Keyword Extraction for Automatic Evaluation of Machine Translation

Lívia Kelebercová and František Forgáč

Department of Informatics, Faculty of Natural Sciences and Informatics,
Constantine the Philosopher University in Nitra,
Trieda Andreja Hlinku 1, 94974 Nitra,
livia.kelebercova@ukf.sk

Abstract. Evaluation plays a key role in the field of machine translation. In general, the evaluation of machine translation can be divided into two types, manual (human) and automatic (machine). Professional human translators can understand and evaluate the text with the best results in terms of measuring quality and analyzing errors but on the other hand, this approach brings a number of disadvantages, including high time consumption, the subjectivity of the translator, and the finan-cial costs associated with hiring professional translators. Automatic evaluation approaches are usually based on the correlation between the sentences or n-grams from human translation and machine translation. The aim of this paper is to capture the semantics of human translation from the English language to the Slovak language and the same text translated by ETransL and Deepl translating engines by extracting the keywords which represent the main phrases from doc-uments to determine how much the machine translations differ from the reference human translation. Based on our results the translations are equal from a seman-tic point of view and the end user should understand the text translated by ETransL and Deepl equally as human translation.

Keywords: Keyword Extraction, Machine Translation, Evaluation of Machine Translation

1 Introduction

Machine translation evaluation is necessary to discover how closely the neural translation language model relates to the reference domain. This process is essential for determining the effectiveness of an existing model and estimating the amount of post-processing the model needs to fulfill the expectations of end-users.

Evaluation approaches can be basically divided into manual approaches and automatic approaches. Manual approaches use professional human translators to evaluate key metrics such as adequacy and fluency scores. The main problem with the manual approach is that evaluation is based on subjective human judgment and this process is time-consuming [1].

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 47–54, 2022. © Tribun EU 2022

1.1 Related Work

To solve the problems with human evaluation researchers developed automatic approaches. These approaches are trying to evaluate how close the machine translation is to one or more human references by using metrics as BLEU [2], NIST [4], or METEOR [5]. These metrics usually score individual segments which are usually sentences and the main concept behind them is that the closer a machine translation is to a reference translation, the better it is [1,2].

Despite the fact that BLEU [2] is the most common metric to evaluate the quality of machine translation, the reliability is questionable. Based on Babych's research, the main problem is that blue and many other commonly used metrics measure lexical identity at the surface level but they are insensitive to linguistic variations [5,6]. Some metrics such as METEOR try to solve this problem by including semantic tools like WordNet lexical database to reduce the dependence on exact matches of words in sentences. WordNet-based approaches also have their disadvantages and may not be able to fully describe word similarity between MT out-puts and references [6]. An interesting way was presented by Mirsarraf and Deghani [7] who, inspired by Lo's [8] research, and like him, proposed a depend-ency-inspired semantic evaluation methodology to quantify how well the underlying meaning of the source is maintained in the translated output using dependency analysis concepts in SRL. Several researchers have attempted to include semantics in machine translation evaluation but neither of them was trying to include keyword extraction in evaluation metrics.

1.2 Proposed Method

In this paper, we focused on finding similarities between machine translation and human translation in terms of text observation rather than in the context of appropriateness and adequacy for each word/phrase/sentence. To capture the meaning of the text we used keyword extraction. Keyword extraction is used to identify the most important phrases from the document [4,5]. From a linguistic point of view, if the machine translation is equal to human translation, then applying a keyword extraction algorithm should give the same keywords for each text. If we don't get the same keywords, it means that the translations are probably different. The Slovak language is one of the inflected languages, and therefore the extracted keywords may differ in endings, which means that there are machine translation errors, but only from the grammar point of view. For this reason, we used a higher level of granularity and determined the base of the word and the root of the word. According to the lemma, we were able to determine the part of speech of the root of the word (stem). There are two reasons why we determined stem. The first reason is that if the stem was the same for the extracted keywords, it could have resulted in the change of part of speech. The second case represents that the meaning is preserved, but the resulting form is incorrect and there is a fluency error. Formally speaking, the neural machine translation model mistakenly transfers the abstract representation of the source word to the abstract representation of the target language and a shift in meaning occurs, or the language model correctly transforms the abstract representation of the word into the target language, but in the target language, it erroneously creates the external form of the given word.

The rest of our paper is organized as follows. In section 2 we describe the keyword extraction process and data preparation for the next step. Section 3 presents the evaluation process and results and in Section 4 we we evaluate the proposed method.

2 Methods

The keyword extraction subsection is intended to introduce you to the function of the chosen approach for extracting key phrases. The subsection data preparation serves to describe the further processing of data for the purpose of subsequent evaluation

2.1 Keyword Extraction

In the case of estimating the quality of machine translation, we used text translated from English to the Slovak language by a human as a reference translation and the same text translated by ETransL [9] and Deepl [10] machine translators. To improve the process of Keyword Extraction it was necessary to preprocess our text data. At first, we converted our data to lowercase. Then we removed tags and special characters. The last step was to identify stop words. For this purpose, we used a library that contains 418 Slovak stop words.

Keyword Extraction is a summarization technique, which uses statistical information from the text to identify the most important phrases [12,15]. In this case, we used a Rapid Keyword Extraction Algorithm (RAKE) [13].

The main concept behind the RAKE algorithm is that keywords are often consisting of multiple words without any interpunction or stop words. The algorithm is based on collocation and co-occurrence, which means the goal is to find words that are frequently occurring together in desired n-gram range [13].

For implementation, we used a rake python package [14]. We created a function that accepts a list of stop words, preprocessed text, and n-gram range as parameters to initialize the RAKE algorithm. The first step of the algorithm is to split the text into words and place them in the word degree matrix. We can imagine this matrix as an Excel table, where every word is placed in separate cells horizontally and vertically. Then each word is assigned a score presenting how frequently a given word co-occurs with another word [13].

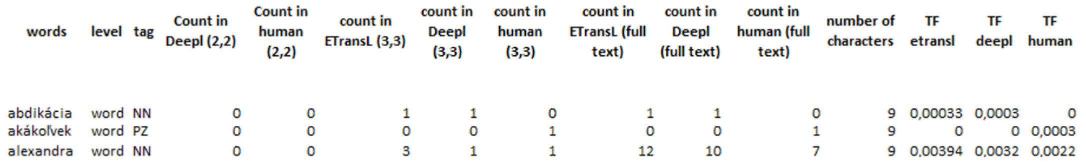
The next step is to calculate the degree of the word in the matrix, which presents the sum of the number of co-occurrences divided by the frequency (how many times a word occurs in the corpus). The final score for keywords in desired n-gram range is then calculated as a sum of degrees of words of its words [13]. We called this function on each translation to obtain the top 100 keywords consisting of two words and the top 100 keywords consisting of three words for each text separately.

2.2 Data preparation

From the Keyword Extraction process, we obtained 100 key phrases consisting of two words and 100 key phrases consisting of three words for human translation, ETransL and Deepl translation. We took all 600 keywords and split them into single words and removed duplicates. We put them in a python data frame column labeled as the word. First, we needed to capture the morphological properties of the words. We used the Stanza [15] library for lemmatization, and we adapted the code of Czech stemmer [16] to obtain the stems in the Slovak language. Lemmas and the stems were then manually controlled and corrected. We placed our lemmas and bases of words in the first column under the original words from keywords and we created a column labeled as level describing whether the word is the original form of the word, lemma, or the stem. The next variable is presented in the tag column. To obtain tags we used MorphoDiTa [17] tagger with a Slovak model.

We calculated the number of times each word occurred in extracted bigrams and trigrams from each translation. These frequencies are presented in columns count in ETransL (2, 2), count in Deepl (2, 2), count in human (2, 2), count in ETransL (3, 3), count in Deepl (3, 3), count in human (3, 3). We also calculated the count of occurrence of each word in the whole translation, so we created columns with labels count in ETransL (full text), count in Deepl (full text), and count in human (full text).

Another value we wanted to capture is the length of the word which is presented in the number of characters in the word column. In the end, we calculated term frequency – the frequency of a word in text divided by the number of words in a document. These values are presented in columns TF etransl, TF deepl, and TF human. We exported our python data frame to an Excel sheet. The First three rows from our sheet are visible on Figure 1.

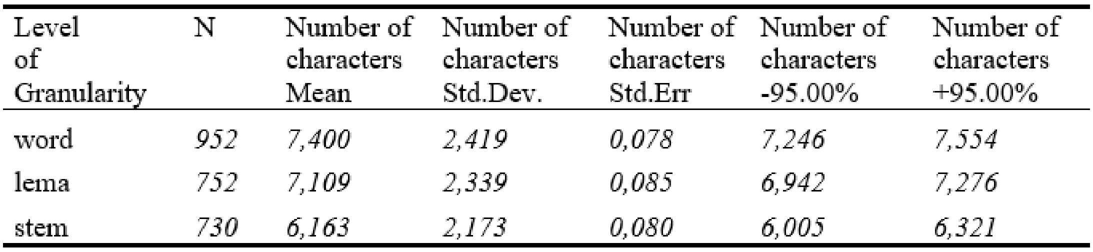


words	level	tag	Count in Deepl (2,2)	Count in human (2,2)	count in ETransL (3,3)	count in Deepl (3,3)	count in human (3,3)	count in ETransL (full text)	count in Deepl (full text)	count in human (full text)	number of characters	TF etransl	TF deepl	TF human
abdikácia	word	NN	0	0	1	1	0	1	1	0	9	0,00033	0,0003	0
akákoľvek	word	PZ	0	0	0	0	1	0	0	1	9	0	0	0,0003
alexandra	word	NN	0	0	3	1	1	12	10	7	9	0,00394	0,0032	0,0022

Fig. 1: First three row from Excel sheet

3 Results

Figure 2 shows the point and interval estimation of the mean in the number of characters. It is logical that the number of characters decreases with the level of granularity, and we have to compare frequencies for each granularity separately.



Level of Granularity	N	Number of characters Mean	Number of characters Std.Dev.	Number of characters Std.Err	Number of characters -95.00%	Number of characters +95.00%
word	*952*	*7,400*	*2,419*	*0,078*	*7,246*	*7,554*
lema	*752*	*7,109*	*2,339*	*0,085*	*6,942*	*7,276*
stem	*730*	*6,163*	*2,173*	*0,080*	*6,005*	*6,321*

Fig. 2: Point and interval estimation of the mean in the number of characters

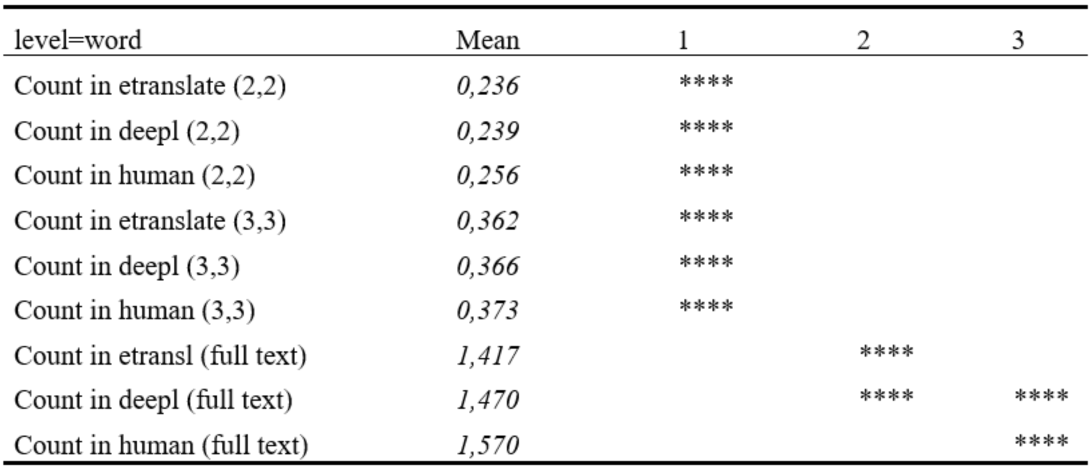
To verify the effectiveness of the proposed models, we used modified tests for repeated measurements (Greenhouse-Geisser adjustment), due to the violation of the sphericity condition of the covariance matrix. If the condition of sphericity of the covariance matrix is not met, the size of the error of the first type increases. Epsilon represents the degree of violation of the sphericity condition. Epsilon equal to one represents the fulfillment of the condition. Conversely, the smaller it is, the more the condition of sphericity is violated. In our case, the Epsilon values were significantly less than one (word: G-G Epsilon = *0,24*, p < *0,001*; lemma: G-G Epsilon < *0,175*, p < *0,001*; stem: G-G Epsilon = *0,190*, p < *0,001*). Null hypotheses with *99,9%* reliability (at the *0,001* significance level) are rejected, which claim that there is no statistically significant difference in word/lemma/stem frequencies in bigrams, trigrams, and whole texts of the ETransL, Deepl, and human translations. Hypotheses were tested at individual levels (word/lemma/stem).

The surrounding of the word is important, it makes a difference whether we compare the frequency of occurrence of the keyword in bigrams, trigrams, or within the entire text. The given keyword/lemma/stem was found in different frequencies, either in the wider area or in the shorter. That's why we compared the translations multiple times to find out between which of them there are statistically significant differences and vice versa between which are not.

From the point of view of multiple comparisons, we identified three homogeneous groups (**** - p > *0,05*) in frequencies at the word level and two at the lemma/stem level. Naturally, a statistically significant difference was demonstrated between the frequencies in whole texts and in bigrams/trigrams. If we look at the frequency separately for bigrams, trigrams, and whole texts, there are no statistically significant differences between the translations machine an human translations except for frequency at the word level, where a statistically significant difference between ETransL and human was demonstrated (p < *0,05*).

Although the averages are low, as some keywords occurred just once, as the context of the phrase expands, the frequency of occurrence increases relative to the word/lemma/stem.

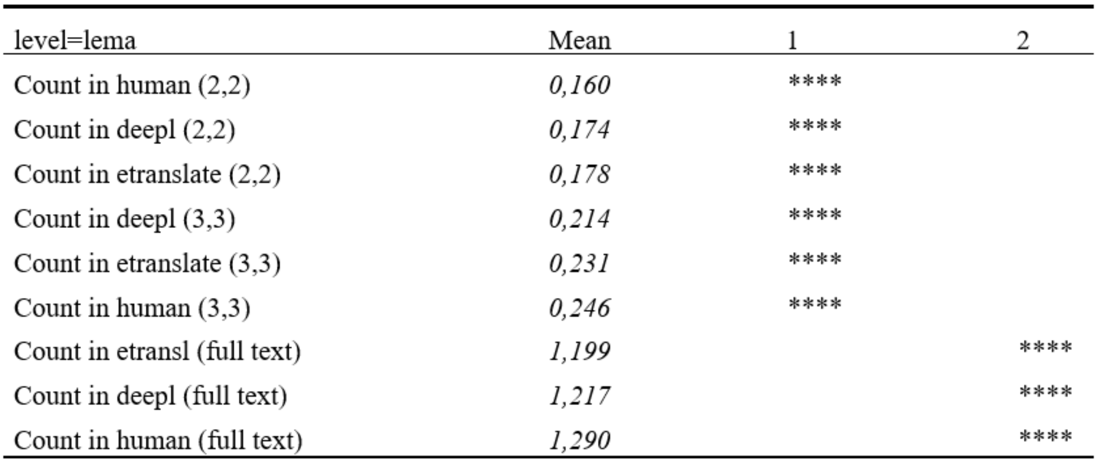
Figure 3 shows that there is no statistically significant difference between human translation and machine translation if we only consider bigrams and trigrams. If we consider the whole text, there is a statistically significant difference in the frequency of occurrence of the keyword between the whole text and the phrases, regardless of whether it is human or machine translation.



level=word	Mean	1	2	3
Count in etranslate (2,2)	*0,236*	****		
Count in deepl (2,2)	*0,239*	****		
Count in human (2,2)	*0,256*	****		
Count in etranslate (3,3)	*0,362*	****		
Count in deepl (3,3)	*0,366*	****		
Count in human (3,3)	*0,373*	****		
Count in etransl (full text)	*1,417*		****	
Count in deepl (full text)	*1,470*		****	****
Count in human (full text)	*1,570*			****

Fig. 3: Frequencies on word level

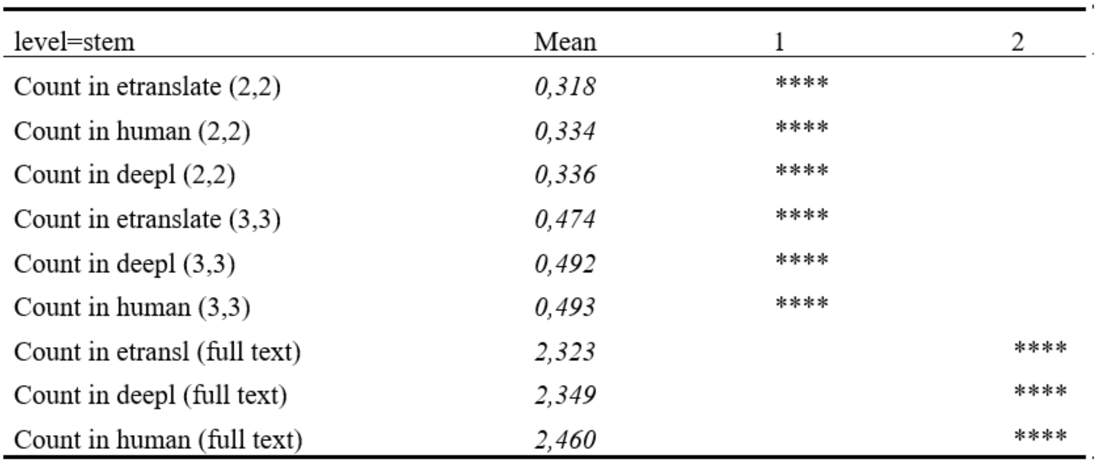
Interestingly if we take a closer look at the whole text, there is no statistically significant difference between the human translation and Deepl in the frequency of keywords, from which we can conclude that both translations reach the same level in understanding the text (they captured the same keywords, i.e. meaning and even their form, i.e. fluidity). A statistically significant difference was demonstrated between ETransL and human translation, and here we can discuss whether the inaccuracy occurred only in the form of the word (i.e. in the ending, fluency) or also in the lemma or at the root of the word (in meaning, i.e. accuracy).



level=lema	Mean	1	2
Count in human (2,2)	*0,160*	****	
Count in deepl (2,2)	*0,174*	****	
Count in etranslate (2,2)	*0,178*	****	
Count in deepl (3,3)	*0,214*	****	
Count in etranslate (3,3)	*0,231*	****	
Count in human (3,3)	*0,246*	****	
Count in etransl (full text)	*1,199*		****
Count in deepl (full text)	*1,217*		****
Count in human (full text)	*1,290*		****

Fig. 4: Frequencies on lemma level

If we look at Figure 4, we can see that due to the lemma, there is no longer a statistically significant difference, probably a grammatical problem, not a semantic one, that occurs with ETransL. This is also confirmed by Figure 5.



level=stem	Mean	1	2
Count in etranslate (2,2)	*0,318*	****	
Count in human (2,2)	*0,334*	****	
Count in deepl (2,2)	*0,336*	****	
Count in etranslate (3,3)	*0,474*	****	
Count in deepl (3,3)	*0,492*	****	
Count in human (3,3)	*0,493*	****	
Count in etransl (full text)	*2,323*		****
Count in deepl (full text)	*2,349*		****
Count in human (full text)	*2,460*		****

Fig. 5: Frequencies on stem level

4 Conclusion

Our research indicates that due to the extracted keywords, or their frequency throughout the text is no difference between machine translations and human translation. Considering the quality of the translation from a semantic point of view, the translations are equal and the end user/reader should understand the text equally and receive the same information. However, due to the form of the given information, a difference between human translation and ETransL was demonstrated in favor of human translation, i.e. keyword was more common in human translation than in ETransL.

The proposed method of determining translation quality differs from existing approaches in the sense that we were not concerned with determining the quality of machine translation in the context of fluency and adequacy for each word/phrase/sentence, but rather with determining the similarity of machine translation and human translation from the point of view of text observation. In our case, it was journalistic texts whose function is to inform the reader and to get an answer to the questions Who? What? When? Where? and how? Basically, we were concerned with the applicability of machine translation in the given context or domain. Through our research, we have shown that DeepL is usable and very similar to human translation in keywords, starting from phrases first, then words, lemmas, and stems.

Acknowledgments This work was supported by the Slovak Research and Development Agency under contract No. APVV-18- 0473.

References

1. Sepesy Maučec, M., Donaj, G.: Machine translation and the evaluation of its quality. Recent Trends in Computational Intelligence. (2020).
2. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. (2001).
3. Doddington, G.: Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In Proceedings of the second international conference on Human Language Technology Research -. (2002).
4. Denkowski, M., Lavie, A.: Meteor Universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation. (2014).
5. Babych, B., Hartley, A: Sensitivity of Automated MT Evaluation Metrics on Higher Quality MT Output: BLEU vs Task-Based Evaluation Methods. In The Sixth International Language Resources and Evaluation (LREC'08) (2008).
6. Wong, B.: Semantic Evaluation of Machine Translation. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (2010)
7. Mirsarraf, M.R., Dehghani, N.: A dependency-inspired semantic evaluation of machine translation systems. Lecture Notes in Computer Science. 71–74 (2013).
8. Lo, C., Wu, D.: MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In Proc. of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 220–229 (2011)
9. Chribonn: ETRANSL, https://www.alanbonnici.com/2022/03/etransl.html.
10. Kutylowski, J.: Deepl, https://www.deepl.com/en/publisher/ (2017)
11. Kelebercová, L., Munk, M.: Analysis of the Popularity Rate of Extracted Keywords From True and Fake News Related to Covid-19. In International Scientific Conference on Distance Learning in Applied Informatics. pp. 384–392. Wolters Kluwer, Prague, Czech Republic (2022).
12. Kelebercová, L., Munk, M.: Search queries related to COVID-19 based on keyword ex-traction. In Procedia Computer Science. pp. 2618–2627 (2022).
13. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text Mining. 1–20 (2010).
14. Sharma, V.: Rake nltk, https://csurfer.github.io/rake-nltk/, (2021).
15. Qi, P., Zhang, Y., Bolton, J., Manning, C.: A Python Natural Language Processing Toolkit for Many Human Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp 101–108 (2020)
16. Gomes, L.: Czech Stemmer https://research.variancia.com/czech_stemmer/ (2010)
17. Straková J., Straka, M., Hajič, J. : Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 13-18. Baltimore, Maryland (2014)