

A Machine Translation Approach for Modernizing Historical Documents Using Backtranslation

Miguel Domingo, Francisco Casacuberta

PRHLT Research Center
Universitat Politècnica de València
midobal@prhlt.upv.es, fcn@prhlt.upv.es

Abstract

Human language evolves with the passage of time. This makes historical documents to be hard to comprehend by contemporary people and, thus, limits their accessibility to scholars specialized in the time period in which a certain document was written. Modernization aims at breaking this language barrier and increase the accessibility of historical documents to a broader audience. To do so, it generates a new version of a historical document, written in the modern version of the document's original language. In this work, we propose several machine translation approaches for modernizing historical documents. We tested these approaches in different scenarios, obtaining very encouraging results.

1. Introduction

Historical documents are an important part of our cultural heritage. With the aim of their preservation, there is an increase need of digitalizing—creating a digital text version which can be searched and automatically processed—of historical documents [1]. However, there is an additional difficulty created by their linguistic properties: On the one hand, human language evolves with the passage of time. On the other hand, due to a lack of a spelling convention, orthography changes depending on the author and time period in which a given document was written. These problems make historical documents harder to read, and makes it harder to digitalize them (since their digital text version needs to be searched and automatically processed).

The orthography problem has been well researched in the literature [2, 3, 4, 5, 6, 7]. The proposed solution that aims to solve this problem is known as spelling normalization, and its goal is to adapt the document's spelling to modern standards in order to achieve an orthography consistency and increase the document's read-

ability. However, while is true that spelling normalization makes historical documents easier to read, they are still hard to comprehend by contemporary people. This problem limits the accessibility of historical documents to scholars specialized in the time period in which the document was written.

Modernization aims at breaking the language barrier, generating a new version of a historical document in the modern version of the language in which the document was originally written (see Fig. 1 for an example). Therefore, not only the orthography is updated. The lexicon and grammar are also modified in order to match the modern use of the document's language. The main drawback of this solution is that part of the document's original intention could be lost in the process (e.g., part of the rhyme in Fig. 1 is lost for the sake of clarity). Nonetheless, the document's clarity is increased and, thus, its accessibility to a broader audience.

To the best of our knowledge, modernization of historical documents has been less researched in the literature. A shared task was organized in order to translate historical text to contemporary language [9]. The shared task's main goal was spelling normalization. However, they also tackle modernization using a set of rules. Finally, there was an approach to modernize historical documents using Statistical Machine Translation (SMT) [10]. In this work, we tackle modernization by using an SMT and Neural Machine Translation (NMT) approach. Additionally, since a frequent problem when working with historical documents is the scarce availability of parallel training data [5], we created two small parallel corpora (see Section 3.1) and generated synthetic data using backtranslation [11]. Our main contributions are the followings:

- First use, to the best of our knowledge, of NMT and backtranslation for historical documents mod-

Shall I compare thee to a summer's day?
 Thou art more lovely and more temperate:
 Rough winds do shake the darling buds of May,
 And summer's lease hath all too short a date:
 Sometime too hot the eye of heaven shines,
 And often is his gold complexion dimm'd;
 And every fair from fair sometime declines,
 By chance or nature's changing course untrimm'd;
 But thy eternal summer shall not fade
 Nor lose possession of that fair thou ow'st;
 Nor shall Death brag thou wander'st in his shade,
 When in eternal lines to time thou grow'st;
 So long as men can breathe or eyes can see,
 So long lives this, and this gives life to thee.

Shall I compare you to a summer day?
 You're lovelier and milder.
 Rough winds shake the pretty buds of May,
 and summer doesn't last nearly long enough.
 Sometimes the sun shines too hot,
 and often its golden face is darkened by clouds.
 And everything beautiful stops being beautiful,
 either by accident or simply in the course of nature.
 But your eternal summer will never fade,
 nor will you lose possession of your beauty,
 nor shall death brag that you are wandering in the underworld,
 once you're captured in my eternal verses.
 As long as men are alive and have eyes with which to see,
 this poem will live and keep you alive.

Figure 1: Example of modernizing a historical document. The original text is *Shakespeare Sonnet 18*. The modernized version of the Sonnet was obtained from [8].

ernization.

- Comparison of approaches based on SMT and NMT.
- Experimented with three historical corpora—two of which were created for this work—from three different time periods and two different languages.

The rest of this document is structured as follows: Section 2 introduces the different Machine Translation (MT) approaches used in our work. Then, Section 3 describes the experiments conducted in order to assess our proposal. After that, Section 4 presents and discusses the results of those experiments. Finally, in Section 5, conclusions are drawn.

2. Machine Translation

In this section, we present the machine translation approaches used in our work.

2.1. Statistical Machine Translation

Given a source sentence \mathbf{x} , SMT aims at finding its best translation $\hat{\mathbf{y}}$ [12]:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} Pr(\mathbf{y} | \mathbf{x}) \quad (1)$$

For years, the prevailing approach to compute this expression have been phrase-based models [13]. These models rely on a log-linear combination of different models [14]: namely, phrase-based alignment models, reordering models and language models; among others [15, 16]. However, more recently, this approach has shifted into neural models (see Section 2.2).

2.2. Neural Machine Translation

NMT is the neural approach to compute Eq. (1). Frequently, it relies on a Recurrent Neural Network (RNN) encoder-decoder framework. In this framework, the source sentence is projected into a distributed representation at the encoding step. At the decoding step, the decoder generates its translation word by word [17].

The system's input is a sequence of words in the source language. Each source word is linearly projected to a fixed-sized real-valued vector through an embedding matrix. These word embeddings are fed into a bidirectional [18] Long Short-Term Memory (LSTM) [19] network, resulting in a sequence of annotations produced by concatenating the hidden states from the forward and backward layers.

The model features an attention mechanism [20], which allows the decoder to focus on parts of the input sequence, computing a weighted mean of annotations sequence. A soft alignment model computes these weights by weighting each annotation with the previous decoding state.

The decoder is another LSTM network, conditioned by the representation computed by the attention model and the last word generated. Finally, a deep output layer [21] computes a distribution over the target language vocabulary.

The model is trained by means of stochastic gradient descent, applied jointly to maximize the log-likelihood over a bilingual parallel corpus. At decoding time, the model approximates the most likely target sentence with beam-search [17].

2.3. Backtranslation

Backtranslation [11] has become the norm when building state-of-the-art NMT systems, especially in resource-poor scenarios [22]. It is a useful technique to increase the training data by creating synthetic text from monolingual data. Given a monolingual corpus in the target language, and an MT system trained to translate from the target language to the source language, the synthetic data is generated by translating the monolingual corpus with the MT system. After that, the synthetic data is used as the source part of the corpus, and the monolingual data as the target part. Finally, this new corpus is mixed with the available training data in order to train a new MT system.

In this work, to generate the synthetic data, we translate the monolingual data using an ad-hoc SMT system trained with the corpus' training partition. Additionally, since the datasets are considerable small, prior to mixing the synthetic corpus with the training partition, we replicate several times the training data in order to match the size of the synthetic data and avoid overfitting [23]. Finally, we trained an NMT system with this new corpus.

3. Experimental Framework

In this section, we present the corpora and metrics, and describe the MT systems used during the experimental session.

3.1. Corpora

The first corpus used to assess our proposal was the **Dutch Bible** [9]. This corpus consists in a collection of different versions of the Dutch Bible. More precisely, a version from 1637, another from 1657, another from 1888 and another from 2010. All versions contain the same text except for the 2010 version, which is missing the last books. Moreover, the authors mentioned that the translation from this last version is not very reliable. Additionally, due to Dutch not evolving significantly during this period, 1637 and 1657 versions are fairly similar. For this reason, we decided to only use the 1637 version—considering this as the original document—and the 1888 version—considering 19th century Dutch as *modern Dutch*.

To create the synthetic corpus (see Section 2.3), we collected all 19th century Dutch books available at the

*Digitale Bibliotheek voor de Nederlandse letteren*¹ and used them as monolingual data.

The second corpus we used was **El Quijote**. We built this corpus using a version [24] of the original 17th century Spanish novel by Miguel de Cervantes, and a 21st century version modernized by Andrés Trapiello [25]. The first step was to split each document into sentences. Since the 17th century version was faithful to the original manuscript (in which each document line is formed by a very few words), we replaced line breaks by spaces to create a single sentence, and removed empty lines. For consistency, we did the same to the 21st century version. After that, we split each document into sentences by adding line breaks to relevant punctuation (i.e., dots, quotation marks, admiration marks, etc). Then, to ensure consistency, we checked special symbols (e.g., quotation marks) and made sure that the same character was used in both versions. Finally, in order to create a parallel corpus, we aligned both documents using *Hunalign* [26].

Since the content of this corpus was a novel, we decided that, to create the synthetic corpus, it would be best to use monolingual data coming from Spanish literature. For this reason and, considering that Spanish hasn't changed significantly over the last decades, we decided to collect free-of-right late 20th century Spanish novels from *Project Gutenberg*².

Finally, as a third corpus, we selected **El Conde Lucanor**. We built this corpus using a version of the original 14th century Spanish novel by Don Juan Manuel, and a 21st century version modernized by Luis López Nieves [27]. To create the parallel version, we followed the same steps than with *El Quijote*. However, unlike with *El Quijote*, the resulting corpus was too small to be able to use for training an MT system. Therefore, we decided to use it only as a test. Unable to find a suitable training corpus, we decided to test *El Conde Lucanor* using the systems created for *El Quijote*—despite the fact that the original documents were written three centuries apart from one another.

Table 1 presents the corpora statistics.

3.2. Metrics

In order to assess our proposal, we made use of the following well-known metrics:

- **BiLingual Evaluation Understudy (BLEU)** [28]: computes the geometric average of the modified

¹<http://dbnl.nl/>

²<https://www.gutenberg.org/>

		Dutch Bible	El Quijote	El Conde Lucanor
Train	S	35.2K	10K	-
	T	870.4/862.4K	283.3/283.2K	-
	V	53.8/42.8K	31.7/31.3K	-
Development	S	2000	2000	-
	T	56.4/54.8K	53.2/53.2K	-
	V	9.1/7.8K	10.7/10.6K	-
Test	S	5000	2000	2252
	T	145.8/140.8K	41.8/42.0K	62.0/56.7K
	V	10.5/9.0K	8.9/9.0K	7.4/8.6K
Monolingual	S	4.1M	567.0K	-
	T	88.3M	9.5M	-
	V	2.0M	470.4K	-

Table 1: Corpora statistics. $|S|$ stands for number of sentences, $|T|$ for number of tokens and $|V|$ for size of the vocabulary. *Monolingual* refers to the monolingual data used to create the synthetic data. M denotes million and K thousand.

n-gram precision, multiplied by a brevity factor that penalizes short sentences.

- **Translation Error Rate (TER)** [29]: computes the number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation.

Confidence intervals ($p = 0.05$) were computed for all metrics by means of bootstrap resampling [30].

3.3. MT Systems

We trained the SMT systems with *Moses* [31], following the standard procedure: we estimated a 5-gram language model—smoothed with the improved KneserNey method—using *SRILM* [32], and optimized optimized the weights of the log-linear model with *MERT* [33]. Additionally, we lowercased and tokenized the corpora using the standard scripts and, later, truecased the translated text using *Moses*’ truecaser.

To train the NMT systems, we used *OpenNMT* [34]. We used LSTM units, following the findings from [35]. We set the size of the word embedding and LSTM units to 1024. We used *Adam* [36] with a learning rate of 0.0002 [37]. The beam size was set to 6. Finally, the corpora were lowercased and tokenized—and, later, truecased and detokenized—using *OpenNMT*’s tools.

In order to reduce the vocabulary, we applied Byte Pair Encoding (BPE) [38] to both SMT and NMT systems. We trained the models with a joint vocabulary of 32000 BPE units.

4. Results

In this section, we present and discuss the results of the experiments conducted in order to assess our proposal. Table 2 presents the experimental results.

Dutch Bible contained an additional baseline which consisted in generating a modernized version of the text by applying a set of rules to the original document [9]. This second baseline improved significantly (close to 40 BLEU points and 30 TER points) the standard baseline of considering the original document as the modernized version. However, the SMT approach improved those results even more (near 30 BLEU points and 15 TER points of improvement with respect to the second baseline, and 70 BLEU points and 50 TER points with respect to the standard baseline). The NMT approach yielded better results than the standard baseline (and improvement of around 25 BLEU points and 5 TER points), but worse results than the second baseline and the SMT approach. Most likely, this is due to the training corpus being too small, which is a well-known problem in NMT. Finally, the backtranslation approach yielded the worst results. These results represent an improvement over the standard baseline in term of BLEU (around 4 points), and a deterioration in terms of TER (around 8 points). Most likely, this is due to the monolingual data used for backtranslation not being similar enough to the training data.

The experiments using *El Quijote* behaved similarly—taking into account that the only available baseline is the standard one—to *Dutch Bible*: The SMT approached yielded the best results (an improvement of

System	Dutch Bible		El Quijote		El Conde Lucanor	
	BLEU	TER	BLEU	TER	BLEU	TER
Baseline	13.5 ± 0.3	57.0 ± 0.3	36.5 ± 0.8	43.3 ± 1.1	5.8 ± 0.3	89.6 ± 1.0
Baseline ₂	50.8 ± 0.4	26.5 ± 0.3	-	-	-	-
SMT	80.1 ± 0.5	9.9 ± 0.3	58.9 ± 1.0	29.4 ± 1.2	8.4 ± 0.3	83.8 ± 1.0
NMT	38.0 ± 0.6	51.7 ± 2.2	37.4 ± 1.2	51.5 ± 2.0	2.7 ± 0.2	99.5 ± 2.0
NMT _{Synthetic}	17.4 ± 0.5	65.6 ± 1.7	45.2 ± 1.3	50.6 ± 3.5	3.1 ± 0.2	165.1 ± 8.2

Table 2: Experimental results. *Baseline* system corresponds to considering the original document as the modernized version. *Baseline₂* came with the *Dutch Bible* and is a modernized version of the text generated by applying a set of rules to the original document [9]. *SMT* and *NMT* are the SMT and NMT approaches respectively. NMT_{Synthetic} is the NMT system trained with the synthetic data generated through backtranslation. Best results are denoted in **bold**.

close to 22 BLEU points and 14 TER points). The results yielded by the NMT approach were not significantly different to the baseline in terms of BLEU, and represented close to a 10 points deterioration in terms of TER. In this case, however, the backtranslation approach yielded nearly a 10 points improvement in terms of BLEU, and the same TER results as the NMT approach.

Not being able to obtain enough suitable training data for *El Conde Lucanor*, we used the same systems than for *El Quijote*. However, these documents were written three centuries apart from one another (*El Conde Lucanor* is written in 14th century Spanish and *El Quijote* in 17th century Spanish). Therefore, the obtained results contained a low translation quality. Nonetheless, it is worth noting that the SMT approach yielded improvements over the baseline (around 3 BLEU points and 6 TER points). However, the NMT and backtranslating approached yielded a deterioration of 3 BLEU points (in both cases) and 10 and 75 TER points respectively.

In general, SMT yielded the best results in all cases. NMT was able to improve *Dutch Bible*’s baseline, yielding similar results to *El Quijote*’s baseline and worse results than *El Conde Lucanor*’s baseline. Finally, despite being successfully used in resources-poor scenarios, backtranslation was only able to improve results for the experiment using *El Quijote*, and these results were worse than the ones yielded by the SMT approach.

4.1. Qualitative Analysis

Table 2 shows some examples of sentences modernized using the different MT approaches.

The first example is a sentence from *El Quijote*. The hypothesis generated by the SMT approach is very closed to the reference. The main differences are a change in the order of actions (the original sentence

says *Y dejando de comer, se levantó*, which is changed by the hypothesis into *Y levantándose, dejó de comer*) and some changes in the conjugation of verbs (e.g., *dejando* is changed into *dejó*). However, the main goal of modernization is not to generate a perfect equivalent version, but to make the document easier to comprehend—making the overall meaning more important than the exact choice of words. While sentences like these are penalized by the automatic metrics, they accomplish modernization’s goal.

The hypothesis generated by the NMT approach follows the same structure than the SMT hypothesis (it makes the same reordering and conjugation changes) but contains non-existent words (e.g., *ancen*) and has some errors (e.g, *a los pies* in stead of *puesto a caballo*). Therefore, some parts are easier to comprehend than in the original version, but the meaning of the sentence is not clear.

Finally, the hypothesis generated by the backpropagation approach is almost the same as the one generated by the SMT approach (the only change is *ferded* in stead of *fiereza*). While this hypothesis is less correct than the SMT one, they are both equally easy to comprehend.

The second example is from *El Conde Lucanor*, whose experiments were conducted using the systems trained with *El Quijote*. As a result, all the hypothesis are hard to comprehend. While the automatic metrics heavily penalize the backtranslation approach, in this case, is the one which is closer to modern Spanish. Moreover, it is the only hypothesis which preserves the name of the main characters (*Lucanor* and *Patronio*). Looking through the whole texts, the SMT and NMT hypothesis frequently changed the characters named into non-existent words, while the backtranslation approach rarely modified them. Finally, having trained the systems with *El Quijote* has a visible effect in the SMT and

El Quijote

Original:	Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lança atrauessada por la boca, con la fiereça que suele pintarse.
Modernized:	Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.
SMT:	Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.
NMT:	Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera ancen, que mostró ser la de San Marorge a los pies y la lanza ahabesada por la boca;
NMT_{Synthetic}:	Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fierded que suele pintarse.

El Conde Lucanor

Original:	-Señor conde Lucanor -dixo Patronio-, vien entiendo que el mí consejo non vos faze grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos consejo sobre ello, fazerlo he luego.
Modernized:	-Señor Conde Lucanor -dijo Patronio-, bien sé que mi consejo no os hace mucha falta, pero, como confiáis en mí,
SMT:	-- Señor conde Lucanor -dijo Patroniorosa, vien entiendo que el mí consejo non vos face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos aconseje en ello, ferlo he luego .
NMT:	Señor conde Olcanor dijo dijo Pacasos -dijo en entiendo que el mí consejo non os fazo felimengua y vuestra merced es que vos diga lo que en esto entiendo.
NMT_{Synthetic}:	-Señor conde Lucanor -dijo Patronio, vien entiendo que el mí consejo non es face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos consejo sobre ello, también yo he dicho.

Table 3: Examples of modernizing a sentence using the different MT approaches. *SMT* and *NMT* are the SMT and NMT approaches respectively. *NMT_{Synthetic}* is the NMT system trained with the synthetic data generated through backtranslation.

NMT approaches (with verbs conjugations such as *ferlo*, or expression such as *vuestra merced*). This effect, however, is not so visible in the backtranslation hypothesis. All in all, neither hypothesis accomplishes the goal of improving the comprehension of the original sentence. Nonetheless, it is worth noting that this was a tricky example since the original sentence is modernized into a much shorter sentence (which reflects that 14th century Spanish used longer expressions than modern Spanish).

5. Conclusions and Future Work

In this work we proposed several machine translation approaches to modernize historical documents in order to break the language barrier and increase their accessibility to a broader audience.

We tested our approaches using three historical datasets (two of which were created for this work) from three different time periods and two different languages.

Our first approach was based in SMT and yielded, for all cases, the best results. With the exception of the dataset for which there were not available any suitable training data, this approach yielded significant improvements of around 22 to 67 BLEU points and 14 to 48 TER points.

Since the available training data was fairly small, the approach based on NMT produced less satisfactory results. While it was able to yield improvements for one dataset, the rest of the experiment resulted in either not significantly different than the baseline, or yielding a deterioration in terms of translation quality.

Finally, despite being successfully used in resource-poor scenarios, backtranslation was only able to improve results for one dataset, and only in terms of BLEU. Our best hypothesis is that historical documents are very language-specific and, therefore, choosing the monolingual corpus to use for creating the synthetic data is extremely important. While we tried to create the monolingual datasets using similar topics, the corpora's topics were too specific: religious texts, a cavalry novel and medieval tales.

In a future work, we would like to research the relation between the domains of the monolingual and training corpora deeper. Additionally, we want to explore the use of data selection techniques for constructing the monolingual corpus to use for backtranslation, and to create a training partition for cases in which we do not have suitable training data available (as was the case with *El Conde Lucanor*).

6. Acknowledgments

The research leading to these results has received funding from the Ministerio de Economía y Competitividad (MINECO) under project CoMUN-HaT (grant agreement TIN2015-70924-C2-1-R). We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for part of this research, and Andrés Trapiello and Ediciones Destino for granting us permission to use their book in our research.

7. References

- [1] M. Piotrowski, *Natural Language Processing for Historical Texts*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2012, no. 17.
- [2] A. Baron and P. Rayson, “VARD2: A tool for dealing with spelling variation in historical corpora,” *Postgraduate conference in corpus linguistics*, 2008.
- [3] J. Porta, J.-L. Sancho, and J. Gómez, “Edit transducers for spelling variation in old spanish,” in *Proceedings of the workshop on computational historical linguistics*, 2013, pp. 70–79.
- [4] Y. Scherrer and T. Erjavec, “Modernizing historical slovene words with character-based smt,” in *Proceedings of the Biennial International Workshop on Balto-Slavic Natural Language Processing*, 2013, pp. 58–62.
- [5] M. Bollmann and A. Søgaard, “Improving historical spelling normalization with bi-directional lstms and multi-task learning,” in *Proceedings of the International Conference on the Computational Linguistics*, 2016, pp. 131–139.
- [6] N. Ljubešić, K. Zupan, D. Fišer, and T. Erjavec, “Normalising slovene data: historical texts vs. user-generated content,” in *Proceedings of the Conference on Natural Language Processing*, 2016, pp. 146–155.
- [7] M. Domingo and F. Casacuberta, “Spelling normalization of historical documents by using a machine translation approach,” in *Proceedings of the Annual Conference of the European Association for Machine Translation*, 2018, pp. 129–137.
- [8] J. Crowther, *No Fear Shakespeare: Sonnets*. SparkNotes, 2004.
- [9] E. Tjong Kim Sang, M. Bollmann, R. Boschker, F. Casacuberta, F. Dietz, S. Dipper, M. Domingo, R. van der Goot, M. van Koppen, N. Ljubešić, R. Östling, F. Petran, E. Pettersson, Y. Scherrer, M. Schraagen, L. Sevens, J. Tiedemann, T. Vanallemeersch, and K. Zervanou, “The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation,” *Computational Linguistics in the Netherlands Journal*, vol. 7, pp. 53–64, 2017.
- [10] M. Domingo, M. Chinea-Rios, and F. Casacuberta, “Historical documents modernization,” *The Prague Bulletin of Mathematical Linguistics*, vol. 108, pp. 295–306, 2017.
- [11] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” *arXiv preprint arXiv:1511.06709*, 2015.
- [12] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
- [13] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2010.
- [14] F. J. Och and H. Ney, “Discriminative training and maximum entropy models for statistical machine translation,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 295–302.
- [15] R. Zens, F. J. Och, and H. Ney, “Phrase-based statistical machine translation,” in *Proceedings of the Annual German Conference on Advances in Artificial Intelligence*, vol. 2479, 2002, pp. 18–32.
- [16] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, pp. 48–54.
- [17] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” pp. 3104–3112, 2014.

- [18] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [19] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [20] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proceedings of the International Conference on Learning Representations (arXiv:1409.0473)*, 2015.
- [21] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, “How to construct deep recurrent neural networks,” *arXiv preprint arXiv:1312.6026*, 2013.
- [22] A. Poncelas, D. Shterionov, A. Way, G. Maillette de Buy Wenniger, and P. Passban, “Investigation backtranslation in neural machine translation,” in *Proceedings of the Annual Conference of the European Association for Machine Translation*, 2018, pp. 249–258.
- [23] R. Chatterjee, M. A. Farajian, M. Negri, M. Turchi, A. Srivastava, and S. Pal, “Multi-source neural automatic post-editing: Fbk’s participation in the wmt 2017 ape shared task,” in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 630–638.
- [24] F. F. Jehle, *Works of Miguel de Cervantes in Old- and Modern-spelling*. Indiana University Purdue University Fort Wayne, 2001.
- [25] A. Trapiello, *Don Quijote de la Mancha Puesto en castellano actual íntegra y fielmente por Andrés Trapiello*. Ediciones Destino, 2015.
- [26] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy, “Parallel corpora for medium density languages,” in *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 2005, pp. 590–596.
- [27] L. López Nieves, *El Conde Lucanor*. Biblioteca Digital Ciudad Seva, 2002.
- [28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [29] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A study of translation edit rate with targeted human annotation,” in *Proceedings of the Association for Machine Translation in the Americas*, 2006, pp. 223–231.
- [30] P. Koehn, “Statistical significance tests for machine translation evaluation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 388–395.
- [31] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for statistical machine translation,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2007, pp. 177–180.
- [32] A. Stolcke, “SRILM - an extensible language modeling toolkit,” in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 257–286.
- [33] F. J. Och, “Minimum error rate training in statistical machine translation,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2003, pp. 160–167.
- [34] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, “OpenNMT: Open-Source Toolkit for Neural Machine Translation,” *arXiv preprint arXiv:1701.02810*, 2017.
- [35] D. Britz, A. Goldie, T. Luong, and Q. Le, “Massive exploration of neural machine translation architectures,” *arXiv preprint arXiv:1703.03906*, 2017.
- [36] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [37] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *arXiv preprint arXiv:1609.08144*, 2016.

- [38] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1715–1725.