transformer architecture (Vaswani et al., 2017) implemented in OpenNMT-tf (Klein et al., 2017). More precisely, we use the following set-up: embedding size: 1,024; number of layers: 6; number of heads: 16; feedforward layer size: 4,096; and dropout rate: 0.1. We share parameters for input and output embedding layers (Press and Wolf, 2017). We train our models using Noam schedule (Vaswani et al., 2017) with 4,000 warm-up iterations. Training is performed over a single V100 GPU during 500k steps with a batch size of 16,384 tokens per step. We apply label smoothing to the cross-entropy loss with a rate of 0.1. Resulting models are built after averaging the last ten saved checkpoints of the training process. For inference, we use CTranslate2. It implements custom run-time with many performance optimization techniques to accelerate decoding execution and reduce memory usage of models on CPU and GPU. We also evaluate our model with weight quantization using 8-bit integer (int8) precision, thus reducing model size and accelerating execution compared to the default 32-bit float (float) precision.

## 5 Evaluation

We evaluate the performance of our synchronization model BiSync compared to a baseline translation model with exactly the same characteristics but trained only on the TRN task over bidirectional parallel data (base). We report performance with BLEU score (Papineni et al., 2002) implemented in SacreBLEU (Post, 2018) over concatenated En-Fr and Fr-En test sets. For tasks requiring an initial target $e$, we synthesize $e$ from $(f', e')$ pairs following the same procedures used for generating the training set (see details in Section 2).

Table 2 reports BLEU scores for our two systems on all tasks. The base system is only trained to...
perform regular translations (TRN) for which it obtains a BLEU score of 36.0, outperforming BiSync, which is trained to perform all tasks. This difference can be explained by the fact that BiSync must learn a larger number of tasks than base. When performing INS, DEL and SUB tasks, BiSync vastly outperforms the results of TRN task as it makes good use of the initial translation $e$. When we use BiSync to generate paraphrases of an initial input (BTI), we obtain a higher BLEU score of 42.6 than the regular translation task (TRN, 34.9). This demonstrates the positive impact of using target side context for paraphrasing.

<table>
<thead>
<tr>
<th>BLEU</th>
<th>TRN</th>
<th>INS</th>
<th>DEL</th>
<th>SUB</th>
<th>BTI</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>36.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BiSync</td>
<td>34.9</td>
<td>87.9</td>
<td>95.5</td>
<td>78.2</td>
<td>42.6</td>
</tr>
</tbody>
</table>

Table 2: BLEU scores over concatenated En-Fr and Fr-En test sets for all tasks.

Next, we evaluate the ability of our BiSync model to remain close to an initial translation when performing synchronization. Note that for a pleasant editing experience, synchronization should introduce only a minimum number of changes. Otherwise, despite re-establishing synchronization, additional changes may result in losing updates previously performed by the user. To evaluate this capability of our model, we take an initial translation $(f, e)$ and introduce a synthetic update (say $f'$) as detailed in Section 3.1. This update leads to a new synchronization that transforms $e$ into $e'$. We would like $e'$ to remain as close as possible to $e$. Table 3 reports TER scores (Snover et al., 2006) between $e$ and $e'$ computed by SacreBLEU.\footnote{Signature: nrefs:1case:lc/tok:tercomlnorm:npunct:yes asian:nolversion:2.0.0} These results indicate that BiSync produces synchronizations significantly closer to initial translations than those produced by base. This also confirms the findings of Xu et al. (2022).

<table>
<thead>
<tr>
<th>TER ↓</th>
<th>INS</th>
<th>DEL</th>
<th>SUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>base</td>
<td>36.5</td>
<td>43.4</td>
<td>34.9</td>
</tr>
<tr>
<td>BiSync</td>
<td>3.3</td>
<td>5.5</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Table 3: TER scores between $e$ and $e'$ issued from different update types. En-Fr and Fr-En test sets are concatenated.

Finally, Table 4 reports inference efficiency for our BiSync model using CTranslate2. We indicate differences in model (Size and Speed) for different quantization, device, batch size and number of threads. Lower memory requirement and higher inference speed can be obtained by using quantization set to int-8 for both GPU and CPU devices, in contrast to float-32. When running on CPUs, additional speedup is obtained with multithreading. Comparable BLEU scores are obtained in all configurations. Note that for the tool presented in this paper, we must retain single batch size and single thread results (bold figures), since synchronization requests are produced for isolated sentences. Therefore, they cannot take advantage of using multiple threads and large batches.

<table>
<thead>
<tr>
<th>Quant</th>
<th>Dev</th>
<th>BS</th>
<th>Threads</th>
<th>Size</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>float</td>
<td>GPU</td>
<td>64</td>
<td>1</td>
<td>10,267</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td>GPU</td>
<td>1</td>
<td>1</td>
<td>232M</td>
<td>2,007</td>
</tr>
<tr>
<td></td>
<td>CPU</td>
<td>64</td>
<td>8 x 4</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>int8</td>
<td>GPU</td>
<td>64</td>
<td>1</td>
<td>12,918</td>
<td>738</td>
</tr>
<tr>
<td></td>
<td>GPU</td>
<td>1</td>
<td>1</td>
<td>59M</td>
<td>738</td>
</tr>
<tr>
<td></td>
<td>CPU</td>
<td>64</td>
<td>8 x 4</td>
<td>2,666</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>CPU</td>
<td>1</td>
<td>1</td>
<td>118</td>
<td>118</td>
</tr>
</tbody>
</table>

Table 4: Inference Speed and model Size when decoding test sets with several settings: quantization (Quant), device (Dev), batch size (BS) and number of Threads. Decoding beam size is set to 3. Speed is measured in tokens/second. GPU is a single V100 GPU with 32Gb memory. CPU has 32 cores with 86Gb memory and CPU$^2$ is an Intel i7-10850H with 32Gb memory.

6 Conclusion and Further Work

In this paper, we presented BiSync, a bilingual writing assistant system that allows users to freely compose text in two languages while always displaying the two monolingual texts synchronized with the goal of authoring two equally good versions of the text. Whenever users make revisions on either text box, BiSync takes into account the initial translation and reduces the number of changes needed in the other text box as much as possible to restore parallelism. BiSync also assists in the writing process by suggesting alternative reformulations for word sequences or alternative translations based on given prefixes. The synchronization process applies several performance optimization techniques to accelerate inference and reduce the memory usage with no accuracy loss, making BiSync usable even on machines with limited computing power.

In the future, we plan to equip BiSync with a
grammatical error prediction model and a better handling of prior revisions: the aim is to enable finer-grained distinction between parts that the system should modify and parts that have already been fixed or that should remain unchanged. Last, we would like to perform user studies to assess the division of labor between users and BiSync in an actual bilingual writing scenario.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable suggestions and comments. This work was performed while the last two authors where affiliated with LISN, CNRS (Orsay).

Jitao Xu has been partly funded by SYSTRAN and by a grant (Transwrite) from the Région Ile-de-France. He was granted access to the HPC resources of IDRIS under the allocation 2022-AD011011580R2 made by GENCI.

References


