# Assessing Back-Translation as a Corpus Generation Strategy for non-English Tasks: A Study in Reading Comprehension and Word Sense Disambiguation

**Fabricio Monsalve**[✠*] **Kervy Rivas-Rojas**[✠*]
**Marco Antonio Sobrevilla Cabezudo**[♣] **Arturo Oncevay**[♢♠]
[✠] Artificial Intelligence Research Group, Pontificia Universidad Católica del Perú
[♣] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
[♠] School of Informatics, University of Edinburgh
f.monsalve@pucp.edu.pe, k.rivas@pucp.pe

## Abstract

Corpora curated by experts have sustained Natural Language Processing mainly in English, but the expensiveness of corpora creation is a barrier for the development in further languages. Thus, we propose a corpus generation strategy that only requires a machine translation system between English and the target language in both directions, where we filter the best translations by computing automatic translation metrics and the task performance score. By studying Reading Comprehension in Spanish and Word Sense Disambiguation in Portuguese, we identified that a more quality-oriented metric has high potential in the corpora selection without degrading the task performance. We conclude that it is possible to systematise the building of quality corpora using machine translation and automatic metrics, besides some prior effort to clean and process the data.

## 1 Introduction

Available data has allowed a steady improvement in Natural Language Processing (NLP) tasks for English. Nevertheless, English is not the broadest native language spoken in the world. According to Ethnologue (Simons and Fenning, 2019), English ranks third, behind Chinese (Mandarin) and Spanish, and is only one of the approximately 7000 currently spoken languages. The relevance of English as the academically universal language has allowed its growth in computational linguistic resources. Even in languages with a large number of speakers, such as Spanish, it is difficult to find specific NLP tools that match the quality and performance as in English. If we want to replicate the development of state-of-the-art models for other languages, we would need large and high-quality corpora analogous to the English ones, and their creation cost would be prohibitive.

In this context, there is a very compelling tool that has reached several languages in commercial systems: Machine Translation (MT). However, it is worth noting that MT works for language-pairs, and therefore, most of the commercial MT tools have obtained excellent results mostly with English as the source or target language. Thus, we still need English in search of robust NLP tools, but at least there is potential for obtaining new data for new languages using high-quality MT systems. As other studies have been focusing on (see §2), we can translate task-specific corpora from English to other languages to leverage an NLP tool without the need of experts in the target language.

Under those circumstances, the next question arises: how can we guarantee the quality of the new corpus by using automatic translations and without recurring to manual validation? Previous work used quality estimation metrics from machine translation, mostly BLEU (Papineni et al., 2002), by applying back-translation and performing the quality evaluation in English. However, we are concerned about the deficiency of using only BLEU as a measurement of a correct translation (Callison-Burch et al., 2006) or text generation in general (Novikova et al., 2017), and currently there are other proposed metrics to cover the correlation gap between BLEU and a human assessment (Denkowski and Lavie, 2014; Fomicheva et al., 2016, among others). Therefore, we believe there is space for improvement in the quality assessment of a back-translation application to the generation of new corpora.

Our study and contribution are not focused in obtaining state-of-the-art results for new languages, but to obtain a new quality corpus that could be used to build state-of-the-art models, such as deep neural networks (Sutskever et al.,

[*] Equal contribution



*Proceedings of the 13th Linguistic Annotation Workshop*, pages 81–89
Florence, Italy, August 1, 2019. 

2014), in new languages. However, we also managed to surpass previous methods on the target languages in monolingual scenarios.

More details about related works are described in §2. Then, we present our methodology for corpus generation in §3, where we also introduce our case studies in Word Sense Disambiguation for Portuguese and Reading Comprehension for Spanish. Furthermore, §4 contains an extrinsic evaluation of the corpora in their respective task. Also, we make publicly available specific code and guidelines to build the new corpora from the original sources[1]. The obtained results enlightens a potential systematisation of new corpora generation for many language-related tasks and opens further work on generalisation and truly low-resource settings.

## 2 Background

Several strategies have been applied to build corpora for different tasks in non-English languages and, thus, to reduce the manual work. Mainly, Machine Translation-based approaches had succeeded in obtaining annotated corpora. A key point to highlight is that results from this approach depend on the availability of an MT system, the quality of the acquired translations, and the precision of the alignments between the two languages (English and non-English).

Jabaian et al. (2011) focused on applying a Phrase-based MT (PBMT) system to deal with the language portability of dialogue systems, whereas Klinger and Cimiano (2015) focused on using PBMT and some quality estimation measures to select the best translations which make up the corpus for the task of Sentiment Analysis. Also, Koehn et al. (2018) reports other works related to corpora selection but for a shared task of parallel corpora filtering, to train better machine translation with fewer noisy data.

Furthermore, back-translation strategies have emerged to improve the quality of corpus in a target language (a non-English language). Misu et al. (2012) used back-translation results to verify whether the translation keeps the semantic meaning of the original sentence in a Spoken Language Understanding System, and they also disregarded BLEU as a good quality measure. Besides, Gaspers et al. (2018) considered metrics from alignments, machine translation and language model as a measure of MT quality, independent of the Natural Language Understanding tasks and, thus, select the best sentences to incorporate into the corpus. Finally, Asai et al. (2018) explored Neural MT models to build a Reading Comprehension model for Japanese and French using English as a source. They consider back-translation as their baseline, and they build a multi-lingual model to assess the task. Our motivation differs from them, as we want to generate large quality corpora that help to build monolingual systems, which can achieve great state-of-the-art results, alike for English.

Previous studies show that translating, automatically measuring the translation, and selecting the best samples are not entirely innovative procedures. However, we want to achieve a systematisation for this procedure and look for general-purpose steps disregarding the NLP task and the language. Next section develops our idea.

## 3 Methodology

We introduce our strategy on back-translation and automatic assessment in a general overview. Then, we extend details specifically for our two case studies: Word Sense Disambiguation and Reading Comprehension. The procedure for the corpus generation and the evaluation (§4) is summarised in Figure 1.

### 3.1 Back-Translation Strategy

Our goal, similar to previous studies (Misu et al., 2012; Gaspers et al., 2018), is to choose the best translations from an automatically translated corpus to train a robust NLP model. For the following description, we consider English as our source language, whereas the target language could be anyone with an MT system available in both directions with English.

If we take a corpus for any task in English NLP and translate it to a new language, we are not going to be able to measure the translation quality in the target language itself due to the lack of a reference translation. Therefore, we automatically translate the text back to English (back-translation[2]) to measure if the semantic information of the source is retained after the process of two automatic translations. For that purpose, we consider that only

[1] https://github.com/iapucp/backcorp

[2] The term has been proposed by Sennrich et al. (2016) in MT, to provide monolingual training data by automatically translate a target sentence into the source language.

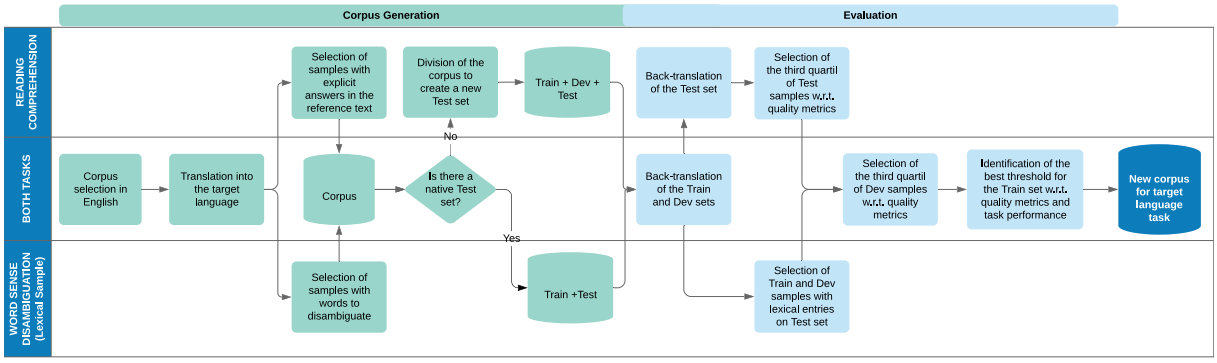


Figure 1: Work-flow of the back-translation strategy with automatic assessment for the generation of corpora for tasks in new languages. We divide general and task-specific steps, as well as the corpus processing and quality evaluation procedures.

BLEU is not a sufficient metric; thus, we attempt the comparison with different approaches.

### 3.2 Automatic Quality Assessment

Given an automatic translation metric, we can compute the score between the source references and the back-translation. We differentiate our work from Misu et al. (2012) by using general metrics, and not task-related ones, to assess quality in the selected translations. However, for this study, we constrained the experiment in the comparison of two word-based/n-gram coverage metrics[3]. We want to evaluate if there is a difference between a baseline metric and one with a higher correlation with human assessment in translation. For the former, we combine BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) in an F-score ($F_{B+R}$)[4], whereas we use the last version of Meteor Universal ($M.U.$) (Denkowski and Lavie, 2014) for the latter.

At this point, we hypothesise as follows: by using one of the metrics mentioned previously, we could extract a good quality corpus if we identify a threshold in the distribution of the translation scores that obtains the best performance score given a test for an extrinsic task related to the corpus. Furthermore, as an extrinsic evaluation, we are going to compare the threshold-based extraction with a random choice of the same corpus size.

#### 3.2.1 Train Set Filtering w.r.t. Metrics

The primary goal is to identify where should be located the best threshold to filter out bad-quality translations. Besides, we can explore whether the quality is more relevant than the potentially-noisy large number of samples to train a model.

Thus, we split the training set in progressive cuts, ranging from top-20% to 70%. We rank the training samples concerning three criteria: $F_{B+R}$, $M.U.$ and a random seed. Therefore, we are going to have several trained models for each cut and criteria to extrinsically measure the quality of the corpus from the performance task.

#### 3.2.2 Development Set Processing

As we want to generate large corpora able to be processed by complex learning algorithms, we require a development (dev) set for our experiments. There is a possibility to filter the development set similar to the train set, but we want to constraint the variable of corpus selection only to the threshold of the train set. However, it is relevant to guarantee high-quality content, so we decided to constrain its content with potentially good translations only regarding our quality metrics:

1. We compute the metrics $F_{B+R}$ and $M.U.$ for all the samples in the development set.
2. For each metric, we obtain the third quartile and intersect both sub-sets.

### 3.3 NLP Tasks and Target Languages

We tested our methodology on two tasks and two languages: Reading Comprehension for Spanish and Word Sense Disambiguation for Portuguese. Both languages are ranked within the top-ten languages with more first-native speakers (Simons

[3]We also tried to distinguish back-translation quality by using pre-trained English document vectors, but the distribution of the scores was not of much use because the $std(\sigma)$ was very small and the mean score was near 1

[4]For this decision, we consider that BLEU and ROUGE complement each other as precision and recall but for measuring overlapping n-grams. Also, we analysed the distribution of the metrics using the formula $\alpha * BLEU + (1 - \alpha) * ROUGE, \quad \alpha \in [0.1, 0.9]$ and found that both of them reached the best F-score when $\alpha = 0.5$, so both had the same importance in the experiment.

| | train | dev | test |
|---|---|---|---|
| SQuAD 1.1 (en) | 88,013 | 10,570 | - |
| SQuAD→es | 62,893 | 6,995 | - |
| SQuAD→es(w/test) | 57,232 | 6,303 | 6,353 |

Table 1: Corpus size for Reading Comprehension. SQuAD 1.1 (en) is translated into Spanish (es), questions without explicit answers are dropped, and the corpus is split to generate a new test.

| | dev | test |
|---|---|---|
| Original | 6,303 | 6,353 |
| Filtered | 1,045 | $Q1 \rightarrow 1,956$ |
| | | $Q2 \rightarrow 1,087$ |
| | | $Q3 \rightarrow 409$ |

Table 2: Size of development and test sets for the evaluation of Reading Comprehension in Spanish

and Fenning, 2019), and they are regularly studied in specific NLP research communities (Portugal, Spain and Latin-America). There are many core NLP tools for both languages, such as morphological analysis, POS-tagging, syntax dependency parsing, word sense disambiguation, among others. However, their performance is not at the same level as their English counterparts, and it is less probable to identify more complex NLP tools such as reading comprehension. There is an exception for machine translation although, as we can find commercial MT systems for both languages to translate from and into English.

#### 3.3.1 Reading Comprehension (es)

In reading comprehension, the fundamental goal is to identify the position of an answer in a reference text given a question. The Stanford Question Answering Dataset (SQuAD; Rajpurkar et al., 2016) is the most famous corpus to evaluate new methods, with more than 80,000 question-answer pairs extracted from Wikipedia documents, but only available in English.

There is not a corpus with the same properties in Spanish. Previous Question-Answering (QA) challenges in Spanish, mainly hosted by the CLEF initiative, consider the extraction of text references from the web before the identification of the answer itself[5]. The most similar datasets were presented in the Question Answering for Machine Reading Evaluation tasks (QA4MRE; Peñas et al., 2013), but they were relatively small and the corpora require additional steps to be entirely similar to the SQuAD task. Nonetheless, the datasets could be processed for future experiments as testing sets directly built in the target language (es).

Therefore, we applied the back-translation strategy to generate a new Reading Comprehension corpus for Spanish. We only use the train and development sub-datasets from the English SQuAD, as the test is not available. Thus, we extract a sample from the train and development to generate a new test for our experiments. Then, we translate the corpus to Spanish and back to English using the Google Translate API[6], and drop the questions that lost their exact answers in the reference. In other words, we do not preserve the samples where the translated answer is not exactly contained in the translated reference. See Table 1 for corpus size details.

In the construction of the dataset, we have already disregarded low-quality translations to preserve the nature of the task (we need an explicit answer in the reference text). Thus, we expect a great difficulty to surpass the proposed random baseline in the selection of the best translation, as there would mostly be high-quality translations to choose. Therefore, there is a must to accompany this study with a different task, to drive more general conclusions from the experimentation.

Furthermore, we must assume that the extracted test set contains high and low-quality translations, as a random seed split it. Thus, we divide our test into quartiles for evaluation purposes with a metric based on $F_{B+R}$ and $M.U.$. We followed a similar process as in the filtering of the development set (see §3.2.2). Table 2 shows the filtered size of the dev set, as well of the different partitions of the test w.r.t. to the quality metrics.

#### 3.3.2 Word Sense Disambiguation (pt)

The ambiguity arises from a linguistic problem that occurs in the language, because a word may assume different meanings depending on the specific context where it is used. In that sense, Word sense disambiguation (WSD) is the task that aims to determine the correct sense of a word given a specific context using a pre-specified sense-repository (Agirre and Edmonds, 2007).

For WSD, there is a considerable amount of English language data; however, they are not avail-

[5]Restricted access: http://catalog.elra.info/en-us/repository/browse/ELRA-E0038/

[6]https://cloud.google.com/translate/

| corpus | number of sentences |
|---|---|
| OMSTI | 813,798 |
| SemCor | 37,176 |
| Senseval-2 | 242 |
| Senseval-3 | 352 |
| SemEval-07 | 135 |
| SemEval-13 | 306 |
| SemEval-15 | 1,138 |
| Total | 852,147 |

Table 3: Corpus size details for the Unified Evaluation Framework or UEF (en)

| | train |
|---|---|
| UEF (en) | 852,147 |
| UEF en→pt (partial trans.) | 73,784 |
| UEF en→pt (after filtering) | 14,376 |

Table 4: Corpus size for Word Sense Disambiguation. The Unified Evaluation Framework or UEF (en) is partially translated into Portuguese (pt), and then we only preserved the samples with one-to-one alignments of ambiguous words.

able data or comparable data (in terms of size) in other languages, such as Portuguese. Thus, we decide to apply back-translation to generate new corpora. Nevertheless, there is a specific problem, as the disambiguation corpus in English may be found in different versions of Wordnet. To overcome this issue, we use the Unified Evaluation Framework (UEF) of Raganato et al. (2017)[7], which includes an standardised corpora aligned with Wordnet 3.0 (Miller, 1995). See Table 3 for corpus size details about the corpora.

In Portuguese, there is an annotated and native WSD corpus: the CSTNews (Cardoso et al., 2011). This is a multi-document corpus composed of 140 news texts (in Brazilian Portuguese) and grouped by 50 collections. The texts in any collection belong to the same topic. Besides, there was an extended annotation for several verbs (Cabezudo et al., 2015), using WordNet 3.0 as sense-repository. In total, there are 5,082 annotated verb instances with 844 different verbs and 1,047 synsets (senses).

Because the CSTNews corpus is a curated corpus in Portuguese, it is convenient to use it as test data, and we do not need to generate a new test set similar to the Reading Comprehension case. So, all the translated sentences from UEF could be used as training and development sets. However, there is an additional consideration for this task if we want to perform an external evaluation later.

To obtain the final set of sentences for the corpus, we follow a two-step procedure. Firstly, we used the Yandex API[8] to partially translate the texts into Portuguese[9]. We decide to use this API due to its provision of word alignments. Secondly, we deal with the alignments between English and Portuguese sentences, as we were only interested in the sentences with one-to-one word alignments for the words to disambiguate. Then, we disregard the samples with many-to-many relationships between ambiguous words, as it could carry some mistakes in the task. Corpus size is detailed in Table 4. Finally, we apply the procedure describe in §3.2.2, generating 10,592 sentences in the training set and 3,784 sentences in the development set.

## 4 Extrinsic Automatic Evaluation

We evaluate each generated corpus by measuring the task performance in a specific test set for the target language. We restrict our experiments in monolingual setups to control the identification of potential results.

### 4.1 Reading Comprehension (es)

With the newly translated corpus, we can evaluate more complex data-driven algorithms, such as deep neural networks. Thus, we adopt the method from Chen et al. (2017)[10] into Spanish, by using pre-trained language-specific models to perform named-entity recognition and part-of-speech tagging from spaCy[11], as well as pre-trained GloVe (Pennington et al., 2014) Spanish word vectors from the Spanish Billion Corpus (Cardellino, 2016). The basic network architecture was not changed and is a sequence-to-sequence with a hidden layer size of 128 and 300-dimensional embedding. We only updated parts of pre-processing modules to work for Spanish.

Following the train set filtering described in §3.2.1, we trained a QA model for each segment of the data and each criterion. We validated the

[7] http://lcl.uniroma1.it/wsdeval/

[8] https://tech.yandex.com/translate/

[9] Due to computational reasons, we were not able to translate all sentences for the corpus. To easier the task, we prepare a list of the most polysemic verbs annotated by Cabezudo et al. (2015). With the list, we filter out entries that did not contain the expected verbs.

[10] https://github.com/hitvoice/DrQA

[11] https://spacy.io/

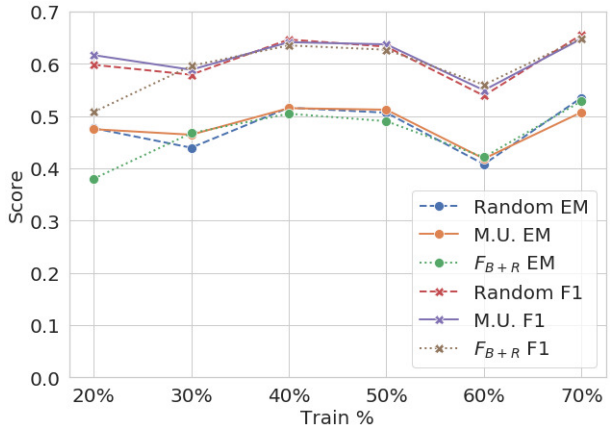


Figure 2: Reading comprehension (es): Exact Match ($EM$) and F1-score ($F1$) on the development set for each partition of the training set

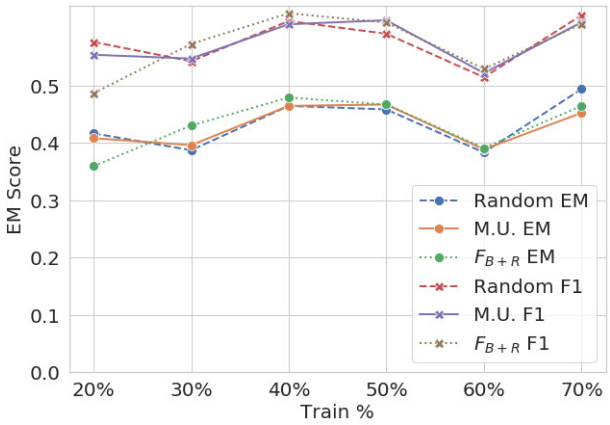


Figure 3: Reading comprehension (es): Exact Match ($EM$) and F1-score ($F1$) on the $3^{rd}$ quartile of the test set for each partition of the training set

results against the development and test sets specified previously. The evaluation metrics for the experiments were Exact Match ($EM$) and F1-score ($F1$). The former one is the percentage of predicted answers that exactly match the original answer, whereas the latter one is the average overlap between the predicted and original answers. The results for both dev and test are shown in Figures 2 and 3, respectively. We use the filtered test partition with the highest quartile.

In both figures, we can observe that there is not a vast difference between any of the metrics and the random selection throughout all the partitions. We expected the previous outcome, as our processed corpus has already been filtered to preserve only the questions with an explicit and exact answer in the reference texts.

We carried out a complementary analysis, where we compared a neural method versus a non-data-driven method. One of the few methods implemented for monolingual reading comprehension in Spanish proposes a straightforward pipeline (Vicedo et al., 2004). They extract keywords from the question, search the web for related passages and identify a potential answer from them. They used the set of 200 questions from the CLEF 2003 Spanish monolingual QA evaluation task (Magnini et al., 2004), which lacks context because of the nature of the challenge. We reproduce the second half of the pipeline, assuming that we already have a related passage to look for the answer.

| | Full Test | | | | |
|---|---|---|---|---|---|
| **Question** | (Vicedo et al., 2004) | | Our model | | **#Q** |
| **Type** | **EM** | **F1** | **EM** | **F1** | |
| Date | 0.2721 | 0.4185 | 0.4545 | 0.5933 | 452 |
| Number | 0.5421 | 0.6377 | 0.4332 | 0.5754 | 404 |
| Other | 0.1376 | 0.1966 | 0.4316 | 0.5841 | 4,119 |
| Not Recognized | | | | | 1,378 |
| Total | 0.1429 | 0.1976 | 0.4347 | 0.5846 | 6,353 |
| **Test Q1** | | | | | |
| Date | 0.2436 | 0.3913 | 0.5513 | 0.7014 | 78 |
| Number | 0.4875 | 0.6038 | 0.6 | 0.6992 | 80 |
| Other | 0.1528 | 0.2059 | 0.4425 | 0.5956 | 687 |
| Not Recognized | | | | | 242 |
| Total | 0.1499 | 0.2026 | 0.4453 | 0.5894 | 1,087 |
| **Test Q2** | | | | | |
| Date | 0.2436 | 0.3913 | 0.5513 | 0.7014 | 78 |
| Number | 0.4875 | 0.6038 | 0.6 | 0.6992 | 80 |
| Other | 0.1528 | 0.2059 | 0.4425 | 0.5956 | 687 |
| Not Recognized | | | | | 242 |
| Total | 0.1499 | 0.2026 | 0.4453 | 0.5894 | 1087 |
| **Test Q3** | | | | | |
| Date | 0.3462 | 0.4559 | 0.5 | 0.6224 | 26 |
| Number | 0.4571 | 0.559 | 0.6857 | 0.7742 | 35 |
| Other | 0.1556 | 0.2299 | 0.4319 | 0.5962 | 257 |
| Not Recognized | | | | | 91 |
| Total | 0.1589 | 0.2212 | 0.4645 | 0.607 | 409 |

Table 5: Reading Comprehension (es): Results from our model trained with the selected threshold versus the method of Vicedo et al. (2004) in all test partitions

For this experiment, we selected the model that achieved the highest F1-score in the development set: the top $40\%$ of the training set arranged by the $M.U.$ score (see Figure 2). Results are shown in Table 5, where we observe a difference between the neural and non-neural model, as the former take advantage of the newly generated corpus.

### 4.2 Word Sense Disambiguation (pt)

After the translation and filtering of the UEF corpus (see §3.3.2 for details about training and de-

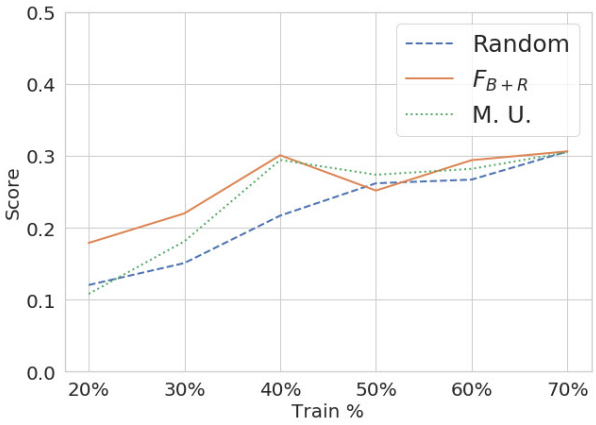


Figure 4: WSD (pt): F1-score on the Development set for each partition of the training set

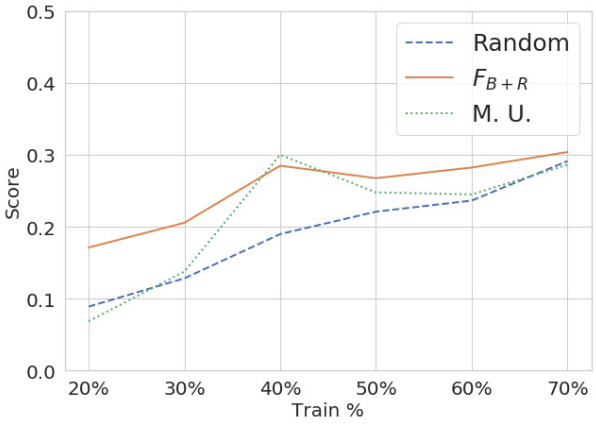


Figure 5: WSD (pt): F1-score on the Test set for each partition of the training set

velopment sets), we proceeded to train the WSD models. Due to the effectiveness of neural networks on several tasks, it was decided to use a Sequence-to-Sequence architecture with an attention mechanism, like the one proposed by Bahdanau et al. (2014). This architecture has been previously used by Raganato et al. (2017). The proposed architecture contains a hidden size of 256 and an embedding size of 300 units. Also, we consider training embeddings from scratch.

Following the training set filtering described in §3.2.1, we trained a different model for each partition of the train data and each criterion. Besides, we used F-score as the validation metric with the formulation of the precision and recall like in Cabezudo and Pardo (2017). The results achieved in the development and test sets are shown in Figure 4 and Figure 5, respectively.

Figure 4 shows that the $F_{B+R}$- and $M.U.$-based filters produce better results (in term of F-score)

| Verb | MFS | Lesk | Our method |
|---|---|---|---|
| *ser* ("to be") | 88.11 | 69.32 | 64.18 |
| *ter* ("to have") | 75.82 | 62.75 | 62.50 |
| *fazer* ("to do") | 31.62 | 11.11 | 21.56 |
| *apresentar* ("to present") | 50.00 | 36.11 | 50.00 |
| *chegar* ("to arrive") | 29.09 | 23.64 | 21.73 |
| *receber* ("to receive") | 61.11 | 42.86 | 36.84 |
| *ficar* ("to stay") | 11.27 | 8.45 | 0.00 |
| *registrar* ("to register") | 3.85 | 7.69 | 0.00 |
| *deixar* ("to leave") | 19.61 | 13.73 | 8.33 |
| *cair* ("to fall") | 17.39 | 17.39 | 20.00 |
| *passar* ("to pass") | 38.30 | 23.40 | 16.67 |
| *fechar* ("to close") | 36.84 | 5.26 | 0.00 |
| *colocar* ("to put") | 63.16 | 31.58 | 62.50 |
| *encontrar* ("to find") | 12.50 | 4.17 | 30.00 |
| *levar* ("to take") | 9.09 | 3.03 | 12.50 |
| *vir* ("to come") | 30.00 | 30.00 | 25.00 |
| *estabelecer* ("to establish") | 8.33 | 16.67 | 25.00 |
| *marcar* ("to mark") | 0.00 | 9.09 | 0.00 |
| *dar* ("to give") | 13.21 | 9.43 | 9.09 |
| *tratar* ("to treat") | 11.11 | 22.22 | 50.00 |
| Precision | 30.52 | 22.39 | 46.44 |

Table 6: Results for the Lexical sample task in WSD

than the Random criterion in the development set. However, the difference between both $F_{B+R}$ and $M.U.$ and random becomes shorter by each part. Given the results, we can observe the best F-score at the 70% partition of the training set.

Nevertheless, due to our primary goal is to build a high-quality corpus, we consider the portion at 40% of the training set better than the one at 70%, as the latter could contain low-quality translations. Moreover, we select the sentences from the 40% partition provided by the $F_{B+R}$ criterion instead of the $M.U.$, because the latter shows less stability in its results.

Besides, Figure 5 is useful to confirm that there is a peak in the validation at the 40% partition of the training data for the two metrics. For that reason, we decided to fix the 40% portion of the data as our definite high-quality corpus.

Finally, we perform a comparison using our selected corpus (at the 40% thresh by $F_{B+R}$) against WSD methods for Brazilian Portuguese in the Lexical Sample task (Cabezudo and Pardo, 2017). Lexical Sample consists of evaluating the 20 most polysemous words in the corpus. Specifically, we compare our results with Most Frequent Sense Heuristic (MFS), which is a strong baseline, and an adaptation of the Lesk algorithm (Lesk, 1986), a knowledge-based method and the best algorithm reported in this work. To analyse the percentage of correctness of the WSD methods on the selected verbs, we only calculate the precision and not the F1-score. Results are presented in Table 6, where

we can see that our method outperforms both MFS Heuristic and Lesk, although not for all the verbs.

## 5 Conclusions and Future Work

We present a study of back-translation and automatic quality evaluation as a corpus generation strategy. Our further goal is to systematise the use of these methods towards a robust and general-purpose corpus generation for new languages. The analysis of several thresholds for corpus filtering and its posterior extrinsic evaluation shows that this strategy is feasible, and it only requires a machine translation system paired with English plus particular processing steps regarding the nature of the target task, but not of the specific language.

We plan to extend the experimentation using less-robust MT systems. Thus, we might assess how far this strategy could work for low-resource languages without commercial MT, as well as to analyse whether the quality-oriented metrics can perform accordingly. There is also potential work in complementing the back-translation strategy with cross-lingual embeddings, supervised or unsupervised, to increase the quality in the corpus generation. Furthermore, an exhaustive exploration could be performed, by including more automatic evaluation metrics as well as additional languages and tasks to draw more general insights.

## Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825299. Besides, we acknowledge the support of the NVIDIA Corporation with the donation of the Titan Xp GPU used for this study. Finally, the first author is granted by the "Programa de apoyo al desarrollo de tesis de licenciatura" (Support programme of undergraduate thesis development, PADET 2018, PUCP).

## References

Eneko Agirre and Philip Edmonds. 2007. *Word Sense Disambiguation: Algorithms and Applications*, 1st edition. Springer Publishing Company.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275v2.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473v7.

Marco A. Sobrevilla Cabezudo, Erick Maziero, Jackson Souza, Márcio Dias, Paula Christina Cardoso, Pedro Paulo Balage Filho, Verônica Agostini, Fernando Antônio Nóbrega, Cláudia de Barros, Ariani Di Felippo, and Thiago Alexandre Pardo. 2015. Anotação de sentidos de verbos em textos jornalísticos do corpus CSTNews. *Revista de Estudos da Linguagem*, 23(3):797–832.

Marco A. Sobrevilla Cabezudo and Thiago A. Salgueiro Pardo. 2017. Exploring classical and linguistically enriched knowledge-based methods for sense disambiguation of verbs in Brazilian Portuguese news texts. *Procesamiento del Lenguaje Natural*, 59(0):83–90.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Cristian Cardellino. 2016. Spanish Billion Words Corpus and Embeddings.

Paula C. F. Cardoso, Erick G. Maziero, Maria L. C. Jorge, Eloize M. R. Seno, Ariani Di Felippo, Lucia H. M. Rino, Maria G. V. Nunes, and Thiago A. S. Pardo. 2011. CSTNews - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, Cuiab MT, Brazil. Sociedade Brasileira de Computao.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. 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.

Marina Fomicheva, Núria Bel, Lucia Specia, Iria da Cunha, and Anton Malinovskiy. 2016. CobaltF: A fluent metric for MT evaluation. In *Proceedings of the First Conference on Machine Translation*, pages 483–490, Berlin, Germany. Association for Computational Linguistics.

Judith Gaspers, Penny Karanasou, and Rajen Chatterjee. 2018. Selecting machine-translated data for quick bootstrapping of a natural language understanding system. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*,

pages 137–144, New Orleans - Louisiana. Association for Computational Linguistics.

Bassam Jabaian, Laurent Besacier, and Fabrice Lefvre. 2011. Combination of stochastic understanding and machine translation systems for language portability of dialogue systems. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5612–5615.

Roman Klinger and Philipp Cimiano. 2015. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 153–163, Beijing, China. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Peñas, Víctor Peinado, Felisa Verdejo, and Maarten de Rijke. 2004. The multiple language question answering track at CLEF 2003. In *Comparative Evaluation of Multilingual Information Access Systems*, pages 471–486, Berlin, Heidelberg. Springer Berlin Heidelberg.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Teruhisa Misu, Etsuo Mizukami, Hideki Kashioka, Satoshi Nakamura, and Haizhou Li. 2012. A bootstrapping approach for SLU portability to a new language by inducting unannotated user queries. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4961–4964.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, and Roser Morante. 2013. QA4MRE 2011-2013: Overview of question answering for machine reading evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 303–320, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Gary F. Simons and Charles D. Fenning, editors. 2019. *Ethnologue: Languages of the World. Twenty-second edition*. Dallas Texas: SIL international. Online version: `http://www.ethnologue.com`.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

José L. Vicedo, Ruben Izquierdo, Fernando Llopis, and Rafael Muñoz. 2004. Question answering in Spanish. In *Comparative Evaluation of Multilingual Information Access Systems*, pages 541–548, Berlin, Heidelberg. Springer Berlin Heidelberg.