Quantitative and Qualitative Evaluation of Human and Machine-Translated EU Economic Texts in the English-Greek Language Pair

Dimitris Boumparis
Master's Student
University of Antwerp, Belgium
dimitrios.boumparis@student.uantwerpen.be

Christos Giannoutsos
Freelance Translator
Kremasti, Rhodes, Greece
christosgiannoutsos@gmail.com

Abstract

Studying the literature regarding EU texts, we see that there are publications on the usefulness of neural machine translation (NMT) for translators in EU institutions, who professionally translate from a major into minor languages. However, there is a research gap regarding the quantitative analysis of human and machine-translated texts in major-to-minor language combinations. This paper explores the quantitative characteristics of both kinds of texts and tries to profile them. We explore the impact of many input textual features, including word n-grams frequencies, Parts of Speech (PoS), function words, punctuation, and a number of stylometric indices (e.g., readability index, type/token ratio, and mean sentence length). We compiled a corpus of 646 English press releases manually translated into Greek by professionals in the European Central Bank, along with their NMT counterparts by state-of-the-art systems. We produce input features for a Support Vector Machines (SVM) algorithm that can predict whether a text is produced by a human or an NMT system and achieves 90% accuracy and f1-macro score. Furthermore, we compare the similarity between the original and the NMT outputs using methods for dimensionality reduction and cluster analysis (PCA, HCA, t-SNE). Finally, we evaluate the quality of the NMT outputs using BLEU.

1 Introduction

"The closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002, 311). Recently, we witnessed a deep learning system outperform translation professionals in the task of translation from a major (English) into a minor language (Czech) (Popel et al., 2020). After studying the literature regarding EU texts, we see that there are publications on the usefulness of neural machine translation (NMT) for translators employed by European Union institutions, who are professionally involved in translating from a major language, such as English (Rado, 1987; Song, 1991; Cronin, 2003; Parianou, 2009), into minor languages, such as Hungarian (Lesznyák, 2019), Polish (Stefaniak, 2020) and Slovene (Arnejšek and Unk, 2020). This raises the question of the main quantitative textual characteristics of translations that have been produced by both human professionals and the MT systems NMT systems, and, by extension, whether there is a significant difference between the two.

2 Previous work

The research conducted on this topic is limited to studying the length (Pouget-Abadie et al., 2014) or the lexical richness (Vanmassenhove et al., 2019, 2021) of the two versions of the texts. In our case, the human-produced texts are, on average, 10.3% longer than the machine-translated ones. While this paper will not go further and discuss why that seems to be the case, it seems that this result can be attributed to the critical thinking that takes place during the translation process in the minds of the professionals (Wu et al., 2016; Stasimioti and Sosoni, 2020). Translators are well aware that the text they are producing is going to be read and hence understood by another human being, so they may dive deeper into the text and translate more freely if it serves the purpose of the translation.

The differences between various MT systems, with regards to the quality of their output and the types of errors included therein, have been reported by several recent studies. Some of them (Bahdanau et al., 2014; Jean et al., 2015; Junczys-Dowmunt et al., 2016; Dowling et al., 2018) relied on automatic evaluation metrics (AEMs) like BLEU (Papineni et al., 2002) and HTER (Snover et al., 2006); others used human evaluations of the MT output quality, employing adequacy and fluency ratings

New Trends in Translation and Technology, pages 248–253
4–6 July, 2022. ©NeTTT 2022

(Bentivogli et al., 2016), manual error analyses (Klubička et al., 2017; Popovic, 2017; Klubička et al., 2018) or a combination of methods (Burchardt et al., 2017; Castilho et al., 2017a,b, 2018; Toral and Sánchez-Cartagena, 2017; Shterionov et al., 2018; Koponen et al., 2019; Jia et al., 2019; Sosoni and Stasimioti, 2019).

In other words, we saw a gap in research regarding the evaluation of human and machine-translated texts when it comes to the quantitative analysis of such texts. This paper explores the various quantitative characteristics of both kinds of texts and tries to profile them. We explore the impact of a large number of input textual features, including, but not limited to, frequencies of character and word n-grams, Part-of-Speech, function words, and punctuation (especially pronouns and full stops, respectively), use of capitalisation, and a number of stylometric indices such as the readability index, type/token ratio (TTR), and mean sentence length. All these indices have been proved to be relevant in quantitative analysis of texts (Read, 2000; Grieve, 2007; Wu et al., 2021).

3 Corpus

To do our analysis, we compiled a parallel corpus of 646 press releases of the European Central Bank (ECB) in English[1] that have been manually translated into Greek by professionals who work or have worked in EU institutions, along with their machine-translated counterparts by a state-of-the-art NMT system. The length of the initial 2,572 texts ranged from 13 to 1,135 words. However, we excluded those with less than 20 words and, after we eliminated those not translated into Greek, our final corpus contained 646 texts.

4 Methodology

We then proceed to quantitatively analyse the texts at hand and produce input features for a Support Vector Machines (SVM) algorithm that can predict whether a text is produced by a human translator or an NMT system with very high accuracy. However, we are not limiting our research to just the quantitative characteristics of the MT texts but also go further and evaluate the translations in terms of the similarity between the original and the NMT output by employing cosine similarity, Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), and t-Distributed Stochastic Neighbouring Entities (tSNE).

[1]Available at https://www.ecb.europa.eu/press/pr/date/html/index.en.html.

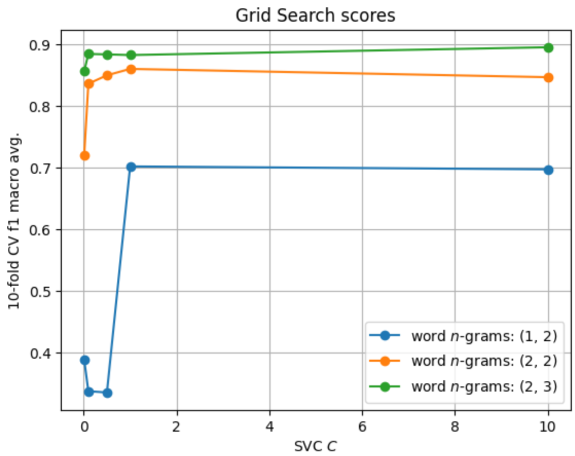

Figure 1: F1-macro scores for different word *n*-grams of the vectorisers.

	Analyser	# features
Vect 1	character bigrams	1000
Vect 2	character trigrams	1000
Vect 3	word bigrams	1000
Vect 4	word trigrams	1000

Table 1: Parameters of the four TF-IDF vectorisers.

4.1 Supervised learning

Our supervised approach included a combination of text vectorizers, provided by the Scikit-Learn open-source Python library[2].

We implemented four term frequency - inverse document frequency (TF-IDF) vectorizers with 1000 features each, as shown in Table 1. We experimented with various *n*-gram ranges and did not preprocess the texts further. As shown in Figure 1, bigrams and trigrams improved the performance of the classifier's 10-fold Cross-Validation significantly. The Support Vector Machine (SVM) classifier achieved .90 f1-macro score on the task of binary classifying translations as produced by a human or Google's NMT system.

4.2 Unsupervised learning

As far as the unsupervised learning is concerned, we applied a variety of dimensionality reduction algorithms to examine whether the two versions of the texts in question form separate enough clusters. In particular, we implemented PCA, HCA, and

[2]Available at https://scikit-learn.org.

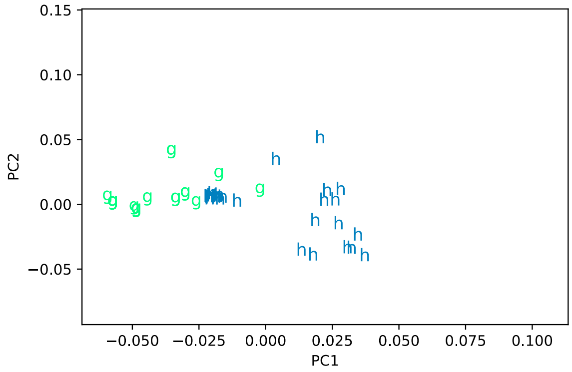

Figure 2: Clusters formed after implementing Principal Component Analysis (PCA) on the unique texts that were translated by humans (h) and Google's NMT API (g).

tSNE. The most apparent, as well as easy to depict, results were given through PCA (Figure 2). It is important to mention that, given the large number of available texts, we excluded those that are almost identical, while keeping only the first one that we came across. To achieve this, we used spaCy's cosine similarity[3] among all the texts. All but the first text that achieved at least 80% similarity were excluded from the dimensionalilty reduction tasks in order to guarantee an easy-to-interpret graph. The HCA graph was also in line with the PCA, although hard to fit given the limited space here.

5 Results and Evaluation

In terms of evaluating the quality of the machine translation, we employed the BiLinugal Evaluation Understudy (BLEU). The MT scored barely under 40, which generally translates to very understandable texts, although not yet matching the quality of a human-produced translation.

5.1 SVM classifier

Our supervised approach scored a very high f1-macro (.90) in binary classification of the texts, just by employing the built-in vectoriser provided by Scikit-Learn with a combination of character and word bigrams and trigrams. That, in it of itself, is a very promising indication that MT has not yet reached human-level quality and that the use Machine Translation Post-editing (MTPE) is having an upward trend.

[3] Read more at `https://spacy.io/api/span#similarity`.

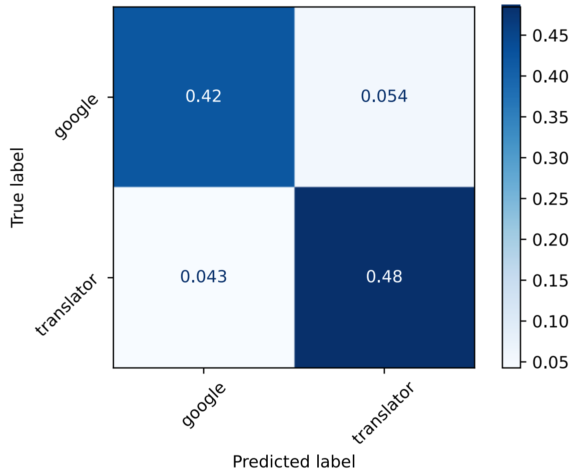

Figure 3: Confusion matrix of the SVM classifier using only the vectorisers (.90 accuracy and f1-macro score).

5.2 PCA

The unsupervised approach has proved as well that translations between human and machine-produced are easy to tell apart. However, there is a very visible cluster forming in the middle and towards the Google's side with translation that are–presumably–produced by translation professionals. The graph in Figure 2 suggests that these texts were translated by an NMT system and post-edited by professionals.

5.3 Quantitative analysis

The quantitative analysis included what stylometric indices, if any, are significantly different between the two versions of the texts. Among the numerous indices and features we calculated, only the following ones show a promising difference between human and machine-produced translations:

- **Average word and sentence length:** Human translations tend to be longer in terms of total words (29.4 in average) and sentences (1.06 in average).
- **Mean number of polysyllabic tokens:** Human translations tend to use longer words more (20.4 in average).
- **Average words per sentence:** Machine translations tend to have more words per sentence (2.56 in average).

The last observation is what captured our attention, mainly because it contradicts the first one, namely that human-produced translations tend to be longer. According to Popovic (2020, 4), "some translators might tend to generate longer sentences in the target text than others".

5.4 Qualitative analysis

Last but not least, we set to define the qualitative aspects that differentiate translations produced by humans with those by a neural translation system. After making an extensive list of Greek function words, we found some examples that we are of the opinion that are worth mentioning. In machine translations:

- Archaic words and phrases were used whereas none is present in the human ones. For example: τω (article used in now-deprecated dative case - only used in fixed phrases), εν όψει (= in view of), πέραν (= beyond, besides).
- There were tokens consisting of words not properly split with the previous or following punctuation sign or a handful of short English words that were not translated into Greek.
- Problems with fluency were observed, as defined in the DQF-MQM error typology (Lommel and Melby, 2018). A common example, that is often an indicator of MT, is the translation of *non-* in front of words (e.g., non-existent) that were translated with a dash (μη-) which is not applicable to Greek.

6 Future work

This is the first time such an analysis is performed in EU texts, at least for the English-Greek language pair. Next steps could include using stylometric features to increase the classifier's f1 score as well as conducting identical experiments but in the Greek-English pair. It would be interesting to see whether the results of the latter differ from the ones above. Finally, a Deep Learning approach to leverage Transformers' efficiency could also yield better results. We hope this paper sparks the curiosity of professors and researchers dedicated to uncovering the differences between human and machine translations.

References

Mateja Arnejšek and Alenka Unk. 2020. Multidimensional assessment of the eTranslation output for English–Slovene. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 383–392, Lisboa, Portugal. European Association for Machine Translation.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan thorsten Peter, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural mt engines. *Prague Bulletin of Mathematical Linguistics*, 108(1):159–170.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017a. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Miceli Barone, and Maria Gialama. 2017b. A comparative quality evaluation of pbsmt and nmt using professional translators. In *Proceedings of MT Summit XVI, vol.1: Research Track*, pages 116–131. Machine Translation Summit XVI 2017, MT XVI 2017 ; Conference date: 18-09-2017 Through 22-09-2017.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Andy Way, and Panayota Georgakopoulou. 2018. Evaluating MT for massive open online courses - A multifaceted comparison between PBSMT and NMT systems. *Mach. Transl.*, 32(3):255–278.

Michael Cronin. 2003. *Translation and Globalization*, first edition. Routledge, London, UK.

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20, Boston, MA. Association for Machine Translation in the Americas.

Jack Grieve. 2007. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal. Association for Computational Linguistics.

Yanfang Jia, Michael Carl, and Xiangling Wang. 2019. Post-editing neural machine translation versus phrase-based machine translation for english–chinese. *Machine Translation*, 33:1–21.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: a case study on english to croatian. *Machine Translation*, 32(3):195–215.

Maarit Koponen, Leena K. Salmi, and Markku Nikulin. 2019. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Mach. Transl.*, 33(1-2):61–90.

Ágnes Lesznyák. 2019. Hungarian translators' perceptions of neural machine translation in the European commission. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 16–22, Dublin, Ireland. European Association for Machine Translation.

Arle Lommel and Alan Melby. 2018. Tutorial: MQM-DQF: A good marriage (translation quality for the 21st century). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, Boston, MA. Association for Machine Translation in the Americas.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Anastasia Parianou. 2009. *Translating from major into minor languages*. Diavlos Publishing Co., Athens, Greece.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(1):4381.

Maja Popovic. 2017. Comparing language related issues for nmt and pbmt between german and english. *The Prague Bulletin of Mathematical Linguistics*, 108:209 – 220.

Maja Popovic. 2020. On the differences between human translations. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 365–374, Lisboa, Portugal. European Association for Machine Translation.

Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merriënboer, Kyunghyun Cho, and Yoshua Bengio. 2014. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 78–85, Doha, Qatar. Association for Computational Linguistics.

György Rado. 1987. A typology of lld translation problems. *Babel*, 33(1):6–13.

John Read. 2000. *Assessing vocabulary*. Cambridge University Press, Cambridge.

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of nmt and pbsmt. *Machine Translation*, 32(3):217–235.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yoonjin Song. 1991. Remarks on cultural transfer from an lld. volume 4, pages 63–79.

Vilelmini Sosoni and Maria Stasimioti. 2019. Mt output and post-editing effort: Insights from a comparative analysis of smt and nmt output and implications for training. In *Fit-For-Market Translator and Interpreter Training in a Digital Age (Language and Linguistics)*, pages 151–176. Vernon Press.

Maria Stasimioti and Vilelmini Sosoni. 2020. Translation vs post-editing of NMT output: Insights from the English-Greek language pair. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 109–124, Virtual. Association for Machine Translation in the Americas.

Karolina Stefaniak. 2020. Evaluating the usefulness of neural machine translation for the Polish translators in the European commission. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 263–269, Lisboa, Portugal. European Association for Machine Translation.

Antonio Toral and Víctor M. Sánchez-Cartagena. 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: Volume 1, Long Papers, pages 1063–1073, Valencia, Spain. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Haiyan Wu, Zhiqiang Zhang, and Qingfeng Wu. 2021. Exploring syntactic and semantic features for authorship attribution. *Applied Soft Computing*, 111:107815.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. volume abs/1609.08144. arXiv.