Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3623–3629
Marseille, 11–16 May 2020

NMT and PBSMT Error Analyses in English to Brazilian Portuguese Automatic Translations

Helena de Medeiros Caseli, Marcio Lima Inácio
Federal University of São Carlos, University of São Paulo
Via Washington Luís, km 235, CP 676, CEP 13565-905, São Carlos, SP, Brazil
helenacaseli@ufscar.br, marciolimainacio@usp.br

Abstract

Machine Translation (MT) is one of the most important natural language processing applications. Independently of the applied MT approach, a MT system automatically generates an equivalent version (in some target language) of an input sentence (in some source language). Recently, a new MT approach has been proposed: neural machine translation (NMT). NMT systems have already outperformed traditional phrase-based statistical machine translation (PBSMT) systems for some pairs of languages. However, any MT approach outputs errors. In this work we present a comparative study of MT errors generated by a NMT system and a PBSMT system trained on the same English – Brazilian Portuguese parallel corpus. This is the first study of this kind involving NMT for Brazilian Portuguese. Furthermore, the analyses and conclusions presented here point out the specific problems of NMT outputs in relation to PBSMT ones and also give lots of insights into how to implement automatic post-editing for a NMT system. Finally, the corpora annotated with MT errors generated by both PBSMT and NMT systems are also available.

Keywords: Machine Translation, Corpus, Evaluation Methodologies

1. Introduction

Machine Translation (MT) is one of the most important Natural Language Processing (NLP) applications. In the last 70 years, many different approaches have been proposed for MT such as the rule-based MT (Senellart and Senellart, 2005; Armentano-Oller et al., 2006), the statistical MT (Brown et al., 1993; Och and Ney, 2004; Koehn et al., 2007) and, recently, the neural MT (Bahdanau et al., 2015; Klein et al., 2017). Regardless of the MT approach applied, a MT system automatically generates an equivalent version (in some target language) of an input sentence (in some source language).

However, despite the huge effort of the MT community, it is not possible yet to generate a perfect completely automatic translation for unrestricted domains. Regarding the Brazilian Portuguese, this statement has been confirmed by some previous works such as (Caseli, 2007) and (Martins and Caseli, 2015). In (Caseli, 2007), the best MT systems for Brazilian Portuguese in that time[1] translated incorrectly more than 50% of the input sentences. Lately, in (Martins and Caseli, 2015), the same occurred for a Phrase-based Machine Translation System (PBSMT), for which 67% of the translated sentences in Brazilian Portuguese had one or more translation errors.

More recently, some analyses have also been done for Neural Machine Translation (NMT) models, such as (Bentivogli et al., 2016; Klubička et al., 2017; Popović, 2018). However, to the best of our knowledge, no comparative analysis of MT errors output by PBSMT and NMT has been done for the English and Brazilian Portuguese language pair.

Thus, in this article we present the first error analysis of a NMT system's output for Brazilian Portuguese. We do that by comparing the errors produced by a NMT system with the errors output of a PBSMT system trained with the same corpus.

As a paper intended to present the evaluation of language resources, our goal is to point out some important insights about the way the different MT approaches behave. So, we place more emphasis on the qualitative analysis of the MT errors than on the technical details about the MT systems and their training process.

This paper is organized as follows. Section 2. presents some related work on MT error analysis. Section 3. briefly describes the machine translation systems investigated in this work together with the values of the automatic evaluation measures calculated based on their outputs. Afterwards, section 4. comprehends the error analyses for both MT systems. Lastly, we reason about some conclusions and future work in section 5.

2. Related Work

Although there are plenty of automatic measures for Machine Translation Evaluation (Papineni et al., 2002; Popović, 2015; Denkowski and Lavie, 2014), the error analysis performed by a human is fundamental to understand the limitations and strengths of different MT approaches.

This sort of analysis was performed more recently by Bentivogli et al. (2016) for German translations from English. They compared PBSMT outputs with NMT ones and stated that neural systems produce better translations. Nonetheless, the authors of that paper also concluded that the quality of NMT's output degrades faster as the input sentence length grows. Another important consideration is that NMT systems deal better with verb placement, although reordering still has room for improvement.

Later, Toral and Sánchez-Cartagena (2017) also did some research for different language directions between English, Czech, German, Finnish and Romanian. The authors provide multiple analyses of NMT outputs compared to PB-

[1] These models were, namely: Systran, FreeTranslation and TranslatorPro.

SMT ones, considering different characteristics, such as fluency and reordering. As a result, Toral and Sánchez-Cartagena (2017) state that translations performed by NMT systems tend to be more fluent, however they also observed that quality degrades faster with the sentence length.
Similarly, Popović (2018) analysed translations for English–German and English–Serbian. Popović (2018) presents a manually annotated analysis of linguistically motivated issues in both PBSMT and NMT systems' outputs. In their conclusions, the NMT model showed better fluency, regarding word order and morphology aspects of the languages, than the PBSMT one. Nonetheless, the NMT system produces errors related to prepositions and ambiguous words. Popović (2018) also focuses on the possibility of combination of both PBSMT and NMT models, as the issues of both approaches are mostly complementary.
Another work which executed some analyses of MT outputs is (Martins and Caseli, 2015), in which both manual and automatic error identification were performed. Martins and Caseli (2015) focused on PBSMT system outputs for the same pair of languages under investigation in this work: English and Brazilian Portuguese. In their work, Martins and Caseli (2015) concluded that the most frequent errors in the PBSMT system output were, in this order: lexical errors (44.48%), inflectional errors (38.61%), reordering errors (8.99%) and errors involving n-grams (7.92%).
This work is an extension of (Martins and Caseli, 2015), now adding the MT error analysis of a NMT system and comparing the kind of errors performed by both approaches.

3. Baseline Machine Translation Systems

In all the history of MT systems we can point out three main approaches: Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT) and, more recently, Neural Machine Translation (NMT). This work focuses on the analysis of errors output by the last two approaches which are/was considered the state-of-the-art in the last 30 years.

3.1. The Phrase-based Statistical Machine Translation System

Statistical Machine Translation systems rely on probabilities to define the best translation for a given source sentence. In its most used strategy, the Phrase-Based SMT (PBSMT), these probabilities are computed through sequences of tokens called **phrases** (Och and Ney, 2004).
The goal of a SMT system is to find the most probable sentence $\hat{\mathbf{e}}$ in the target language given the sentence $\mathbf{f}$ in the source language according to the probability distribution $p(\mathbf{e}|\mathbf{f})$ through all possible sentences $\mathbf{e}$ in the target language.
This probability distribution can be modeled in many ways, being Log-Linear models the state of the art for several years (Koehn, 2009). The estimation for the translation probability is calculated from three different base distributions: a translation model, a distortion model and a language model.
The translation model focuses on finding which phrases are the most suitable translation for a given phrase. The distortion model aims at reordering the phrases by calculating the probability that a phrase has to be moved. Finally, the language model guarantees the fluency of the sentence being produced in the target language. Subsequently all models are weighted and combined together in the Log-Linear model to perform the translation.
This approach has the advantage of being able to be applied to, possibly, all language pairs and corpora types. However, this method cannot map structural aspects of the languages and the depth of linguistic knowledge it considers is limited.
In this work, we chose the PBSMT system presented in (Martins and Caseli, 2015) as one of our baseline MT systems. The PBSMT was trained using Moses[2] toolkit as detailed in (Martins and Caseli, 2015).

3.2. The Neural Machine Translation System

The NMT model used to generate the translations analysed in this work follows the standard **attention-based** architecture (Bahdanau et al., 2015). This model is based on the **sequence-to-sequence Recurrent Neural Network Model** (Sutskever et al., 2014), which consists of two Recurrent Neural Network cells, often a Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) or a Gated Recurrent Unit (GRU) (Cho et al., 2014), which are named as encoder and decoder units of the translation system.
The encoder unit is in charge of generating an intermediate vectorial representation for the input sequence of tokens in the source language (English in this work), the input sentence. Afterwards, the Decoder turns this internal representation into a sequence of tokens in the target language (Brazilian Portuguese in this work) equivalent to the translation of the input sentence.
Additionally, the model used for this work implements an attention layer between the encoder and decoder units (Bahdanau et al., 2015). This layer is responsible for weighing the representations generated by the Encoder at each time step t.
The encoder unit generates a representation for each time step t, when it receives the token w_t from the source sentence $\{w_1, \cdots w_t \cdots w_T, (\text{EOS})\}$. This means that the final representation – when it receives the "end-of-sentence" (EOS) token – the intermediate vector can represent the meaning of the whole sentence, but it is more likely that it "forgets" the initial tokens as the length of the sentence grows.
To avoid the problem of forgetting the meaning of the first tokens, the attention model can apply weights into each representation at each time step of the Encoder. This selects which part of the input sentence is most important for the Decoder to generate the current token, i.e. the decoder unit "pays attention" to specific parts of the input sentence to generate each word in the target language.
The NMT system was trained using the Open-NMT tool[3] (Klein et al., 2017) implemented in Python, as it is widely known and optimized.

[2] http://www.statmt.org/moses/
[3] http://opennmt.net/

3.3. MT Baseline Systems' Automatic Evaluation

Both MT baseline systems were trained with the Brazilian Portuguese – English train split of the FAPESP parallel corpora[4] (Aziz and Specia, 2011) and tested with the FAPESP's test-a and test-b files. The automatic evaluation was performed based on BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) measures and the obtained values are shown in Table 1.

Test	MT system	BLEU	ChrF
Test A	PBSMT	60.00	88.51
	NMT	48.75	66.34
Test B	PBSMT	49.36	74.93
	NMT	39.25	67.18

Table 1: BLEU and ChrF values for both PBSMT and NMT systems

From the values in Table 1, it is possible to see that the PBSMT outperforms the NMT model regarding the values of both automatic measures. We believe that the worst performance of the NMT system was mainly due to the small size of the training corpus, which contains only about 162k sentences. As a neural model needs a huge amount of data to be trained properly, this is the most probable reason for its lower values in these automatic measures.

4. Error Analysis

In this section we describe the methodology adopted for the error analysis as well as the main quantitative and qualitative results.

4.1. Methodology

The first step for the error analysis was to identify the errors in the automatically translated sentences and to associate them to the correspondent error category. To do so, we adopted the same error taxonomy of (Martins and Caseli, 2015). Briefly, the errors were divided into four main categories each one possibly including subcategories. The main categories were:

- **Syntactic errors** are those involving just one word where the occurrence is related to other neighbouring words. Subcategories of syntactic errors are: wrong number agreement (singular, plural), wrong gender agreement (female, male), wrong verbal inflection (time, person, etc.) and wrong part-of-speech (PoS) assumption during translation.
- **Lexical errors** are those involving just one word where the occurrence can be analysed as less dependent on the neighbouring words and more dependent on the source sentence. Subcategories of lexical errors are: extra word, absent word, not translated word and incorrectly translated word.
- Errors involving **n-grams** are those involving two or more words in similar subcategories of lexical errors.
- **Reordering** errors are those regarding wrong word order.

The error analysis was performed in a sample of 300 sentences from FAPESP's test-a (Aziz and Specia, 2011) translated from English to Brazilian Portuguese.
The annotation process was carried out by a native speaker of Brazilian Portuguese with good knowledge of English. Although this is not the ideal situation – that the MT error annotation be performed by only one person – it is important to mention that the same person has participated in the annotation of the same sentences, some years ago. Furthermore, since the annotation was carried out taking into account the NMT and the PBSMT outputs in parallel (that is, at the same time), there is a strong accordance in the annotations performed for the NMT output and the previous, now revised, PBSMT output.
To annotate all the occurrences of errors, the specialist used the BLAST[5] (Stymne, 2011) annotation tool and followed the guidelines of Martins (2014).

4.2. Quantitative Evaluation

Table 2 shows the numbers and percentages of MT errors found in the PBSMT and the NMT outputs for the 300 sentences automatically translated by both approaches and manually analysed by our human specialist.
Our main conclusions derived from the values on Table 2 are:

- The total amount of errors identified in the NMT output (902) is bigger than the total amount of errors identified in the PBSMT one (573).
- In both MT translation approaches, the majority of errors lies on the lexical category (40.83% in PBSMT and 66.75% in NMT) followed by syntactic errors (37.69% in PBSMT and 17.51% in NMT), errors involving n-grams (12.57% in PBSMT and 13.63% in NMT) and reordering errors (8.90% in PBSMT and 2.11% in NMT).
- There is a more similar distribution between the total amount of lexical and syntactic errors in PBSMT (234 $\times$ 216) than in NMT (602 $\times$ 158).

4.3. Qualitative Evaluation

Although the total amounts of errors listed in Table 2 can gives us the big picture of how the MT approaches behave, in this section we point out the main insights derived from our MT qualitative error analysis.

4.3.1. Syntactic Errors and Better Language Fluency in the NMT's Output

Some values in Table 2 catched our attention. It is interesting to notice that while in the PBSMT output there is a small difference (only 18 occurrences) between the amount of errors annotated as lexical (234) and syntactic categories

[4] http://www.nilc.icmc.usp.br/nilc/tools/Fapesp%20Corpora.htm

[5] http://stp.lingfil.uu.se/~sara/blast/

		PBSMT		NMT	
Error category	Error subcategory	Amount	%	Amount	%
Syntactic Errors	Number agreement	76	13.26	54	5.99
	Gender agreement	63	11.00	64	7.09
	Verbal inflection	65	11.34	35	3.88
	PoS	12	2.09	5	0.55
	TOTAL	**216**	**37.69**	**158**	**17.51**
Lexical errors	Extra word	25	4.36	99	10.98
	Absent word	71	12.39	218	24.17
	Not translated word	45	7.85	137	15.19
	Incorrectly translated word	93	16.23	148	16.41
	TOTAL	**234**	**40.83**	**602**	**66.75**
N-gram	Absent n-gram	4	0.70	26	2.88
	Not translated n-gram	0	0.00	2	0.22
	Incorrectly translated n-gram	68	11.87	95	10.53
	TOTAL	**72**	**12.57**	**123**	**13.63**
Reordering	Order	51	8.90	19	2.11
	TOTAL	**51**	**8.90**	**19**	**2.11**
TOTAL		**573**	**100**	**902**	**100**

Table 2: Manual annotation – amount and percentage of MT errors by error category in the output of each approach (PBSMT and NMT)

(216), this difference is huge (444 occurrences) in the NMT one (602 lexical errors against only 158 syntactic errors identified).

In our qualitative evaluation we noticed that this fact is reflected in the quality of the translations since the ones produced by the NMT system are more fluent than those generated by the PBSMT one.

To illustrate this fact Table 3 brings an example of a source (Src) sentence, in English, translated by both baseline systems in which there is a syntactic error (number agreement) in the PBSMT output and no error in the NMT output. The wrong word (*adequado*) is shown in bold accompanied by the indication of the error subcategory (syn_numberConc).

Table 4 brings another example with two syntactic errors (number and gender agreement) in the sentence translated by the PBSMT system and no error in the NMT system's output.

Src	The conclusions open up the possibility of indicating the most **suitable** exercises for specific diseases , something until now done only on the basis of intuition , without any experimental evidence .
Ref	As conclusões abrem a possibilidade de indicar os exercícios mais **apropriados** para doenças específicas , algo feito até agora com base apenas na intuição , sem evidências experimentais .
PBSMT Sys	As conclusões abrem a possibilidade de indicar os exercícios mais **adequado**$_{\text{syn\_numberConc}}$ para certas doenças , algo feito até agora somente a partir de intuição , sem nenhuma evidência experimental .
NMT Sys	As conclusões abrem a possibilidade de indicar os exercícios mais **adequados** para doenças específicas , algo até agora feito apenas com base na intuição , sem nenhuma evidência experimental .

Table 3: Example of a sentence translated by both baseline systems with one syntactic error (number agreement) annotated in the PBSMT system's output and no error in the NMT one's

4.3.2. Lexical Errors and the NMT's Vocabulary Limitations

On the other hand, our NMT baseline system seems to have more problems to deal with infrequent words than the PBSMT one.

Table 5 brings examples of lexical errors of absent (*running back*) and incorrectly translated words in the NMT system's (*em, no*) output. From those errors, only one also occurs in the PBSMT system's output (*em*)

Other lexical/n-gram errors were due to rare (infrequent) words mainly related to technical terms of a specific domain. Table 6 shows an example of an incorrectly translated n-gram in the translations generated by both baseline systems.

We believe that the infrequent words and n-grams have a bigger impact in the NMT approach than in the PBSMT one due to the way the final model is built. Due to computational optimization, the infrequent tokens (words and n-grams) are discarded (considered UNKNOWN tokens) during the model's training.

As a consequence of this NMT approach's limitation, the most frequent error subcategories in the NMT system's output are all of the lexical error category: absent word (24.17%), incorrectly translated word (16.41%), not translated word (15.19%) and extra word (10.98%).

To deal with this problem, one alternative is to control terminology in NMT systems. However, we strongly believe

Src	But there is one part of this research whose results can already be **adopted** .
Ref	Mas há uma vertente dessa pesquisa cujos resultados já podem ser **adotados** .
PBSMT Sys	Mas há uma parte dessa pesquisa cujos resultados já podem ser **adotada**$_{\text{syn\_numberConc syn\_genderConc}}$.
NMT Sys	Mas há uma parte dessa pesquisa cujos resultados já podem ser **adotados** .

Table 4: Example of a sentence translated by both baseline systems with two syntactic errors (gender and number agreement) annotated in the PBSMT system's output and no error in the NMT one's

Src	He arrives at the laboratory **at** eight , and whenever he manages to free himself from his family duties , ends **the** day **running**$_{\text{lex\_abstWord}}$ the 12 to 15 kilometers **back**$_{\text{lex\_abstWord}}$ home .
Ref	Chega ao laboratório **às** 8 e , sempre que consegue se liberar dos compromissos familiares , termina **o** dia **correndo** 12 a 15 quilômetros na **volta** para casa .
PBSMT Sys	Ele chega ao laboratório **em**$_{\text{lex\_incTrWord}}$ oito e , quando ele consegue livrar - se da sua família , termina o dia correndo 12 a 15 quilômetros de volta para casa .
NMT Sys	Ele chega ao laboratório **em**$_{\text{lex\_incTrWord}}$ oito , e sempre que consegue se livrar da família , termina **no**$_{\text{lex\_incTrWord}}$ dia 12 a 15 quilômetros de casa .

Table 5: Example of a sentence translated by the NMT baseline system with four lexical errors (two incorrectly translated words and two absent words)

Src	Another species that lives in the coastal strips and also has its survival at stake is the **tropical mockingbird** (Mimus gilvus) .
Ref	Outra espécie que vive nas restingas , cuja sobrevivência também está em jogo , é o **sabiá - da - praia** (Mimus gilvus) .
PBSMT Sys	Outra espécie que vive nas faixas litorâneas e também tem sua sobrevivência em jogo é **tropicais mockingbird**$_{\text{grm\_incTrGram}}$ (Mimus gilvus) .
NMT Sys	Outra espécie que vive nas faixas costeiras e também tem sua sobrevivência em jogo é a **mockingbird tropical**$_{\text{grm\_incTrGram}}$ (Mimus gilvus) .

Table 6: Example of a sentence translated by both baseline systems with a n-gram error (incorrectly translated n-gram) annotated in the NMT system's output and in the PBSMT one's

that it is necessary to invest in the identification of word sequences as multiword expressions as a way to correctly translate them as a whole unit. We also see the use of a bilingual lexica with (single and multiword) terms and corpora of the specific domain as fundamental tools to better handle terminology in both MT approaches.

4.3.3. Other Qualitative Insights

During the annotation of the test corpus, the human specialist reported some insights that were confirmed by the numbers on Table 2 such as: (i) the NMT system has more problems with gender agreement and the PBSMT one's with the number agreement and (ii) the NMT system has less reordering errors than the PBSMT one. Table 7 shows a reordering error in the PBSMT system's output that do not occur in the NMT one's.

Src	The actions regarded as a priority are : expanding the extent of the regions now protected by law and carrying out more **wide - ranging surveys** of the species of plants and animals found in the coastal strips , besides developing environmental education programs in the coastal areas .
Ref	As ações consideradas prioritárias : aumentar a extensão das áreas já protegidas por lei e realizar **levantamentos mais abrangentes** das espécies de plantas e animais encontrados nas restingas , além de desenvolver programas de educação ambiental nas regiões litorâneas .
PBSMT Sys	As ações consideradas prioritárias são : Ampliar a extensão das áreas já protegidas por lei e fazer **mais amplos levantamentos**$_{\text{ord}}$ de espécies de plantas e animais encontrados nas faixas litorâneas , além de desenvolver programas de educação ambiental nas áreas costeiras .
NMT Sys	As ações consideradas prioritárias são : ampliar a extensão das regiões já protegidas por lei e realização de **levantamentos mais abrangentes** das espécies de plantas e animais encontrados nas faixas costeiras , além de desenvolver programas de educação ambiental nas áreas costeiras .

Table 7: Example of a sentence translated by both baseline systems with one reordering error annotated in the PBSMT system's output and no error in the NMT one's

However, to be able to point out possible causes for these insights we need to augment our annotated corpus with more occurrences of these error subcategories. This is one of our future works.

4.3.4. Regarding the Related Work

However other insights were not possible to be confirmed based on the values on Table 2. One example of this is the fact that the PBSMT system seems to deal better with prepositions than the NMT one. Popović (2018) had already pointed out that the NMT system evaluated by her produced errors related to prepositions.

Table 8 shows an example of a sentence for which the PBSMT system was able to correctly translate the source preposition (*to*) and the NMT one was not. In this case, the error category was n-gram because the source sequence (to the) was translated as just one word in the reference sentence (*ao*, which is the combination of the preposition *a* and the determiner *o*).

Src	In the course of five months , 20 biologists covered 1,600 kilometers of Brazilian coastline , from the south of Rio de Janeiro **to the** south of Bahia .
Ref	Durante cinco meses , 20 biólogos percorreram 1.600 quilômetros do litoral brasileiro , do sul do Rio de Janeiro **ao** sul da Bahia .
PBSMT Sys	Durante cinco meses , 20 biólogos percorreram 1.600 quilômetros da costa brasileira , do sul do Rio de Janeiro **ao** sul da Bahia .
NMT Sys	Ao longo de cinco meses , 20 biólogos percorreram 1.600 quilômetros da costa brasileira , do sul do Rio de Janeiro **para o**$_{\text{n-gram\_incTrGram}}$ sul da Bahia .

Table 8: Example of a sentence translated by both baseline systems with a n-gram error (incorrectly translated n-gram, involving a preposition) annotated in the NMT system's output and no error in the PBSMT one's

Another conclusion of Popović (2018) also corroborated by this work is that the NMT produces errors regarding ambiguous words. Table 9 brings an example of an ambiguous source word (arms) incorrectly translated by the NMT (as *armas*) but correctly translated by the PBSMT (as *braços*).

Src	**Known** as an **ischemic** stroke or cerebro vascular accident (**CVA**) , this problem can lead to the immobility of **arms** and legs , and even to the loss of speech .
Ref	**Conhecido** como acidente vascular cerebral isquêmico (**AVC**) ou **isquemia** cerebral , esse problema pode levar à imobilidade de **braços** e pernas e até mesmo à perda da fala .
PBSMT Sys	**Conhecida**$_{\text{syn\_genderConc}}$ como **ischemic**$_{\text{lex\_notTrWord}}$ **infarto**$_{\text{lex\_incTrWord}}$ ou acidente encefálico (**CVA**$_{\text{lex\_notTrWord}}$) , esse problema pode levar à imobilidade de **braços** e pernas e até mesmo à perda da fala .
NMT Sys	**Conhecida**$_{\text{syn\_genderConc}}$ como derrame **ischemic**$_{\text{lex\_notTrWord}}$ ou acidente vascular cerebral (**CVA**$_{\text{lex\_notTrWord}}$) , esse problema pode levar à imobilidade de **armas**$_{\text{lex\_incTrWord}}$ e pernas e até à perda de fala .

Table 9: Example of a sentence translated by the NMT baseline system with one syntactic (gender agreement) and three lexical errors (two of not-translated words and one of incorrectly translated word)

5. Conclusions and Future Work

In this work we presented the first error analysis of a neural machine translation (NMT) system for Brazilian Portuguese. This analysis was carried out in parallel with the output of other MT system: a phrase-based statistical machine translation (PBSMT) one.

As show in section 4., the most frequent category of errors are the same in both MT approaches: lexical errors followed by syntactic errors. However, we also presented that the NMT system is better than the PBSMT one to deal with syntax and order of words; but the PBSMT is able to better handle words with less errors of extra word, absent word and not-translated word.

Our conclusions also corroborate the ones of Popović (2018): the NMT system showed better fluency, regarding word order and morphology aspects of the languages, than the PBSMT one.

From the point of view of post-editing, since the lexical errors are easier to correct than the syntactic ones, both manually or automatically, we think that the results present in this paper show a good perspective for the automatic post-editing for NMT output.

As our future work we intend to augment the annotated corpus with more sentences from the same domain. With a bigger corpus we think that would be possible to go deeper in the finds about MT errors and propose some approaches to automatically correct them. Thus, it is also a future work to use the annotated corpus to train an automatic post-editing tool for correcting the main MT errors.

The annotated corpora are available at `https://github.com/LALIC-UFSCar/FAPESP-PBSMT-NMT`.

Acknowledgements

This work has been developed with the support from São Paulo Research Foundation (FAPESP), grants #2016/21317-0 (Undergraduate research grant) and #2016/13002-0 (MMeaning Project).

6. Bibliographical References

Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Scalco, M. (2006). Open-source Portuguese-Spanish machine translation. In *Proceedings of the VII Encontro para o Processamento Computacioal da Língua Portuguesa Escrita e Falada (PROPOR)*, pages 50–59, Itatiaia, RJ, Brazil.

Aziz, W. and Specia, L. (2011). Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL)*, page 234–238, Cuiabá, MT, Brazil.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Yoshua Bengio et al., editors, *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 257–267, Austin, Texas, November. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Caseli, H. M. (2007). *Indução de léxicos bilíngües e regras para a tradução automática*. Doutorado em Computação e Matemática Computacional, ICMC-USP.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In Dekai Wu, et al., editors, *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST@EMNLP)*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Klubička, F., Toral, A., and Sánchez-Cartagena, V. M. (2017). Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 108:121–132, June.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press, Cambridge.

Martins, D. B. J. and Caseli, H. M. (2015). Automatic machine translation error identification. *Machine Translation*, 29(1):1–24.

Martins, D. B. J. (2014). Pós-edição automática de textos traduzidos automaticamente de inglês para português do brasil. Master's thesis, PPGCC/UFSCar.

Och, F. J. and Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th workshop on statistical machine translation (WMT)*, pages 392–395, Lisbon, Portugal.

Popović, M. (2018). Language-related issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation*, 32(3):237–253.

Senellart, P. and Senellart, J. (2005). SYSTRAN translation stylesheets: machine translation driven by XSLT. In *XML Conference and Exposition*, pages 1–15, Atlanta, USA.

Stymne, S. (2011). Blast: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pages 56–61. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.