Analysis of Edit Operations for Post-editing Systems

Jozef Kapusta¹ · Ľubomír Benko¹ · Dasa Munkova¹ · Michal Munk^{1,2}

Received: 21 June 2021 / Accepted: 10 November 2021
© The Author(s) 2021

Abstract

Post-editing has become an important part not only of translation research but also in the global translation industry. While computer-aided translation tools, such as translation memories, are considered to be part of a translator's work, lately, machine translation (MT) systems have also been accepted by human translators. However, many human translators are still adopting the changes brought by translation technologies to the translation industry. This paper introduces a novel approach for seeking suitable pairs of n -grams when recommending n -grams (corresponding n -grams between MT and post-edited MT) based on the type of text (manual or administrative) and MT system used for machine translation. A tool that recommends and speeds up the correction of MT was developed to help the post-editors with their work. It is based on the analysis of words with the same lemmas and analysis of n -gram recommendations. These recommendations are extracted from sequence patterns of the mismatched words (MisMatch) between MT output and post-edited MT output. The paper aims to show the usage of morphological analysis for recommending the post-edit operations. It describes the usage of mismatched words in the n -gram recommendations for the post-edited MT output. The contribution consists of the methodology for seeking suitable pairs of words, n -grams and additionally the importance of taking into account metadata (the type of the text and/or style and MT system) when recommending post-edited operations.

Keywords Machine translation · Natural language processing · Recommender tool · Post-editing · POS tagging · Sequence rules

1 Introduction

Tasks that were exclusively based on human thinking and intelligence are gradually starting to be operated by machines. Nowadays, post-editing (PE) constitutes a significant element of the translation industry [1]. Its situation and quality can be improved with the improvement of

technologies and machine translation (MT) systems. One of the main components of the translation process is the time the translator spends “behind the keyboard”. It is a time that involves active writing in the target language, i.e. translating and rewriting text from the source language into the target language. The satisfaction of growing demands for translation is complicated for translators, thus they are starting to facilitate their work by using computer-assisted translation (CAT) tools and MT systems [2].

The possibility to escape from the routine and from dull translations, and the ability to use technology while translating technical or administrative documents undoubtedly increases translation productivity in multilingual societies—societies opting for rapid and high-quality translation services [3–5].

The core of CAT tools lies in translation memory (TM); the program which stores parallel text aligned into segments (e.g. popular Trados, Memsource or MemoQ). A segment does not necessarily equal a sentence as a unit, sometimes it is just a phrase. The parallel text consists of the original and its translation usually aligned 1 to 1 (but it could be also

✉ Jozef Kapusta
jkapusta@ukf.sk

Ľubomír Benko
lbenko@ukf.sk

Dasa Munkova
dmunkova@ukf.sk

Michal Munk
mmunk@ukf.sk

¹ Department of Computer Science, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, Nitra 949 74, Slovakia

² Science and Research Centre, University of Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic

aligned 1:2, 1:0 or vice versa). The parallel texts saved in TM may be later used by the translator in similar translations, fully (100% match) or partially (fuzzy match), depending on the context of the project. TM is used as a translation support tool for decreasing the need for re-translating segments of text that have already been translated [6]. In the case of extensive text and time-consuming translation, TMs and terminological databases are highly appreciated.

The issue arises when an exact match is not available. Then it is possible to use fuzzy-match repair method to correct the proposal of the translation. Similarly, the idea of TM can be applied for parallel post-edited MT (PEMT) texts during the PE [7, 8].

An interactive online system (OSTPERE—Online System for Translation, Post-editing, Revision and Evaluation) was created by [9, 10] and used not only for translation and revision such as Trados or Memsource, but also for manual evaluation of MT output. It offers an interface for effective PE, i.e. a logged-in user can correct a translation and can also determine basic MT errors. It stores parallel texts of the source language, their MT outputs and also their PEMTs (post-edited by human translators) in the target language. At the time of writing, the online interactive system contained around 62,000 sentence segments (around 570,000 MT words and 720,000 PEMT words).

The OSTPERE system works with translations from English, German, French, and Russian into Slovak. The OSTPERE system, unlike CAT tools, also has tools for manual evaluation. Besides the PE itself, the post-editor has an opportunity to evaluate the quality of MT output on a scale from 1 to 5 with respect to the adequacy and fluency of machine translation, and also to determine the extent of MT errors (critical, major, and minor) in the sphere of language, accuracy, terminology, and style. A semi-automatic metric HTER (Human-mediated Translation Error Rate) was also implemented in the system, which is based on the lexical similarity of MT and PEMT.

HTER is based on the Levenshtein distance [11], i.e. the distance between two segments is determined as the minimum number of characters/words that we must change (insert, substitute, delete or shift) in one segment to match with the other corresponding segment (reference/PEMT). Likewise, the basic techniques of natural language processing (NLP) such as segmentation, tokenization, alignment and part-of-speech (POS) tagging were used to compare segments and/or to determine the lexical similarity.

Analysis of MT outputs into the inflectional Slovak language [8, 12–14] proved that the most MT errors are related to the incorrect form of the word, e.g. wrong gender, number or case. For this reason, a tool was developed to help and increase post-editor productivity, i.e. it recommends and speeds up the corrections of MT output. The tool is a part

of the OSTPERE system. The tool is based on the assumption that most of the work done by post-editors is connected with corrections of word form. The tool uses knowledge obtained from post-edit memory (PEM). Knowledge-based recommender systems give recommendations based on the explicit specification of the kind of content that is required or needed [15]. PEM consists of parallel segments of MTs and their PEMTs. MT output was obtained using two different MT engines: Google Translate (GT) and MT@EC. For this study, two types of texts were examined, i.e. MTs and subsequently PEMTs: news (press texts) and manuals (technical documentation).

The paper is aimed at the three following objectives: (1) It describes an approach how PEM, containing parallel MT output–PEMT output, can be used during PE, i.e. it deals with how the information from PEM can be used to recommend the corrections within the PE.; (2) It describes the algorithm of sequence analysis for n -gram recommendation; (3) It analyzes whether it is necessary to include metadata into the recommendation mechanism (information concerning the used MT engine and text type), i.e., whether information about MT or style of source text can help to improve the recommendations for post-editors.

The contribution of the study corresponds to our stated objectives. The theoretical contribution of research to the subject area lies in a new approach to the creation of PEM and to the evaluation of MT quality through sequence analysis and fuzzy matches for n -gram recommendation. It introduces the use of PEM recommendations for post-editors. The second, practical contribution consists of procedures to obtain the corresponding n -grams between MT output and PEMT output and of verifying the impact of metadata on n -gram recommendation. Adding metadata to the recommendation system seems to be a possible improvement of the n -gram recommendation system.

The paper is structured as follows: section 2 introduces a related work focusing on PE and fuzzy-match repair. Section 3 describes the materials and methods to achieve the aim of this paper. Section 4 summarizes the results of experimental verification of the appropriateness of metadata included in created post-editing systems with a recommender tool. Section 5 offers a discussion of the obtained results and subsequently, the conclusion is offered in the last section.

2 Related Work

DePalma [1] notes that similarly to TMs in the 1990s, nowadays PE has become an important part not only of translation research but also of the global translation industry. When PE, the translator does not rewrite the

whole text, he/she makes only the necessary corrections in the MT output. According to Daems et al. [16], the PE effort can be assessed through product or process analysis, i.e. by comparing MT output to a reference translation or PEMT output, or by observing aspects of the PE process e.g. time/durations or pauses. Krings [17] distinguishes three aspects of PE effort: temporal effort (time to turn the MT output into a high-quality translation, the easiest to define and measure), technical effort (physical actions—edit operations, in PE including deletions, insertions, substitutions, reordering), and cognitive effort (mental processes and cognitive load during PE). Steiert and Mariniello [18] pointed out that the types of translations in which MT has truly been successful are technical. Plitt and Masselot [19] carried out productivity tests to find out whether PE or human translation is more productive. They used Moses, an open engine for MT and PE, they created their own system which recorded time for decision-making and editing for each post-edited MT sentence. The results have shown that MT saved 43% of a translator's working time (PE performance was higher than traditional human translating).

Green et al. [20] show that POS counts (nouns and adjectives) and language complexity, in case of English as a source language, influence time of translation.

Khan et al. [21] dealt with a morphologically rich language, they examined and evaluated the effectiveness of machine learning and deep learning for morphologically rich language part-of-speech tagging.

However, the idea of a recommender system in the field of MT is not new, still there is a lack of recommender systems for PE, which can help the post-editors with their work and additionally increase their productivity. Such a system should offer a recommended translation (segment) for a post-editor. Espla-Gomis et al. [22] proposed an approach to CAT systems users to identify the target word in the translation that needed to be changed or kept unedited. Integration of MT into a TM system allows efficient possibilities for human translators. Such a recommender system is offered by Dara et al. [23]. He et al. [24] proposed a translation recommender system which integrates statistical MT output within TM systems. He et al. [25] conducted a user study with professional post-editors to integrate the MT output with TM systems. This study is connected to their previous study focusing on an evaluation of the effectiveness of translation recommendation. The results support the validation of the previous method [24] using automatic evaluation metrics to approximate the PE effort. Bonet [26] worked on a case study-oriented on a recommender system for MT quality estimation. The study dealt with the issue of MT quality estimation by designing a classifier based on the source text's linguistic features. The proposed model used a manually annotated

linguistic features from the source text which were used to create a recommender system that classified sentences for PE or translation. Magdin et al. [27] used multiple regression analysis based on the text input of users and recommended their emotional states. Aranberri and Pascual [28] focused on a specific language pair for translation: Spanish-Basque. They found, that the language pair has poor support for MT. They examined the possibility to create a comprehensive recommender system to speed up the decision time between PE and translation from scratch. The results provided PE effort indicators and also a trained classification model that would recommend translators which editing approach to take into consideration. The results showed that comprising the PE effort indicators helped to improve the poor performance of the baseline models. Kagita et al. [29] deal with a novel approach to a recommender system. They have used conformal prediction which is a recent approach in machine learning for reliable prediction. A conformal predictor is guaranteed that the error probability is limited by a predetermined confidence level. They introduced a Conformal Recommender System which they evaluated against 12 state-of-art recommender algorithms on different datasets. The conformal prediction could offer a different approach to deal with recommendations. Escartin and Arcedillo [30] experimented to test a fuzzy score given to an MT output to get an alternative evaluation metric of the MT. They compared the fuzzy score with other metrics, such as TER or BLEU.

Ortega et al. [31] extended their previous approach [32] which used an external source of bilingual information to repair fuzzy matches from TM. Ortega et al. [32] introduced a novel approach to fuzzy-match repair using any MT system called patching. Patching means to offer accurate translation proposals to CAT tool users. It can be used to repair and improve fuzzy-match proposal from a TM. Ortega et al. [31] improved and updated a fuzzy-match repair algorithm which is used to generate a set of fuzzy-match repaired segments from a translation unit and the source language segment to be translated by using any source of bilingual information. The methodology consists of building a list of patching operators that are used in the fuzzy-match repair process. Following the list of patching operators is an algorithm that explores all possible combinations of patching operators to generate the set of candidate fuzzy-match repair segments. They focused on Spanish and examined the effectiveness of their approach to generating a set of all possible fuzzy-match repair target segments. They have chosen the best translations and evaluated them. This was done on a sandbox environment and in a real setting, the best fuzzy-match repaired segment would have to be chosen using another method (e.g. using MT quality estimation methods). Carrasco et al. [33]

used a fuzzy linguistic approach in research to obtain qualitative information.

Knowles et al. [34] presented one of the most novel studies of fuzzy-match repair and evaluated a rule-based (Apertium), phrase-based statistical (Moses), and neural (Nematus) MT systems as black-box sources of bilingual information. As in previous research [31, 32], they have chosen translation direction from English into Spanish. The evaluation of the MT systems was done using the BLEU and WER metrics. The results showed that the neural MT system performs better on fuzzy-match repair on matches than other MT systems and also the neural MT is more often successful at repairing the segments. However, the phrase-based statistical MT produced almost similar results with a slightly worse measures.

3 Materials and Methods

3.1 Corrections for PE System

The OSTPERE system was created in the PHP framework CodeIgniter. The data is stored in the MariaDB database. The system is held on Sun Fire x4170 server with 16 Intel(R) Xeon(R) E5540 processors clocked at 2.53 GHz. The system is equipped with 72 GB or RAM and runs on Debian operating system.

For the analysis and subsequent recommendation of corrections for PE, it is necessary to conduct the basic steps of data preparation using NLP methods. The first step is text segmentation. This is a basic NLP task consisting of splitting a text into a set of segments. This step is important because most of the NLP processes are done at the sentence level. To create the correct segment pairs (MT segment with corresponding PEMT segment) it is necessary to perform segment alignment using HunAlign [35] and subsequently to perform tokenization and morphological text annotation.

It is a fundamental and common linguistic method that can be found in corpus linguistics or NLP [36]. It comprises the grammatical and morphological features of a word in context [37]. MT and PEMT segments were annotated using the TreeTagger tool [38–40], which resulted in tokens (words) with specific morphological annotations (POS tags) and lemmas.

The individual processes of segment preparation are shown in the Fig. 1. The result of the whole process is a segment pair (mostly segment=sentence) consisting of an MT segment and its corresponding PEMT segment. Segment pairs prepared in this way can be used:

- for MT errors analysis and
- for a recommendation system for PE.

Fig. 1 Workflow of segment preparation processes

In our study, we focus on the usage of these segments for recommendations for PE. The recommended corrections can serve the post-editor by making their work more efficient.

It was necessary to extract the required knowledge from the existing PEM to create the recommendation of tokens. The created tool for recommendation is based on the selection of all tokens that may be recommended for PE the token(s). The tool was created regarding the inflectional language type where the most MT errors are in the incorrect word form (errors of gender, number, and case). The first step was a selection of the token (word) pairs for the recommended correction, i.e. the original token—the token correction.

The used PEM contained 46,681 segments of the source texts. These segments were consequently translated by MT engines and post-edited. Using tokenization 454,528 tokens were created from the segments.

Later, two sets for each segment (MT and PEMT) were created; the first set T consists of tokens from MT segments together with corresponding tags and lemmas; the second set P consists of tokens from PEMT segments together with corresponding tags and lemmas.

Afterwards, an automatic parser for further analysis was created which analyses the pairs of segments (MT sentences with corresponding PEMT sentences). The pairs of tokens

were extracted from the matches of segments as follows (Fig. 2):

- (i) (WTL)—100% match—in the first step the same word in both sets (MT segment and PENT segment) was chosen together with its tag and lemma (the used TM contained 52.47% matches of WTL type);
- (ii) (WT)—fuzzy match—the pair with the same word and tags were chosen from both remaining sets (the used TM contained 0.1% matches of WT type);
- (iii) (W)—fuzzy match—the pair only with the same word (the used TM contained 4.38% matches of W type);
- (iv) (L)—fuzzy match—the pairs with the same lemma were chosen (the used TM contained 12.07% matches of L type);
- (v) (MisMatch)—remaining tokens (not chosen or mismatched) after the analysis in both sets (the used TM contained 30.97% matches of MisMatch type).

The parser results were interesting for the categories (WT) and (W), i.e. groups of words with the same word and morphological tags but with different lemma or with different lemma and also tag (the used TM contained 4.48% matches of these types). A more detailed analysis of these groups revealed these as parser errors and errors of the tool used for morphological tagging. Mostly it was wrong assigned tags by foreign words, e.g. word HDMI mentioned in Fig. 2.

3.2 Recommendations Based on the Same Lemma (L)

Category (L) was used in the further analysis, i.e. tokens that had a similar lemma but words and morphological tags were different. Precisely these tokens captured the corrections in TM typical for a fleective language and reflected the MT error correction of gender, number, and case. Based on token pairs, a recommendation matrix was created for the category (L).

Based on the parser results, the category (L) contains tokens pairs $t_i \rightarrow p_i$; where $t_i \in T$ (tokens from MT segments) and $p_i \in P$ (tokens from post-edited MT segments)

where tokens t_i and p_i have the same lemma but different words and tags.

Let M be $n \times m$ matrix for the recommendation (Fig. 3):

- matrix element $m_{i,j}$ equals the occurrence of the tokens pair $t_i \rightarrow p_i$ for category (L),
- n is the count of all tokens from the set T ,
- m is the count of all tokens from the set P .

Zero values were omitted in the case of the token pairs of the category (L). The matrix is saved in a table, in which the tokens of set T represent columns and tokens of set P represent rows (Fig. 3).

The tool for recommendation carries out tokenization for each segment. Each token (word) from the segment is then searched in the matrix M (column) and based on the occurrence it recommends the user suitable tokens for correction [16].

3.3 n -Gram Recommendations from Mismatch Category

The new challenge was to cover the words that were not matched based on tags and lemmas into the recommendation, i.e. words from the group (MisMatch). In terms of text processing, it does not recommend a word or sequence of words but n -grams, and also the recommender tool can recommend a different n -gram (where $n \geq 1$) for a given n -gram.

In our recommendation system, it was a relatively simple task to find a recommendation for tokens with the same lemma such as (Fig. 2) recommendations for correcting the token 'zvukové' ('audio', adjective in plural and nominative case) to token 'zvukového' ('audio', adjective in singular and genitive case) or the token 'signály' ('signals', noun in singular and genitive case) to token 'signálu' ('signal', noun in singular and genitive cases). These recommendations are aimed at correcting the unigram to unigram.

A more complex problem was to find suitable recommendations for correcting n -grams such as (Fig. 2) a

Fig. 2 Categorization of segments matching

recommendation for correcting tokens ‘v’ (‘in’) and ‘jednom’ (‘one’) into only one token ‘jediný’ (‘the only one’).

Sequence rule analysis was used to find suitable n -grams pairs from the category (MisMatch). The condition was to respect the order, i.e. the order of the tokens from the category (MisMatch) was assigned based on the tokens’ order in the sentence. Also, the condition that the tokens from set T had to be before tokens from set P .

The condition of respecting the order of tokens was very important. Without this condition, it would be possible to use other analysis such as association analysis, but in NLP, considering the word order in a sentence is vital in finding language or linguistic patterns. For such applications, association rules are no longer appropriate and sequential patterns are required.

In the case of n -gram recommendations, sequential rules will help us to find the most frequent n -gram corrections, i.e. from all the corrections of tokens from the MisMatch category, sequential rules select the most numerous patterns of corrections, while the order of tokens within the segments will be kept (the order of the words within the sentence).

Implementation of the A priori algorithm was used for the extraction of the sequence rules [41]. After the extraction of the sequence rules, it was necessary to select rules that contain only tokens from set T in the conditional part and in the action part rules containing tokens from set P . We obtained the rules for recommending n -grams using the group (MisMatch).

4 Results

PEMT, from which the pairs of tokens for the recommendation (edit operations for post-editors) were created, contains two types of texts: journalistic and technical texts (news and manuals). Both texts were translated by GT engine and MT@EC engine. During the development of the tool for correction, it was interesting to know how metadata—the type of texts and different MT engines—can influence the recommendation. If the importance of both parameters (text type and MT engine) for the recommendation was recorded; the recommender matrix and also the rules of the MisMatch group for the created tool would have to take into account the importance of both parameters.

The following assumptions were stated:

- (i) It is expected that MT output translated by GT, will have a significant impact on the quantity of extracted rules (the number of edit operations needed to transform MT output into required translation quality-into PEMT).
- (ii) It is expected that MT output translated by GT, will have a significant impact on decreasing the portion

of rules (the number of edit operations needed to transform MT output into required translation quality- into PEMT).

- (iii) It is expected that the style/type of text, will have a significant impact on the quantity of extracted rules (the number of edit operations needed to transform MT output into required translation quality- into PEMT).

The abovementioned assumptions, as well as hypothesis, will be tested for both categories, category (L) and category (MisMatch).

4.1 Results of the n -gram Recommendation of MisMatch

The analysis (Table 2) resulted in sequence rules, which were obtained from frequent sequences fulfilling their minimum of variable support (min support = 0.5%). Frequent sequences were obtained from identified sequences, i.e. segments (sequences of tokens) of MT and PEMT output, distinguishing MT system (GT or MT@EC) (Table 1).

There is a high coincidence between the results (Table 2) of sequence rule analysis in terms of the portion of rules found in the segments of MT output translated by GT and MT@EC, and subsequently post-edited by human translators, where 1 means the rule was found in the examined segments translated by examined MT system and vice versa, 0 means the rule was not found in the examined segments. The most rules were extracted from segments translated by MT@EC; specifically, 186 rules which represent almost 56%. From the segments translated by GT, 159 rules were extracted which represents over 47%. Based on the Q test results, the zero hypotheses, reasoning that the incidence of rules (represents a number of edit operations needed to transform MT output into required translation quality) does not depend on MT system, is not rejected (Table 2).

Generally, if the style of MT output is not distinguished, GT (fewer rules) is a bit better but not statistically significant, the number of operations (required in terms of post-editing) expressed by the rules is roughly the same (Table 3).

The analysis (Table 4) resulted in sequence rules, which were obtained from frequent sequences fulfilling their

Table 1 Number of tokens, identified sequences and frequent sequences of tokens for examined MT systems of source text for MisMatch category

	GT	MT@EC
Number of tokens	5112	6929
Number of identified sequences (segments)	622	674
Number of frequent sequences	458	619

Fig. 3 Sample of the matrix M

$$M = \begin{bmatrix} \dots & \text{signál} & \text{signálov} & \text{signálu} & \text{signály} & \dots & \text{zvuková} & \text{zvukové} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & 18 & 12 & 29 & \dots & 0 & 0 & \dots \text{signál} \\ \dots & 24 & 0 & 14 & 21 & \dots & 0 & 0 & \dots \text{signálov} \\ \dots & 12 & 4 & 0 & 4 & \dots & 0 & 0 & \dots \text{signálu} \\ \dots & 8 & 21 & 7 & 0 & \dots & 0 & 0 & \dots \text{signály} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & 0 & 0 & 0 & \dots & 0 & 17 & \dots \text{zvuková} \\ \dots & 0 & 0 & 0 & 0 & \dots & 12 & 0 & \dots \text{zvukové} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$
Table 2 Incidence of extracted sequence rules in examined segments considering MT system for MisMatch category

Body	= >	Head	GT	MT@EC
(otvoriť (MisMatch_MT))	= >	(Otvorenie (MisMatch_PE))	1	1
⋮	= >	⋮	⋮	⋮
(pre (MisMatch_MT)), (pripojenie (MisMatch_MT))	= >	(pripojte (MisMatch_PE))	1	0
⋮	= >	⋮	⋮	⋮
(TV (MisMatch_MT))	= >	(Televízor (MisMatch_PE))	1	1
Count of derived sequence rules			159	186
Percent of derived sequence rules (Percent 1's)			47.46	55.52
Percent 0's			52.54	44.48
Cochran Q test			$Q = 2.2431$; $df = 1$; $p < 0.1342$	

Table 3 Number of tokens, identified sequences and frequent sequences of tokens for examined MT systems and styles/types of the source text for MisMatch category

	GT news	GT manual	MT@EC news	MT@EC manual
Number of tokens	1705	3407	1999	4930
Number of identified sequences (segments)	156	466	160	514
Number of frequent sequences	122	181	121	107

Table 4 Incidence of extracted sequence rules in examined texts considering MT system and a style/type of source text for MisMatch category

Body	= >	Head	GT news	GT manual	MT@EC news	MT@EC Manual
(hladovaniu (MisMatch_MT))	= >	(hladuje (MisMatch_PE))	0	0	1	0
⋮	= >	⋮	⋮	⋮	⋮	⋮
(Press (MisMatch_M), (opakovane (MisMatch_M))	= >	(opätovným (MisMatch_PE)), (stláčáním (MisMatch_PE))	0	1	0	0
⋮	= >	⋮	⋮	⋮	⋮	⋮
(menu (MisMatch_MT))	= >	(ponuky (MisMatch_PE))	0	0	0	1
Count of derived sequence rules			7	88	7	30
Percent of derived sequence rules (Percent 1's)			5.43	68.22	5.43	23.26
Percent 0's			94.57	31.78	94.57	76.74
Cochran Q test			$Q = 134.9538$; $df = 3$; $p < 0.0000$			

minimum of variable support (min support = 1.5%). Frequent sequences were obtained from identified sequences, i.e. segments (sequences of tokens) of MT and PEMT output, distinguishing MT system (GT or MT@EC) and style/type of source text (news or manuals) (Table 3).

There is a coincidence between the results (Table 4) of sequence rule analysis in terms of the portion of rules found in the case of press texts translated by GT and MT@EC. The most rules were extracted from technical texts translated by GT; specifically, 88 rules which represent over 68%. On the other hand, only 30 rules were extracted from technical texts translated by MT@EC which represents over 23%.

Based on the Q test results of the zero hypotheses, which reasons that the incidence of rules (expressing required edit operations in terms of post-editing) does not depend on MT system and style/type of source text, is rejected at the 0.1% significance level (Table 4).

Kendall's coefficient of concordance represents the degree of concordance in the number of rules found among examined texts for examined MT systems and types of the source text. The value of the coefficient (Table 5) is approximately 0.35 while 1 means a perfect concordance and 0 represents a discordance. Low values of the coefficient confirm the Q test results (Table 6).

From multiple comparisons (Tukey test) three homogeneous groups (Table 5) were identified regarding the average incidence of found rules. Statistically significant differences were proved at the 5% significance level in the average incidence of rules found between GT and MT@EC in a case of technical texts and also between technical and press texts regardless of used MT system. On the other hand, the statistically significant difference was not proved between GT and MT@EC in the case of press texts.

Generally, if the styles/types of text were distinguished, it would be found out that there is no difference in the press texts (news), but in case of manuals, there is a statistically significant difference in the number of edit operations (in terms of post-editing) expressed by the rules in favour of MT@EC.

Looking at the results in more detail, Table 6 shows that:

Table 5 Homogeneous groups for incidence of extracted sequence rules in examined texts considering MT system and a style/type of source text for MisMatch category

	Incidence mean	1	2	3
MT@EC_news	0.0543	****		
GT_news	0.0543	****		
MT@EC_manual	0.2326		****	
GT_manual	0.6822			****
Kendall's Coefficient of Concordance		0.34872		

Table 6 Crosstabulation: GT_manual × MT@EC_manual for MisMatch category

GT_manual\MT@EC_manual	0	1	Σ
0	14 10.85%	27 20.93%	41 31.78%
1	85 65.89%	3 2.33%	88 68.22%
Σ	99 76.74%	30 23.26%	129 100.00%
Phi-square	−0.6883		
McNemar (A/D)	<i>Chi-square</i> = 5.8824; <i>df</i> = 1; <i>p</i> = 0.0153		
McNemar (B/C)	<i>Chi-square</i> = 29.0089; <i>df</i> = 1; <i>p</i> = 0.0000		

- almost 66% of the new rules were found in technical texts translated by GT (GT_manual) and
- only 21% of the new rules were found in technical texts translated by MT@EC (MT@EC_manual).

In the case of technical texts translated by GT (Table 6: McNemar (B/C)), it is 85 new rules. The statistically significant difference was proved in the number of rules found between the used MT systems in favour of texts translated by MT@EC (fewer edit operations required to transform MT output into the text of high quality, in terms of post-editing).

The condition (validity assumption) of the use of the chi-square test is sufficiently high expected frequencies. The condition is violated if the expected frequencies are lower than 5. In our case, the assumption is not violated (Table 7).

The graph (Fig. 4) visualizes interaction frequencies—GT_manual × MT@EC_manual. In this case, the curves did not copy mutually; they had a different course—which only proved the results of the analysis (Table 6).

4.2 Results of the Recommendation Based on Similar Lemma (Category L)

Similar results were obtained also for the category (L), where the match was only for lemma.

Table 7 Expected frequencies: GT_manual × MT@EC_manual for MisMatch category

G_manual\ MT@EC_manual	0	1	Σ
0	31.4651	9.5349	41.0000
1	67.5349	20.4651	88.0000
Σ	99.0000	30.0000	129.0000

Fig. 4 Interaction Plot: GT_manual × MT@EC_manual for Mis-Match category

Table 8 Incidence of extracted sequence rules in examined segments considering MT system for category L

Body	= >	Head	GT	MT@EC
(nastavenie (L_MT))	= >	(nastavenia (L_PE))	1	0
⋮	= >	⋮	⋮	⋮
(diaľkové (L_MT))	= >	(diaľkovom (L_PE))	0	1
⋮	= >	⋮	⋮	⋮
(klávesnicu (L_MT))	= >	(klávesnice (L_PE))	1	0
Percent of derived sequence rules (Percent 1's)			55.56	25.93
Percent 0's			44.44	74.07
Cochran <i>Q</i> test			$Q=5.3333$; $df=1$; $p<0.0209$	

The same methods were used to verify the assumptions as for the category (MisMatch). The rules were extracted on the same principle. The difference was only in the sequence length that was restricted on length 2. In the case that there is a match in lemma it is not necessary to analyze longer sequences as it was for the MisMatch category.

Based on the *Q* test results, the zero hypotheses, reasoning that the incidence of rules (representing the number of edit operations needed to transform MT output into required translation quality) does not depend on MT system, is rejected at the 5% significance level (Table 8).

In the case of category (L), if the style of MT output is not distinguished, MT@EC (fewer rules) is statistically significantly better.

Based on the *Q* test results the zero hypothesis, reasoning that the incidence of rules (expressing required edit operations in terms of post-editing) does not depend on MT system and style/type of source text, is rejected at the 5% significance level (Table 9).

The value of Kendall's coefficient of concordance (Table 10) is approximately 0.11 while 1 means a perfect concordance and 0 represents a discordance. Low values of the coefficient confirm the *Q* test results. Two homogeneous groups (Table 10) from multiple comparisons were identified regarding the average incidence of found rules. Statistically significant differences were proved at the 5% significance level in the average incidence of rules found between GT and MT@EC in a case of technical texts. On the other hand, the statistically significant difference was not proved between GT and MT@EC in the case of press texts.

In the case of category (L), if the styles/types of text were distinguished, it would be found out that there is no difference in the press texts (news), but in case of manuals, there is a statistically significant difference in the number of edit operations (in terms of post-editing) expressed by the rules in favour of MT@EC.

From this point of view, the results are similar to the one obtained for the category (MisMatch).

Table 9 Incidence of extracted sequence rules in examined texts considering MT system and a style/type of source text for category L

Body	= >	Head	GT news	GT manual	MT@EC news	MT@EC Manual
(inteligencie (L_MT))	= >	(inteligencia (L_PE))	1	0	0	0
⋮	= >	⋮	⋮	⋮	⋮	⋮
(k (L_MT))	= >	(ku (L_PE))	1	0	0	0
⋮	= >	⋮	⋮	⋮	⋮	⋮
(siete (L_MT))	= >	(Siet' (L_PE))	0	1	0	1
Percent of derived sequence rules (Percent 1's)			18.52	55.56	22.22	25.93
Percent 0's			81.48	44.44	77.78	74.07
Cochran <i>Q</i> test			$Q=8.6552$; $df=3$; $p<0.0343$			

Table 10 Homogeneous groups for incidence of extracted sequence rules in examined texts considering MT system and a style/type of source text for category L

	Incidence mean	1	2
GT_news	0.1852	****	
MT@EC_news	0.2222	****	
MT@EC_manual	0.2593	****	
GT_manual	0.5556		****
Kendall's coefficient of concordance		0.10685	

5 Discussion

Using parallel texts—MT output and their corresponding PEMT output can help post-editors to improve their work and/or increase their productivity and also save the time and labor. Our findings are in line with Vieira [42] and Bonet [26], who showed that linguistic features influence the PE effort and have an impact on translation quality. It is not interesting to examine 100% matches, the words that are matched 1-to-1 (source word to target word) but those words that are not 100% matches (fuzzy-matches or mismatches). Using a sequence analysis we did not examine 1-to-1 match, but n to n , where $n \geq 1$.

The main idea behind this study is to show the use of mismatched words in the n -grams recommendations for PE. Specifically, to investigate how the mismatch words can be paired and what are the possibilities of matching regarding MT system and style/type of the source text, i.e. when we have a given MT output (a specific type of text, translated by a certain MT system), what are the possibilities of its PE based on PEM.

In case of category (MisMatch), the first two assumptions (assumption 1 and assumption 2) in favour of GT have not been confirmed; there is no difference between the MT systems when the style/type of the source text is not distinguished. Besides, MT@EC is a significantly better system when the style/type of text is taken into consideration. In case of technical texts such as manuals, a fewer, statistically significant number of corrections (edit operations) are required during the PE in terms of found rules. The assumption concerning the style/type of the source text (assumption 3) was confirmed.

The results for category (L) correspond with the results obtained for the category (MisMatch) if the text style and MT engine are taken into consideration.

Manuals are specific texts in any languages. Typical attributes of manuals are documentary, monologue, public, conceptual, accuracy, clearness, expertise, official, absence of the addressee, and communicative function. Many of these attributes have also administrative texts. If we take common administrative texts such as forms, directives,

regulations, there are schematized or stereotypical sentences (e.g. we have received, we appeal/invite you, nonobservance of, default on, nonperformance, violation of), i.e. the same sentence constructions. In terms of lexicons, in both text types, technical terms frequently occur. In terms of the sentence structure, both text types are very similar, both are accurately structured, using numbering, structured into paragraphs, etc. MT@EC system (engine) is trained on administrative EU texts. The results of our research have shown that the domain (text database) is important for training the MT engine. A smaller range of specialized, parallel or comparable texts such as EU document (EU database) is more appropriate for MT engine than a large range of general texts (Google database).

6 Conclusion

The paper aims to introduce a new approach to the PEM creation and to the evaluation of MT quality. It is based on the identification of n -gram corrections using sequence rule analysis and fuzzy matches between MTs and PEMTs. It deals with improving the correction recommender tool for post-editors of MT output. The recommendation was created based on PEM. After tokenization of parallel texts from PEM, tokens were POS annotated and lemmas were identified. Precisely, lemma was essential to find a suitable recommendation for word correction (for wrong gender, number or case where the lemma is the same). Words/tokens from the category (MisMatch) and category (FuzzyMatch) were subjected to a detailed analysis that was done to improve the recommendation tool. Tokens from both categories (MisMatch and FuzzyMatch) represent a significant part of the text that is not always usable and here can be seen is the PE effort (productivity of the translator/post-editor). This motivated us to examine segments that are not identical 100% matches.

The last objective was to discover whether it is significant to include words based on metadata (the text type or used MT engine) into the recommendation tool.

The results showed that the created rules (edit operations) for the recommendation are disjunctive sets if the metadata are taken into consideration. For example, in the case of the technical text (text type) and MT engine (GT_manual $\times$ MT@EC_manual), only three identical operations were found (Table 6).

For this reason, it is necessary to include into our PE recommender tool, a recommendation based on metadata (the selection of the MT system and PEMT style/type). It will also be important to use multiple sets (for each MT system

and text style/type) for extracting edit operations to recommend n -grams for PE.

Acknowledgements This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and Slovak Academy of Sciences (SAS) under the contract No. VEGA-1/0792/21, also by the scientific research project of the Czech Sciences Foundation Grant No:19-15498S and by the Slovak Research and Development Agency under the contract no. APVV-18-0473.

Funding Grant name: vedecká grantová agentúra mšvvaš sr a sav, grant number: VEGA-1/0792/21, grant name: Michal Munk, agentúra na podporu výskumu a vývoja, grant number: APVV-18-0473, grant name: Michal Munk, grantová agentúra českej republiky, grant number: 19-15498S.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- De Palma, D., Pielmeier, H., Stewart, R. G., Henderson, S.: Common Sense Advisory's Annual Report, 2016. [Online]. Available: <http://www.commonsenseadvisory.com/AbstractView/tabid/74/ArticleID/36540/Title/TheLanguageServicesMarket2016/Default.aspx>
- Chérargui, M.A.: Theoretical overview of machine translation. In: Proceedings of ICWIT 2012, pp. 160–169 (2012) [Online]. Available <http://ceur-ws.org/Vol-867/Paper17.pdf>
- Wallis, J.: Interactive Translation vs. Pre-translation in the Context of Translation Memory Systems: Investigating the Effects of Translation Method on Productivity, Quality and Translator Satisfaction. University of Ottawa, Ottawa (2006)
- Webb, L.E.: Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis. Monterey Institute of International Studies, Monterey, CA (1998)
- Benis, M.: Translation Memory from O to R. ITI Bull. (1999) [Online]. Available <http://utkl.ff.cuni.cz/~rosen/VYUKA/MT/tm-review01.htm>
- Simard, M.: Translation spotting for translation memories. In: Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts Data Driven Machine Translation and Beyond, vol. 3, pp. 65–72, May 2003. <https://doi.org/10.3115/1118905.1118918>
- Munk, M., Munková, D.: Detecting errors in machine translation using residuals and metrics of automatic evaluation. J. Intell. Fuzzy Syst. **34**(5), 3211–3223 (2018). <https://doi.org/10.3233/JIFS-169504>
- Munk, M., Munková, D., Benko, L.: Identification of relevant and redundant automatic metrics for mt evaluation. In: Multi-disciplinary Trends in Artificial Intelligence (MIWAI 2016) Book Series: Lecture Notes in Computer Science, vol. 10053, pp. 141–152. Springer International Publishing, Cham (2016)
- Munková, D., Munk, M., Benko, L., Absolon, J.: From old fashioned 'one size fits all' to tailor made online training. Adv. Intell. Syst. Comput. **916**, 365–376 (2020). https://doi.org/10.1007/978-3-030-11932-4_35
- Benko, L., Munková, D.: Application of POS tagging in machine translation evaluation. In: DIVAI 2016 : 11th International Scientific Conference on Distance Learning in Applied Informatics, Sturovo, May 2–4, pp. 471–489 (2016)
- Snober, M.G., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: 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, August 8–12, pp. 223–231 (2006)
- Báňík, T., Benko, L., Máčková, R., Munk, M., Munková, D.: Wie irrt die Maschine? Probleme der maschinellen Übersetzung. Verlag Dr. Kovac, Hamburg (2019)
- Munk, M., Munková, D., Benko, L.: Towards the use of entropy as a measure for the reliability of automatic MT evaluation metrics. J. Intell. Fuzzy Syst. **34**(5), 3225–3233 (2018). <https://doi.org/10.3233/JIFS-169505>
- Munková, D. et al.: Mýliť sa je ľudské (ale aj strojové): analýza chýb strojového prekladu do slovenčiny. UKF, Nitra (2017)
- Aggarwal, C.C.: Recommender Systems: The Textbook. Springer International Publishing, Cham (2016)
- Daems, J., Vandepitte, S., Hartsuiker, T.J., Macken, L.: Identifying the machine translation error types with the greatest impact on post-editing effort. Front. Psychol. **8**, 1282 (2017). <https://doi.org/10.3389/fpsyg.2017.01282>
- Krings, H.P.: Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Process. Kent State Univ Pr (2001)
- Mariniello, E., Steiert, A.: The Human Role in a Machine-Translated World. TCworld (2016)
- Plitt, M., Masselot, F.: A productivity test of statistical machine translation post-editing in a typical localisation context. Prague Bull. Math. Linguist. **93**, 7–16 (2010)
- Green, S., Heer, J., Manning, C.D.: The efficacy of human post-editing for language translation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 439–448 (2013)
- Khan, W., et al.: Part of speech tagging in Urdu: comparison of machine and deep learning approaches. IEEE Access **7**, 38918–38936 (2019). <https://doi.org/10.1109/ACCESS.2019.2897327>
- Espla-Gomis, M., Sánchez-Martínez, F., Forcada, M.L.: Using machine translation in computer-aided translation to suggest the target-side words to change. In: Machine Translation Summit, pp. 172–179 (2011) [Online]. Available <https://rua.ua.es/dspace/bitstream/10045/27578/1/espla-gomis11b.pdf>
- Dara, A., Dandapat, S., Groves, D., van Genabith, J.: TMT-prime: a recommender system for MT and TM integration. In: Proceedings of the NAACL HLT 2013 Demonstration Session, pp. 10–13 (2013) [Online]. Available <http://www.aclweb.org/anthology/N13-3003>
- He, Y., Ma, Y., van Genabith, J., Way, A.: Bridging SMT and TM with translation recommendation. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 622–630 (2010) [Online]. Available <http://www.mt-archive.info/ACL-2010-He.pdf>

25. He, Y., Ma, Y., Roturier, J., Way, A., van Genabith, J.: Improving the post-editing experience using translation recommendation: a user study. In: Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers, October 31–November 4, 10 p (2010)
26. Bonet, O. G.: To post-edit or to translate... That is the question: A case study of a recommender system for Quality Estimation of Machine Translation based on linguistic features. Euskal Herriko Unibertsitatea (2018)
27. Magdin, M., Držík, D., Reichel, J., Koprda, S.: The possibilities of classification of emotional states based on user behavioral characteristics. *Int. J. Interact. Multimed. Artif. Intell.* **6**(4), 97–104 (2020). <https://doi.org/10.9781/ijimai.2020.11.010>
28. Aranberri, N., Pascual Saiz, J.A.: Towards a post-editing recommendation system for Spanish–Basque machine translation. In: Proceedings of the 21st Annual Conference of the European Association for Machine Translation, pp. 21–30 (2018)
29. Kagita, V.R., Pujari, A.K., Padmanabhan, V., Sahu, S.K., Kumar, V.: Conformal recommender system. *Inf. Sci. (NY)* **405**, 157–174 (2017). <https://doi.org/10.1016/j.ins.2017.04.005>
30. Parra Escartín, C., Arcedillo, M.: A fuzzier approach to machine translation evaluation: a pilot study on post-editing productivity and automated metrics in commercial settings. In: Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra), Dec. 2015, pp. 40–45. <https://doi.org/10.18653/v1/w15-4107>
31. Ortega, J.E., Sánchez-Martínez, F., Forcada, M.L.: Fuzzy-match repair using black-box machine translation systems: what can be expected? (2016). Accessed: Feb. 09, 2020. [Online]. Available <http://www.atril.com/software/dj-vu-x3-professional>
32. Ortega, J., Sánchez-Martínez, F., Forcada, M.: Using any machine translation source for fuzzy-match repair in a computer-aided translation setting (2014). <https://doi.org/10.13140/2.1.3306.1121>
33. Carrasco, R.A., Blasco, M.F., García-Madariaga, J., Herrera-Viedma, E.: A fuzzy linguistic RFM model applied to campaign management. *Int. J. Interact. Multimed. Artif. Intell.* **5**(4), 21–27 (2019). <https://doi.org/10.9781/ijimai.2018.03.003>
34. Knowles, R., Ortega, J.E., Koehn, P.: A comparison of machine translation paradigms for use in black-box fuzzy-Match Repair (2018). Accessed: Feb. 09, 2020. [Online]. Available <http://www.casmacat.eu/corpus/>
35. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. *Proc. RANLP* **2005**, 590–596 (2005)
36. Garside, R., Leech, G., Mcenery, A.M.: Corpus Annotation: Linguistic Information from Computer Text Corpora. Routledge, London (1997)
37. Forróová, M., Horák, A.: Morfológická anotácia textov. In: Slovenčina na začiatku 21. Storočia v Prešove, pp. 1–12 (2003)
38. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, pp. 44–49 (1994)
39. Schmid, H., Baroni, M., Zanchetta, E., Stein, A.: The Enriched TreeTagger System (2007)
40. “TreeTagger—A Language Independent Part-of-Speech Tagger.” <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>. Accessed Oct. 10 (2019)
41. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993). <https://doi.org/10.1145/170035.170072>
42. Vieira, L.N.: Indices of cognitive effort in machine translation post-editing. *Mach. Transl.* **28**(3–4), 187–216 (2014). <https://doi.org/10.1007/s10590-014-9156-x>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.