The Influence of a Machine Translation System on Sentiment Levels

Jaroslav Reichel and Ľubomír Benko

Constantine the Philosopher University in Nitra
94901 Nitra, Slovakia
{jreichel, lbenko}@ukf.sk

Abstract. The aim of the paper is to verify the influence of the used machine translation system on the level of sentiment in the translated text from Slovak to English using the available systems Google Translate and DeepL. The experiment was carried out on a parallel corpus created from subtitles of movies of different styles. The raw parallel corpus contained subtitles in Slovak and English. IBM Watson Natural Language Understanding service was used to identify the sentiment in the subtitles of ten movies of different genres. The paper also describes the methodology of preparing the dataset suitable for sentiment analysis using the IBM NLU service. The research showed a high correlation between human text and machine translation of subtitles for both translation systems. The research results show a high level of consistency of sentiment levels in both forms of translation. Based on the results obtained, the results of sentiment in machine translation can be generalized for the two most widely used translation systems.

Keywords: Machine translation, Natural language processing, Sentiment analysis, Slovak language

1 Introduction

The quality of machine translation depends on many factors. The text has many characteristics that need to be preserved in translation. However, there are also properties such as gender [1] or other regional formal habits that are not transferable between languages. A human translator can transfer some but machine translation has a problem with them.

An important characteristic of the text is the sentiment and the related emotion that the text should evoke. This is especially important in artistic texts such as poetry, prose, or film scripts. Emotion can also be captured in a text by observing an actor's performance. Sentiment and emotion can be identified in different ways. One of the most widely used is a tool from IBM that uses the IBM Watson supercomputer. For this purpose, the IBM Tone Analyzer tool was a service launched as an application programming interface (API) by IBM Corporation [2]. The IBM Watson™ Tone Analyzer service will be completely shut down in 2023 and is currently being replaced by the IBM Watson™

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 201–208, 2022.

Natural Language Understanding service on IBM Cloud as part of IBM's service offerings. In the area of Natural Language Processing, researchers [3] compared DialogFlow, LUIS, and Watson, where Watson performed the best. IBM Watson Natural Language Understanding (IBM NLU) is a machine learning system that uses linguistic models to break free text into important words and phrases, called keywords. The program then calculates a general sentiment score for each keyword [4]. The *sentiment_score* variable indicates the sentiment measure, which takes values from -1 to 1.

Sentiment analysis is the field of studying and analyzing people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions [5]. Many researchers are investing a lot of energy in developing a sentiment analysis tool for different languages. This analysis has to take into account the creation of a large dictionary, the use of artificial intelligence, and so on. If it would be possible to use an already established tool, such as IBM Watson NLU service, for the Slovak language, these resources could be used more efficiently. The problem of using resource-rich languages [6,7] (typically English) for text identification in low-resource languages is dealt with in the area of cross-lingual sentiment analysis or classification [8].

The aim of this paper is to compare whether the results of sentiment analysis in machine translation differ between the two most widely used tools. For the analysis, it is necessary to obtain a parallel corpus of the Slovak language texts and the corresponding English human-written text. For this purpose, in this research movie subtitles will be used. However, this research will not use community-created subtitles but professional subtitles from a streaming service. These are high-quality human translations, which are also used, for example in English language teaching [9].

For the human translations, machine translations were obtained using the two most widely used online machine translation systems, Google Translate and DeepL. Their outputs were compared in the analysis. Using the IBM Watson™ Natural Language Understanding service were identified the sentiments of each segment in different versions of the translations: human-written text (EN), Google Translate machine translation from Slovak to English (GT), and DeepL machine translation from Slovak to English (DL).

The structure of the paper is as follows. The second section contains related work in the field of sentiment analysis. The third section describes the experimental setup, used dataset and applied research methodology. The subsequent section focuses on the research results based on the sentiment analysis and evaluation of the research problem. The fourth section offers a discussion of the results.

2 Related work

Sentiment analysis can be considered as a sub-field of information extraction [10]. Several commercial systems exist for sentiment analysis as Amazon Web Services Amazon Comprehend, Dandelion Sentiment Analysis API,

Google Cloud Platform Natural Language API, IBM Watson Natural Language Understanding, Lexalytics Semantria API, MeaningCloud Sentiment Analysis API, Microsoft Azure Text Analytics, ParallelDots Sentiment Analysis, Repustate Sentiment Analysis, Text2data Sentiment Analysis API, TheySay PreCeive API or twin word Sentiment Analysis API [11]. IBM Watson NLU is one such system, which provides sentiment analysis scores with great accuracy based on the information presented to it [12] and that was the reason it was used in the research. IBM Watson has been used by researchers in sentiment analysis often for different types of reviews [4] or social media posts [13].

Kapusta et al. [14] aimed to explore the influence of sentiment analysis on fake news identification. The most important finding was that there are statistically significant differences in the article sentiment where the fake news articles were identified with more negative sentiment. The authors used a basic sentiment classification method. Evaluating the assessment of the truthfulness of a text and its sentiment has also been addressed by Reichel et al. [15].

The scientific field that deals with sentiment analysis using multiple languages and machine translation is called Cross-lingual sentiment analysis (CLSA). CLSA leverages one or several source languages to help the low-resource languages perform sentiment analysis tasks. The models used in the CLSA methodology can be significantly refined if it is possible to find the best range of source languages for a given target language. The authors see the limitation mainly in the wrong models and data available for some languages, such as the Slovak language [16]. This area has been addressed by researchers for different languages or combinations of languages.

A comparative study [17] verifies sentiment classification from Chinese texts (reviews) using sentiment analysis in English. The authors compare sentiment from human translation and machine translation of publicly available services such as Google Translate, Yahoo Babel Fish, and Windows Live Translate.

3 Experimental setup

We addressed the following research question in the experiment:

RQ: Is there a significant difference between the translation systems Google Translate and DeepL in the accuracy of identifying sentiment scores compared to human texts?

We can infer the null hypothesis from it.

H0: There is no statistically significant difference between the correlation of sentiment level in human text and in Google Translate machine translation compared to the correlation of sentiment level in human text and in DeepL machine translation.

Through an experiment, we measure how much the sentiment score in the source text and the sentiment score in the machine translation from Google Translate match. In the same way, we will evaluate the level of agreement in sentiment between the original text and the machine translation from the DeepL system. We then compare these values. If the sentiment rates match across

translators, this will mean that the results of further analyses of sentiment in machine translation can be generalized to all common translation systems. A more detailed description of the research process is given in the following steps:

1. Data preparation
   (a) Source corpus preparation
      i. Alignment of Slovak and English subtitles into a coherent parallel corpus.
      ii. Removal of erroneous, inconsistent, repetitive, or unnecessary records.
      iii. Segmentation - Merging sentences that have been split to multiple subtitles back into a single segment.
   (b) Generating a machine translation for each of the subtitles using Google Translate and DeepL machine translation systems.
   (c) Identification of keywords and their sentiment using IBM Watson NLU service.
   (d) Transforming the sentiment of the keywords into a coherent dataset of sentiment scores of each segment for the three sets:
      i. Human text (EN),
      ii. Machine translation from Google Translate (GT),
      iii. Machine translation from DeepL (DL).
2. Data analysis
   (a) Verification of the level of correlation of the identified sentiment of the machine translations (GT, DL) with the reference sentiment from the human text (EN).
   (b) Comparison of results from Google Translate and DeepL.
3. Verification and interpretation of results
   (a) Verification of research hypothesis H0.

Table 1: Sample of the dataset with subtitles and their machine translations from Google Translate and DeepL

id	Text_sk	Text_en	Text_gt	Text_dl
0	Blake.	Blake.	Blake.	Blake.
5	Zabalili nám jedlo?	Did they feed us?	Did they pack our food?	Did they pack us food?
6	Nie. Len poštu.	No. Just mail.	Not. Just mail.	No. Just mail.
7	Je čas na čaj!	Time for some tea! Tea's up!	It's tea time!	It's tea time!
8	Myrtle bude mať šteniatka.	Myrtle's having puppies.	Myrtle will have puppies.	Myrtle will have puppies.
10	Som hrozne hladný. Ty nie?	Oh, I'm bloody starving. Aren't you?	I'm terribly hungry. You do not?	I'm terribly hungry. Aren't you?

3.1 Machine translation generation

The corpus that was used contained 11 601 subtitles from 10 movies of different styles (war, fairy tale, action, sci-fi, comedy). The raw data had to be cleaned of erroneous entries, incorrectly paired parallel corpus pairs, and duplicate pairs. After resolving all errors, we obtained a parallel corpus of subtitles in English and Slovak. The created dataset contained 8551 segments.

To obtain the machine translation, the two most used online machine translation systems were chosen: Google Translate, and DeepL. Therefore, in the case of equality of results, it will be possible to generalize the results for machine translation obtained from different machine translation systems.

For further analysis, only the variables *id* (id of the segment), *Text_sk*, *Text_en*, and two variables *Text_gt* (machine translation obtained from Google Translate) and *Text_dl* (machine translation obtained from DeepL) were needed (Table 1).

3.2 Sentiment analysis

A tool IBM NLU was used for sentiment analysis. Each segment's sentiment analysis resulted in the identification of keywords and the determination of their sentiment. These results were transformed from JSON format into 3 matrices for each translation group (EN, GT, DL). There were 3 files with identified keywords for each segment and an associated *sentiment_score* value (Table 2). The output matrix from IBM NLU contains an identified *sentiment_score* for each keyword but for the same sentence (*id*) they match. Thus this is the *sentiment_score* of the sentence not of the keyword itself.

Table 2: Sample output from IBM NLU in the form of a matrix

id	keyword	text_en	sentiment_score
5	food	Did they feed us?	0
6	mail	No. Just mail.	0
7	tea time	Time for some tea! Tea's up!	0.842084
8	Myrtle	Myrtle's having puppies.	0.849348
8	puppies	Myrtle's having puppies.	0.849348
12	priesthood	It was the only reason I decided against the priesthood.	-0.839655

The results from the analysis using IBM NLU showed that there are approximately 18% fewer keywords in the machine translation compared to the human text (*Text_en* – 8419, *Text_gt* – 6886, *Text_dl* – 6923). After combining all three categories based on identified/unidentified sentiment, 4076 segments were extracted. These records were further manually cleaned of erroneous unpaired segments from sentences split into multiple subtitles. The resulting dataset contained 3768 segments.

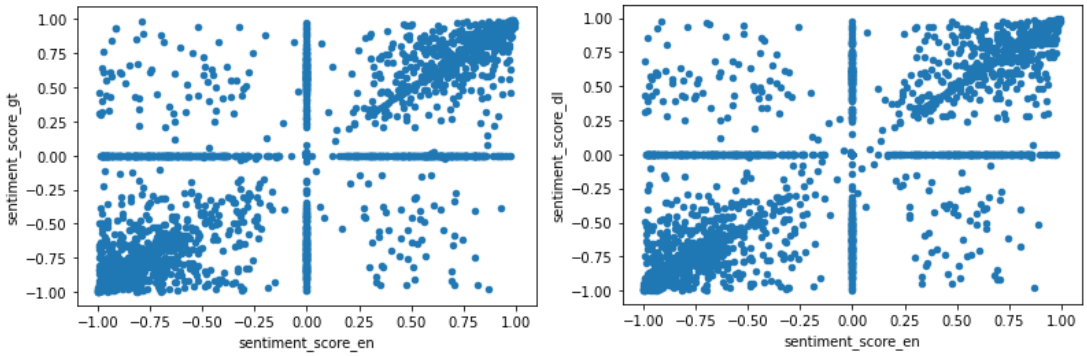

Fig. 1: 2D scatterplots for the correlation of the variable *sentiment_score_en* and a) *sentiment_score_gt*, b) *sentiment_score_dl*.

3.3 Results

RQ aims to verify whether there is a difference between Google Translate and DeepL in the results obtained. H0 predicts that there is no statistically significant difference between the correlation of sentiment level in human text and in Google Translate machine translation compared to the correlation of sentiment level in human text and in DeepL machine translation. Correlation analysis was used to verify the dependence. Correlation analysis verifies, in a simplistic way, that if sentiment is high in human text, it is also high in machine translation and vice versa. To determine the correct method for correlation analysis, the distribution of each group of EN, GT and DL was verified. The *sentiment_score* variable does not have a normal distribution. This is confirmed by the results of the Kolmogorov-Smirnov test for all 3 variables: *sentiment_score_en* ($D(3768) = 0.256, p < 0.01$), *sentiment_score_gt* ($D(3768) = 0.276, p < 0.01$) and *sentiment_score_dl* ($D(3768) = 0.272, p < 0.01$). Since enough cases are available, the parametric method can be used: Pearson's correlation coefficient was used. The calculation was performed at a 5% significance level.

Results of correlation analysis (Fig. 1):

- EN/GT: $r(3768) = 0.73, p < 0.01$,
- EN/DL: $r(3768) = 0.74, p < 0.01$.

From the graph (Fig. 1), quite a large number of pairs contain the value 0 in at least one of the variables. This means that sentiment has not been identified in any of the translations (of the pair). If we exclude these segments from the analysis in order to mainly evaluate the match in identified sentiment, then the results look like the following (Fig. 2):

- EN/GT: $r(1497) = 0.86, p < 0.01$,
- EN/DL: $r(1539) = 0.86, p < 0.01$.

Considering the correlation results between the human text and the machine translations (0.73 and 0.74 in the unadjusted dataset (Fig. 1); 0.86 and 0.86 in the

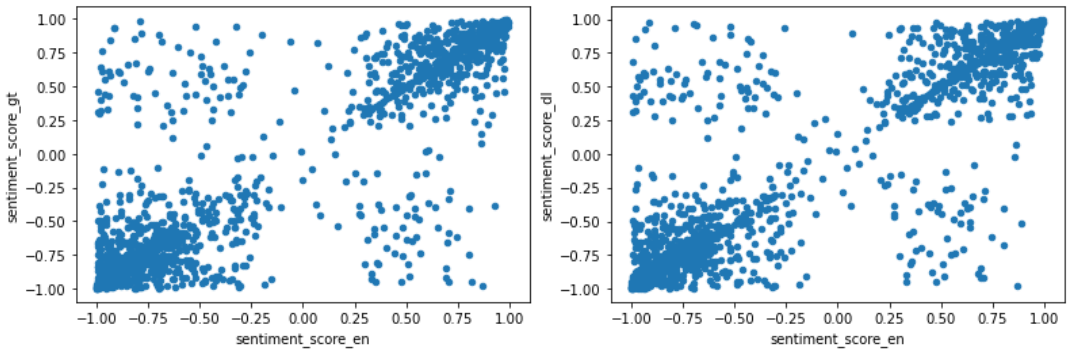

Fig. 2: 2D scatterplots for the correlation of the variable *sentiment_score_en* and a) *sentiment_score_gt*, b) *sentiment_score_dl* except for segments with neutral sentiment.

adjusted dataset (Fig. 2)), it can be argued that there is no significant difference between them. H0 is thus not rejected. Hence, the results of Google Translate are significantly similar to the results of DeepL and therefore it is relevant to use only one of these systems in further analysis. Based on the rejection of H0, these results can be generalized.

4 Conclusion

We tested whether there is a significant difference in sentiment analysis for texts translated by Google Translate and DeepL. The results say that there is no difference. This means that for further sentiment analysis in machine translation, it is not necessary to do duplicate analyses for multiple translation systems but just choose which one suits better based on text style. The results of the research can be generalized for machine translation from Slovak to English.

By evaluating RQ, it was verified that there is no significant difference in sentiment transfer in machine translation between the most widely used machine translation systems, i.e. Google Translate and DeepL. It is therefore possible to generalize the results for machine translations in general.

Acknowledgement This work was supported by the Slovak Research and Development Agency under the contract No. APVV-18-0473 and by the projects UGA VII/20/2022 and UGA VII/1/2022.

References

1. Rabinovich, E., Mirkin, S., Patel, R. N., Specia, L., Wintner, S.: Personalized machine translation: Preserving original author traits In: 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, vol. 2, pp. 1074–1084 (2017), doi: 10.18653/v1/e17-1101.

2. Al Marouf, A., Hossain, R., Kabir Rasel Sarker, Md. R., Pandey, B., Tanvir Siddiquee, S. Md.: Recognizing Language and Emotional Tone from Music Lyrics using IBM Watson Tone Analyzer. In: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Feb. 2019, pp. 1–6. doi: 10.1109/ICECCT.2019.8869008.
3. Canonico, M., de Russis, L.: A comparison and critique of natural language understanding tools. In: https://thinkmind.org/download.php?articleid=cloud_computing_2018_6_20_20057, last accessed 2022/06/23.
4. Lu, T. J., Nguyen, A. X. L., Trinh, X. V., Wu, A. Y.: Sentiment Analysis Surrounding Blepharoplasty in Online Health Forums. In: Plastic and Reconstructive Surgery - Global Open, vol. 10, no. 3, Mar. 2022, doi: 10.1097/GOX.0000000000004213.
5. Liu, B.: Sentiment analysis and opinion mining. In: Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–184, 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
6. Ghafoor, A. et al.: The Impact of Translating Resource-Rich Datasets to Low-Resource Languages through Multi-Lingual Text Processing. In: IEEE Access, vol. 9, pp. 124478–124490, 2021, doi: 10.1109/ACCESS.2021.3110285.
7. Baisa, V.: Czech Grammar Agreement Dataset for Evaluation of Language Models. In: Horak, A., Rychly, P., Rambousek, A. (eds.) RASLAN 2016, pp. 63-67. Karlova Studanka, Czech republic (2016).
8. Zhou, X., Wan, X., Xiao, J.: Cross-Lingual Sentiment Classification with Bilingual Document Representation Learning. In: Association for Computational Linguistics, pp. 1403–1412 (2016)
9. Dizon, G., Learning, B.: Language learning with Netflix: Exploring the effects of dual subtitles on vocabulary learning and listening comprehension. In: Computer Assisted Language Learning Electronic Journal. no. 3, pp. 52–65 (2021)
10. Sazzed, S., Jayarathna, S.: A sentiment classification in bengali and machine translated english corpus. In: Proceedings - 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science, IRI 2019, pp. 107–114 (2019) doi: 10.1109/IRI.2019.00029.
11. Ermakova, T., Henke, M., Fabian, B.: Commercial Sentiment Analysis Solutions: A Comparative Study. In: 17th International Conference on Web Information Systems and Technologies. pp. 103–114. (2021) doi: 10.5220/0010709400003058.
12. Dash, A. S., Pathare, S. B.: Survey of Sentiment Analysis Through Machine Learning for Forecasting Indian Stock Market Trend. In: SSRN Electronic Journal. (2022) doi: 10.2139/SSRN.4057274.
13. Daneshfar, Z., Asokan-Ajitha, A., Sharma, P., Malik, A.: Work-from-home (WFH) during COVID-19 pandemic – A netnographic investigation using Twitter data. In: Information Technology & People, (2022), doi: 10.1108/ITP-01-2021-0020.
14. Kapusta, J., Benko, Ľ., Munk, M.: Fake News Identification Based on Sentiment and Frequency Analysis. pp. 400–409, 2020, doi: 10.1007/978-3-030-36778-7_44.
15. Reichel, J., Magdin, M., Benko, Ľ., Koprda, Š.: Can Fake News Evoke a Positive/Negative Affect (Emotion)? In: DIVAI 2020, pp. 563–572. (2020)
16. Xu, Y., Cao, H., Du, W., Wang, W.: A Survey of Cross-lingual Sentiment Analysis: Methodologies, Models and Evaluations. In: Data Science and Engineering, vol. 1, p. 3, (2022), doi: 10.1007/s41019-022-00187-3.
17. Wan, X.: A comparative study of cross-lingual sentiment classification. In: Proceedings - 2012 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2012, pp. 24–31. (2012). doi: 10.1109/WI-IAT.2012.54.