Evaluating machine translation in a low-resource language combination: Spanish-Galician

María do Campo Bayón
Grup Tradumàtica
Universitat Autònoma de Barcelona
maria.docampo@e-campus.uab.cat

Pilar Sánchez-Gijón
Grup Tradumàtica
Dept. de Traducció i d'Interpretació i
d'Estudis de l'Àsia Oriental
Universitat Autònoma de Barcelona
pilar.sanchez.gijon@uab.cat

Abstract

This paper reports the results of a study designed to assess the perception of adequacy of three different types of machine translation systems within the context of a minoritized language combination (Spanish-Galician). To perform this evaluation, a mixed design with three different metrics (BLEU, survey and error analysis) is used to extract quantitative and qualitative data about two marketing letters from the energy industry translated with a rule-based system (RBMT), a phrase-based system (PBMT) and a neural system (NMT). Results show that in the case of low-resource languages rule-based and phrase-based machine translations systems still play an important role.

1 Introduction

In the last couple of years, Neural Machine Translation is gaining more attention in the translation industry and becoming more popular thanks to the considerably good results obtained in certain language combinations. Nevertheless, low-resource languages and minoritized languages represent some challenges for machine translation (MT) usage and training. This paper describes the process followed to test and evaluate three different MT systems in a closely related language combination such as Spanish-Galician.

2 Aim of this study

This study aims to determine which type of MT system (RBMT, PBMT or NMT) is perceived as more adequate in the context of a minoritized language such as Galician in an MT+Post-editing (PE) workflow. For that purpose, the quality of all three raw outputs was established with the following metrics:

- Evaluating which type of MT system obtains better results applying the BLEU metric.
- Evaluating which type of MT system obtains better results in a human evaluation (quality perception survey conducted among professional post-editors).
- Evaluating which type of MT system obtains better results following an error analysis framework (MQM).

3 Background

3.1 NMT Evaluation

With the outbreak of NMT, many studies have tried to shed some light on the real and the perceived quality of this kind of MT systems. Shterionov et al. (2018) show that a few translators see NMT as a booster of their productivity. Some translators even see (N)MT as a handicap for their productivity while others perceive it the other way around (Sánchez-Gijón et al., 2019). In terms of NMT quality perception, Castilho et al. (2017) conclude that raw NMT segments may not be preferred by translators. In the same paper, they concluded that, compared to PBMT, NMT represents a step forward but it implies also some limitations. The same idea of strengths and weaknesses on NMT with respect to PBMT can be found in Popovic, 2017. Most of these studies describe tests involving language combinations of high-resource languages. This paper approaches this topic from the perspective of a low-resource language: Galician.

3.2 MT in Galician

As a minoritized language, Galician represents a serious challenge to develop MT systems due to the lack of technological and data resources. In recent years, there has been an enormous effort, mainly from the academic community, to develop Natural Language Processing (NLP) resources and compile corpora such as GalNet, the Galician WordNet (Gomez Guinovart & Solla Portela, 2017), SemCor (Solla Portela & Gomez Guinovart, 2017), several terminology projects (Solla Portela & Gomez Guinovart, 2015), big corpus annotation (Gomez Guinovart & Lopez Fernández, 2009), Freeling (Padro & Stanilovsky, 2012) and Linguakit (Gamallo & Garcia, 2017).

There are also some MT systems specifically created for Galician: the RBTA MT system of the *Centro Ramón Piñeiro para a Investigación en Humanidades* (Diz Gamallo, 2001), OpenTrad Apertium (Armentano-Oller & Forcada, 2006) and Carvalho PBMT system (Pichel Campos et al., 2009).

Nevertheless, the need to keep investigating in NLP and Deep Learning (DL) in Galician is very clear in order to develop the corpora and the strategies needed to train phrase-based and neural systems and obtain better results (Agerri et al., 2018: 2322).

4 Methodology

The investigation is divided into three different phases. The first one consisted of choosing the source document to be processed by the three different MT systems. Two marketing letters of approximately 500 words in total with specific terminology from the energy industry were chosen. After that, RBMT (OpenTrad Apertium) and PBMT (ModernMT v. 2.5) systems were created and trained. In the case of Apertium, the stable version of the pair Spanish-Galician was downloaded into an Ubuntu environment and trained with specific terminology of the source field. Similarly, a new engine was created in MMT v. 2.5 and trained with a thematic translation memory (TM) of 4315 translation units and a parallel corpus of 6 million words from the legal and administrative field. Lastly, regarding NMT system, due to the lack of enough high-quality training data, we selected Google Neural Spanish-Galician engine to perform the texts.

Once all three MT systems in the language combination Spanish-Galician were available, a set of 32 Spanish segments was translated with each of them. The quality of the raw MT segments obtained was measured in the next phase of the investigation following different approaches.

The second phase of the investigation involved the evaluation of the quantitative data results obtained applying the automatic metric Bilingual Evaluation Understudy, abbreviated as BLEU (Papineri et al., 2001). Then, a survey was designed to compile qualitative information about the quality perception of Spanish-Galician post-editors. For that purpose, a sample of 14 segments from the whole set was used. 69 professional translators with experience in Spanish-Galician post-editing were selected from the CPSL Language Solutions resource database and Proz portal. Finally, 15 people participated in that survey. To complete the qualitative results, an error analysis was performed following the MQM framework. Once all the individual results were analysed, a global evaluation was performed to triangulate the resulting data.

5 Results

5.1 Automatic evaluation

The BLEU score on the whole set of segments is as follows:

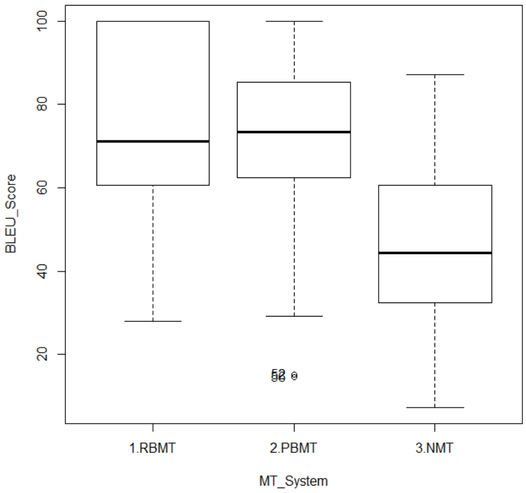

Figure 1. BLEU Score

RBMT and PBMT segments show higher scores than NMT. There is not significant difference between RBMT and PBMT scores, but differences are significant between these two systems and NMT:

	RBMT	PBMT	NMT
RBMT	**1**	0.831	**<0.0001**
PBMT	0.831	**1**	**0.000**
NMT	**<0.001**	**0.0004**	**1**

Figure 2. p-values per pairs

Finally, 14 of the source segments contains more than 30 words. These segments were identified as long segments and analysed separately. This is the BLUE score obtained in the subset of 14 long segments analysed by post-editors:

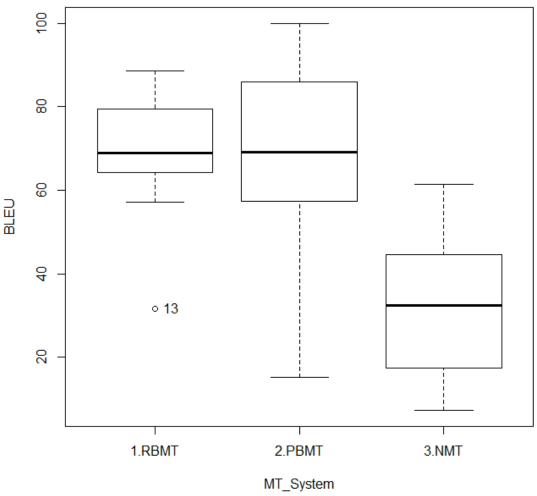

Figure 3. BLEU Score of long segments

RBMT segments show higher and more homogeneous scores than NMT and PBMT. Differences between NMT and both PBMT and RBMT are significant. Differences between PBMT and RBMT are not significant:

	RBMT	PBMT	NMT
RBMT	**1**	1.000	**0,0002**
PBMT	0.831	**1**	**0.004**
NMT	**0.0002**	**0.004**	**1**

Figure 4. p-values per pairs in long segments

5.2 Human evaluation

The human evaluation was designed to gather two different pieces of information segment by segment: ranking of MT system and which MT system is considered good enough to be post-edited. 14 translated segments, one by each MT system, were selected as sample. Equal translation results from different MT systems or too bad translations were excluded from the survey in order not to distort the survey results.

5.2.1 Global evaluation

Human evaluators were asked to answer 2 questions. In each question, they had to rank the three different raw machine translations as 1st, 2nd and 3rd place. Then, they had to specify if they would use or not the machine translation to post-edit (binary response). In relation with BLEU scores, the results of usable/not usable segments show that RBMT and PBMT would be used to be post-edited.

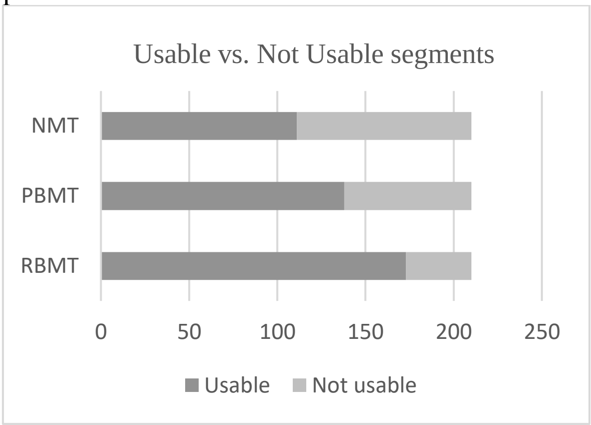

Figure 5. Segments usable vs. not usable to post-edit

To analyse this data, the non-parametric statistical test Cochran 's Q test is applied.

Variable	Categories	Frequencies	%
RBMT	0	37	17.619
	1	173	82.381
PBMT	0	72	34.286
	1	138	65.714
NMT	0	99	47.143
	1	111	52.857

C (Observed value)	42.014
C (Critical value)	5.991
FD	2
p-value	**<0.0001**
Alfa	0.05

Figure 6. Cochran's Q test results

Differences are significant (p-value < 0.0001). Proportions among the three groups are statistically significant (Marascuilo procedure):

Contrast	Value	Critical Value	Significance
|p(RBMT – p(PBMT)|	0.167	0.103	Yes
|p(RBMT – p(NMT)|	0.295	0.106	Yes
|p(PBMT – p(NMT)|	0.129	0.116	Yes

Figure 7. Marascuilo procedure results

And the proportions show that the three groups are different:

Sample	Proportion	Groups		
NMT	0.529	A		
PBMT	0.657		B	
RBMT	0.824			C

Figure 8. Proportions of MT systems

Regarding the ranking, these are the results from each segment. As Figure 9 shows, RBMT and PBMT are better positioned that NMT. Post-editors also agreed that all segments selected as 1st place would be used to post-edit.

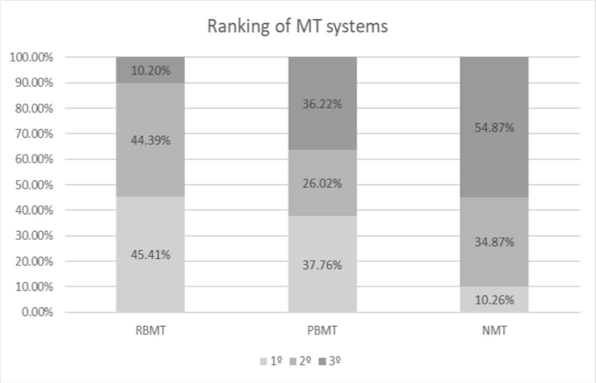

Figure 9. Ranking of MT systems

To establish whether these differences among MT systems are significant, the Kruskal-Wallis test was applied. Comparing results per pairs, in all cases p-value was under 0.05, meaning that differences are significant.

5.2.2 Long segments

Four of the source segments contains more than 30 words. These segments were identified as long segments and analysed separately.

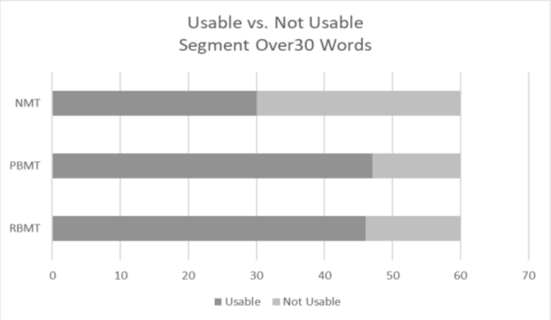

Figure 10. Usable vs. not usable in long segments

To analyse this data, the non-parametric statistical test Cochran 's Q test is applied.

Variable	Categories	Frequencies	%
RBMT	0	14	23.333
	1	46	76.667
PBMT	0	13	21.667
	1	47	78.333
NMT	0	30	50.000
	1	30	50.000

C (Observed value)	15.167
C (Critical value)	5.991
FD	2
p-value	**0.001**
Alfa	0.05

Figure 11. Cochran's Q test results in long segments

Differences are significant (p-value = 0.001). Proportions are statistically significant, but not among all three groups (Marascuilo procedure):

And the proportions show that there are differences between NMT and the other two MT systems:

Contrast	Value	Critical Value	Significance
|p(RBMT – p(PBMT)|	0.017	0.187	No
|p(RBMT – p(NMT)|	0.267	0.207	Yes
|p(PBMT – p(NMT)|	0.283	0.205	Yes

Figure 12. Marascuilo procedure results in long segments

Sample	Proportion	Groups	
NMT	0.500	A	
PBMT	0.767		B
RBMT	0.783		B

Figure 13. Proportions in long segments

Usable scenario in long segments differs from the whole document. Figure 14 shows which segment from each MT system would be chosen to be post-edited in 1st, 2nd and 3rd place.

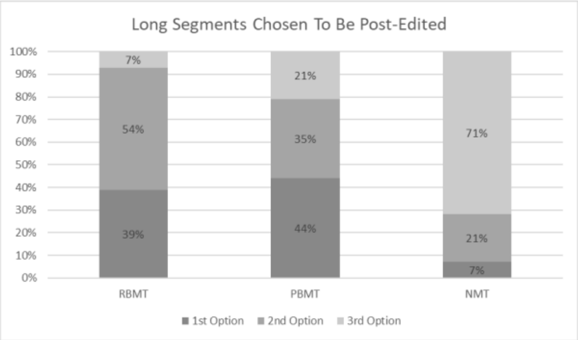

Figure 14. Ranking of MT systems in long segments

To establish whether these differences among MT systems are significant, the Kruskal-Wallis test was applied. Comparing results per pairs, the p-value was under 0.05 only between NMT and the other to MT systems:

	RBMT	PBMT	NMT
RBMT		No	**Yes**
PBMT	No		**Yes**
NMT	**Yes**	**Yes**	

Figure 15. Statistical differences in long segments

5.3 Error analysis

A Multidimensional Quality Metrics (MQM) customized framework was used to identify the errors made by each MT system. Only relevant types of errors from accuracy, fluency, style and terminology were selected. Figure 16 shows the total number of errors obtained per segment in each MT system:

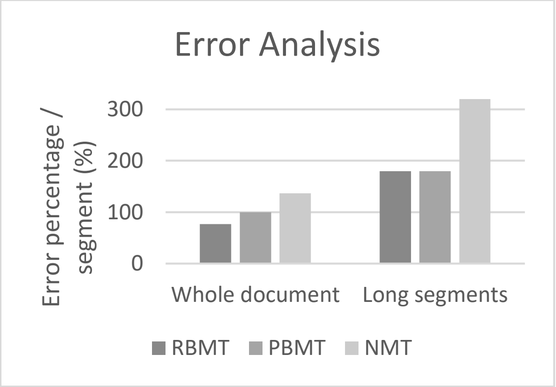

Figure 16. Error percentage per segment

Regarding the type of errors, there's a clear prevalence in all MT systems of mistranslations, gender and number agreement errors, function-words errors, word-order errors and unidiomatic expressions. Also, PBMT and NMT made more addition, omission, orthography, typography and part-of-speech errors, and domain terminology inconsistencies. RBMT and NMT registered verb concordance errors and awkward constructions. Finally, the only system with register errors was NMT.

The clearest example of error in RBMT is the wrong identification of the preposition *para* (*for*, in English) and the undefined feminine form of the article: *una* (*a* in English). RBMT interprets these words as verbs so they are translated as *parar* (*stop*) and *unir* (*join*). PBMT sometimes makes errors in verbal constructions such as the wrong translation of *hemos venido* by *comezamos viñesen (we started coming* instead of *we have come*).

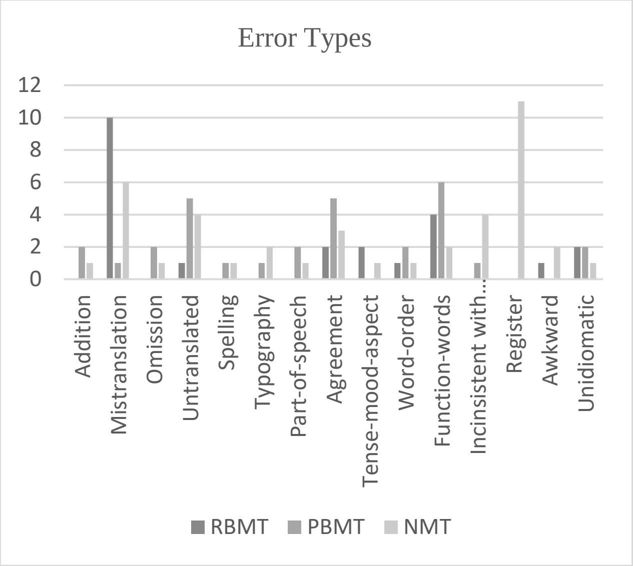

Figure 17. Error types

Finally, a repetitive error in NMT is the change of register. In this phrase, not only the verbal form is incorrect but also changes from the second person plural to the singular even if it is the same in the source text:

Spanish: Instala ahora el gas y disfruta de todas sus ventajas [...] en todos los rincones de tu casa

Galician: Agora *instálalle* [incorrect verbal construction] *o gas e goce* [second person plural] *de todas as súas* [second person plural] *[...] en todos os recunchos da túa* [second person singular] *casa.*

6 Conclusions

The main conclusion is that although NMT seems promising in frequent language combinations, especially if English is involved, it is not obtaining the desired results in low-resource languages such as the pair Spanish-Galician. NMT has not yet unseated RBMT and PBMT, performing, in fact, worse than these systems.

This small study reveals that more tests should be done to replicate results and evaluate special needs to have a competitive NMT. Subsequent investigations must consolidate error patterns of each system to address some of the most prominent issues. Thus, there's a clear need to work in the access to the resources and parallel data needed to train MT systems, especially in PBMT and NMT.

Consequently, any future lines of investigation in MT and minoritized languages should be focused on searching and optimizing NLP and text resources.

Acknowledgement

This work was supported by the ProjecTA-U project "Machine Translation Users", grant number FFI2016-78612-R [MINECO / FEDER, UE].

References

Agerri, R., Gómez Guinovart, X., Rigau, G. & Solla Portela, M. A. 2018. Developing New Linguistic Resources and Tools for the Galician Language. *Proceedings of the 11th Language Resources and Evaluation Conference (LREC'18)*: *2322-2325.*

Armentano-Oller, C. & Forcada, M. L. 2006. Open-source machine translation between small languages: Catalan and Aranese Occitan. *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages).* May 22-28, p. 51-54.

Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S. et al. 2006. Open-source Portuguese-Spanish machine translation. In *Computational Processing of the Portuguese Language: 7th Workshop on Computational Processing of Written and Spoken Portuguese*, PROPOR. Lecture Notes in Artificial Intelligence 3960. Springer-Verlag, 50–59.

Burchardt, A. & Lommel, A. 2014. *Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality*. Available at <http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf >.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. 2017. Is Neural Machine Translation the New State of the Art? In *The Prague Bulletin of Mathematical Linguistics*, 108:1, 109–120. <https://doi.org/10.1515/pralin-2017-0013>.

Diz Gamallo, I. 2001. The importance of MT for the survival of minority languages: Spanish-Galician MT system. *Proceedings of MT Summit VIII*, November 2001, Spain.

Gamallo, P. & Garcia, M. 2017. Linguakit: a multilingual tool for linguistic analysis and information extraction. *Linguamática*, 9(1): 19–28.

Gómez Guinovart, X. & López Fernández, S. 2009. Anotación morfosintáctica do Corpus Técnico do Galego. *Linguamática*, 1(1): 61–71.

Gómez Guinovart, X. & Solla Portela, M. A. 2017. Building the galician wordnet: methods and applications. *Language Resources and Evaluation,* 52 (1): 317–339.

Iglesias, G., Rodríguez Liñares, L., Rodríguez Banga, E., Campillo Díaz, F. L. & Méndez Pazó, F. 2010. Perspectivas de la traducción automática castellano-gallego mediante técnicas estadísticas y por transferencia. *IV Jornadas en Tecnología del Habla*, November 8-10 of 2006, Zaragoza. pp. 111-116.

Padró, L. & Stanilovsky, E. 2012. Freeling 3.0: Towards wider multilinguality. In Nicoletta Calzolari et al. (Eds.). *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 2473–2479, Istanbul, Turkey.

Papineni, K., Roukos, S., Ward, T. & Zhu, W. 2001. *BLEU: A Method for Automatic Evaluation of Machine Translation*, IBM Research Report RC22176 (W0109−022).

Popović, M. 2017. Comparing Language Related Issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, 108(1): 209-220.

Pichel Campos, J. R., Malvar Fernández, P., Senra Gómez, O., Gamallo Otero, P. & García González, A. 2009. Carvalho: English-Galician SMT system from EuroParl English-Portuguese parallel corpus. *Procesamiento del Lenguaje Natural*, 23: 379-381.

Sánchez-Gijón, P., Moorkens, J., & Way, A. (2019). Post-editing neural machine translation versus translation memory segments. *Machine Translation*, 31-59. <https://doi.org/10.1007/s10590-019-09232-x>.

Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'Dowd, T. & Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translaiton*, 32, 217–235. <https://doi.org/10.1007/s10590-018-9220-z>.

Solla Portela, M. A. & Gómez Guinovart, X. 2015. Termonet: Construcción de terminologías a partir de WordNet y corpus especializados. *Procesamiento del Lenguaje Natural*, 55:165–168.

Solla Portela, M. A. & Gómez Guinovart, X. 2016. Dbpedia del gallego: recursos y aplicaciones en procesamiento del lenguaje. *Procesamiento del Lenguaje Natural*, 57:139–142.

Solla Portela, M. A. & Gómez Guinovart, X. 2017. Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con wordnet 3.0. *Procesamiento del Lenguaje Natural*, 59:137–140.