Evaluating Machine Translation for Cross-Lingual Fact-Checking

Irina Temnikova[1 [0000−−0002−−5601−−2540]], Silvia Gargova[1 [0009−−0009−−4178−−0158]], Tsvetelina Stefanova[1 [0000−−0001−−8079−−2965]], Iva Marinova[2], Ruslana Margova[1 [0000−−0001−−6243−−104X]], Nevena Grigorova[1], Alexander Komarov[1], Dan Sultanescu[3 [0000−0002−6439−3523]], and Kalina Bontcheva[1,4 [0000−−0001−−6152−−9600]]

[1] Big Data for Smart Society Institute (GATE), Sofia, Bulgaria
{irina.temnikova, silvia.gargova, ruslana.margova, nevena.grigorova, alexander.komarov, tsvetelina.stefanova}@gate-ai.eu
[2] Identrics, Sofia, Bulgaria
iva.marinova@identrics.ai
[3] National University of Political Studies and Public Administration, Bucharest, Romania
dan.sultanescu@snspa.ro
[4] University of Sheffield, Sheffield, United Kingdom
k.bontcheva@sheffield.ac.uk

Abstract. While cross-lingual manual and automatic fact-checking are important, and Machine Translation (MT) is among the methods, used for them, there are no orienting guidelines, nor evaluation metrics that could assist with determining which MT engine would be appropriate for such a task. This article presents an evaluation approach that fills a gap by providing a numerical estimate of Machine Translation (MT) engines' suitability for translating texts for cross-lingual claim matching and fact-checking. The approach focuses on elements important for the task, such as the correct translation of Named Entities (NEs), while making others less important (for example the style and the fluency of the translations). Our contributions include an MT error classification, evaluation guidelines, formulas to obtain a normalized numerical score, and a Python script for calculating it. The numerical weights of the score's components can be modified, which allows flexibility, reflecting what is possible for the subsequent stage. We also present the results of two experiments in which we apply our approach (with a choice of weights) and determine the most suitable freely accessible MT tools for translating Bulgarian and Romanian news articles and social media texts into English. Our results show that eTranslation is the best MT tool for Romanian-English translation direction, while the HuggingFace Helsinki Opus MT model is best for Bulgarian-English.

Keywords: Evaluation of machine translation · Bulgarian · Romanian.

Proceedings of New Trends in Translation and Technology, pages 210–225
3–6 July, 2024, Varna, Bulgaria.
https://doi.org/10.26615/issn.2815-4711.2024_018

1 Introduction

Detecting incorrect information, disinformation, and verifying content have become important tasks in the last decade [15]. While journalists and fact-checkers daily perform fact-checking, this is a hard and time-consuming work, which they have to often shorten, to cope with their stringent deadlines. To assist journalists and other specialists with fact-checking, (semi-)automatic applications and platforms are being created. Such applications either assist manual fact-checking, or propose more automatic operations, using methods from Computer Vision and Natural Language Processing (NLP).

The NLP methods for automatic fact-checking somewhat differ from manual fact-checking. One difference (frequently observed by one of our co-authors, who is a journalist with long experience) is that automatic fact-checking most often relies on detecting explicitly formulated *claims* (statements), while for manual fact-checking, a journalist does not necessarily need the text to contain an explicitly formulated claim, and may compose a claim by applying a mixture of summarizing, paraphrasing, and extracting the main elements from different parts of a text. Due to this difference, the steps of NLP methods for fact-checking most frequently include the following stages [1,28]: detecting check-worthy claims, containing factual and verifiable information, determining whether the detected claims have been already fact-checked, retrieving evidence, and verifying the claims. In both manual and automatic cases, however, fact-checking is performed on important real-world elements (such as names of people, organizations, locations, events, time expressions, etc.) and the relationships between them. Most such elements correspond to the so-called Named Entities (NEs) in NLP. An example of an automatic platform, assisting journalists with fact-checking is Truly Media[5], which is a web-based collaboration platform developed to support primarily journalists and human rights workers in the verification of digital content (including news articles and social media texts). Truly Media supports many languages, and new languages are added by combining the detection of Named Entities (NEs) in the specific language and Machine Translation (MT), allowing cross-lingual fact-checking. The Truly Media creators are the Athens Technology Center (ATC) and Deutsche Welle (DW). ATC is a consortium member of the international project BROD[6] (Bulgarian-Romanian Observatory of Digital Media), coordinated by the Big Data for Smart Society Institute (GATE) in Sofia, Bulgaria. BROD is part of the EDMO (European Digital Media Observatory[7]) network. It aims to create a multinational, multi-stakeholder, and multidisciplinary regional hub for fact-checking, and media literacy campaigns, supplied with newly created robust technical infrastructure for the detection, analysis, and combating disinformation circulating in Bulgaria and Romania. To achieve this last objective, one of the project's tasks is to provide publicly

[5] https://www.truly.media/solution/. Last accessed on June 14th, 2024.
[6] More information about BROD can be found on its website: https://brodhub.eu/. Last accessed on June 14th, 2024.
[7] https://edmo.eu/ Last accessed on June 14th, 2024.

available monolingual and cross-lingual language-specific tools for the automatic verification and fact-checking of Bulgarian and Romanian news articles and social media texts. Such tools will be on one hand integrated into Truly Media, and on the other - shared for free public usage.

Monolingual fact-checking is enough when claims are country-specific, and there exist databases of already checked claims in the respective language. However, it is often necessary to check Bulgarian and Romanian claims in English-language databases and other fact-checking resources[8] (known as cross-lingual fact-checking). Such procedure is necessary for claims, related to topics with information, available in two or more countries (e.g. the European Union elections, or Covid-19 vaccinations). It is also motivated by the fact that English-language databases are usually much richer and more frequently updated than the ones in less-resourced languages.

Although different methods for cross-lingual fact-checking exist [23,24,25,26,27] (including using multilingual models, cross-lingual information retrieval, and transfer learning), to achieve integration with Truly Media, BROD project's task involved determining which are the best (for fact-checking) open-source/open-access MT tools for Bulgarian-English and Romanian-English language pairs and translation directions. While manual translation would be more precise, MT is necessary as large amount of texts will be processed.

Even if MT has already been used for cross-lingual fact-checking [23,25,18,19,20], we could not find any specific criteria which an MT engine needed to satisfy, to be appropriate for such a task. We could not find either any evaluation metric, showing how adequate an MT tool is for this task. To address this gap and achieve our objective, in this article, we present a new MT evaluation approach for MT engines, providing a measure of their appropriateness for cross-lingual fact-checking. The approach includes a targeted error classification, manual evaluation guidelines, and a score. We also present the settings and the results of two evaluation experiments with which we determined the best MT engines for translating Bulgarian and Romanian news articles and social media texts into English for manual and automatic fact-checking. We are confident that this new evaluation approach will be of use to researchers, developers, and journalists, applying MT tools for cross-lingual fact-checking in any language direction.

Next, Section 2 mentions the relevant Related work, Section 3 presents our MT evaluation approach, Section 4 describes the experiments. Section 5 presents the Conclusions, Ethical Aspects, and Future Work, and finally - Section 6 provides the Acknowledgements.

2 Related Work

Evaluation of Machine Translation is a widely addressed topic [3,5,7,4,8], with a large number of approaches, including manual and automatic ones, or a combination of them. Manual evaluation usually includes ranking translations from

[8] Examples are: https://www.politifact.com/, https://www.snopes.com/, https://firstdraftnews.org/, and https://efcsn.com/.

different engines, or assigning (a) score(s) (with the most frequent criteria looked for being *fluency* and *adequacy*). Analyzing translations for categories of errors with different severity weights is another manual evaluation approach. Automatic metrics (such as BLEU, TER, HTER, METEOR, etc.) usually include comparing MT-produced translations with manual ones or counting the amount of post-editing (manual or automatic correction) needed. **The appropriateness of MT tools for the tasks of manual and automatic fact-checking is characterized by specific requirements** (see Section 3), which makes the application of the existing general MT evaluation metrics and approaches not suitable. Due to the same requirements, we cannot rely on manual reference translations, and post-editing is not included. Automatic metrics do not provide a detailed view of the MT engines' performance for specific aspects, and the usual error classifications (such as the Multidimensional Quality Metrics, MQM[9]) are unnecessarily detailed, take into account aspects which we ignore (such as the change in style), and do not consider details, important for our task (such as the correct translation of NEs).

Both manual translation [28] and MT [23,25,18,19,20] have already been used for translating the titles [20] or the whole texts [18,19] for cross-lingual claim matching and fact-checking. However, no work appears to describe the characteristics of an MT engine, which would make it suitable for these tasks. Some works [2] just mention the importance of MT quality for translating texts, without providing any details.

As the correct translation of NEs is an important point in our approach, similar to us are task-specific MT evaluations involving NEs. Such are works that check if NEs have been correctly translated, and whether the translated NEs can be matched against knowledge bases [11,12,13,14]. This type of works, however, do not include the other aspects important for fact-checking.

3 Evaluation Approach

We consider the fact that the good quality translation of source into target texts for the task of subsequent manual and automatic claims matching and fact-checking must abide by specific requirements. These requirements are less strict (detailed) than those of texts, which need to be translated by translation agencies, such as official documents, legislation, and books. However, there are some specific aspects, which need to be checked for. Specifically, the MT tools used for this task have to be able to:

- **transfer the exact meaning of the source text** (as much as possible), without modifying, or omitting parts of the source text's meaning, or adding any new meanings, not present in the source text,
- **not consider synonyms, style change, or paraphrases as errors** (including the alternative translations of NEs), but as acceptable translations,

[9] https://themqm.org/. Last accessed on June 14th, 2024.

- **consider the correctness and the fluency of the target English translations with less importance** (lower weight),
- **take into account both the processes of human fact-checking, and automatic fact-checking**,
- **correctly translate important elements, such as the Named Entities (NEs), events, time expressions, and the relationships between them (such as the logical and correct temporal ordering).** Consider this aspect with a higher weight.
- **Allow flexibility of weights, needed to reflect the available components of cross-lingual fact-checking pipelines.**

Based on the above requirements, we created manual evaluation guidelines that can be used for evaluating the quality of MT engines for this task. In this first version, we consider only the correct translation of NEs of the types *people's names*, *organisations*, *locations*. We plan to include time expressions and the other aspects in future work. We apply the evaluation approach in two experiments, described in Section 4. We also share the settings of the annotation tool used.

The manual evaluation of MT tools following our approach is best applied when the human evaluators are specialized in translating texts in the specific language pair and language direction and preferably have experience with using and evaluating MT. We share a Python script which allows to calculate a final score, characterizing the performance of each MT engine. These scores can be used to compare different MT engines, in order to select the most appropriate one for this task. The scores include numerical weights, which can be modified.

4 Evaluation Experiments

We conducted two evaluation experiments, aiming to determine the most appropriate MT tools for translating news articles and social media posts from Bulgarian and Romanian to English for manual and automatic claim matching and fact-checking. The settings of **Experiment 1 for Bulgarian** and **Experiment 2 for Romanian** slightly differed in terms of the source texts used and the approaches, followed for collecting them. The Sections below describe the various details.

4.1 Preliminary Selection of the MT Tools

The MT tools to be integrated had to abide to criteria, part of which were specified in BROD's grant agreement, and part by us:

- **License:** to be open-source/open access tools licensed for commercial use or non-open access tools that provide access without a fee,
- **Activity level:** to be maintained or in active development,

– **Integrability:** to provide REST API access or Docker Image access (for local hosting or hosted image), which facilitates integration,
– **Languages and directions:** to support both Bulgarian and Romanian, as well as translations in both directions

We reviewed all MT tools in the European Language Grid (ELG) platform[10]. The MT engines matching all criteria were **HelsinkiNLP - OPUS-MT** [6] (created by the University of Helsinki, based on Marian Neural Machine Translation, trained on OPUS), the Neural MT (**NTEU**[11]) for public administration, and European Commissions's **eTranslation** system[12]. We ran a preliminary fast check of 3 texts with a Bulgarian and a Romanian translators and discovered that NTEU was producing too low-quality translations for Bulgarian. For this reason, we removed it and subsequently added **UEDIN-MT**[13] (also based on Marian NMT and trained on OPUS), which was supporting only Romanian, but was matching the other criteria, and was strongly recommended to us by Romanian MT users). From all available HelsinkiNLP - OPUS-MT options, we took the most recently updated versions, which appeared to be the models, uploaded to the HuggingFace platform[14]. These models, however, had 512 token restrictions for their input for both Bulgarian and Romanian engines. For this reason, we split the source texts into sentences and run separately each sentence through these engines (we did this only for the Helsinki OPUS MT tools). To measure whether the resulting translations sounded like linked texts, we added an MT evaluation aspect for all MT engines - text cohesiveness. This aspect was also important, as it was necessary to preserve correctly any existing relationships between key elements in separate sentences of the texts, including the anaphoric ones.

4.2 Selection of the Source Texts

For each language, we used 100 short texts on a variety of topics of different types (political, sports, cultural, health-related), important in both countries for the period 2020-2024. The granularity of topics differed - the topics, important for Bulgaria were covering shorter periods of time (from a few days to several months), while those for Romanian - were periods of several years.

For each topic, we created language-specific keywords, which were used to retrieve the source texts. The methods for creating the keywords were different for the two languages. Specifically, the keywords for Romanian were collected by the Romanian co-author of this article, and those for Bulgarian - by three of the Bulgarian co-authors, who used different methods, including combinations of the words, composing the topics, or replacing some of them with synonyms, and

[10] https://live.european-language-grid.eu/. Last accessed on June 14th, 2024.
[11] https://nteu.eu/. Last accessed on June 14th, 2024.
[12] https://commission.europa.eu/resources-partners/etranslation_en.
[13] https://live.european-language-grid.eu/catalogue/tool-service/5328
[14] https://huggingface.co/

sometimes specifying time period to restrict search. For some Bulgaria-related topics, several approaches were applied.

Tables 1 and 2 show the topics for Bulgarian and Romanian. Due to the difference in the keywords selection, the Bulgarian topics table does not specify them. On the other hand, as mentioned, the Romanian topics are much more general, and all cover several years from 2020 to 2024, and for this reason, the table does not specify these periods. In total, the Bulgarian source texts contained 13573 words, while the Romanian ones – 10129 words. Table 3 shows the minima, maxima, and average length of texts in number of words in Bulgarian and Romanian and their English translations by each MT engine. The Bulgarian texts were a mixture of news articles of at most one paragraph (collected from the BROD website and via Google search), and social media (Facebook) messages. The Romanian texts were only Facebook messages. The Facebook messages were collected using CrowdTangle[15].

Table 1: Topics and time periods of the Bulgarian source texts

Year	**Date**	**Topics (translated in English)**
2022	14 April	The members of the Parliament closed the specialized justice
	27 April	Russia stopped the gas for Bulgaria
	9 May	Procession/March for Ukraine
	16-22 June	Protests in defence of Nikola Minchev and the government of Kiril Petkov
	22 June	Petkov's government was overthrown
	24 June	We have lifted the veto over North Macedonia for the EU
	3 July	Bulgaria expels 70 Russian diplomatic staff
	5 July	The car crash on "Cherni Vrah" Boulevard
	5 August	Protests against the Russian gas
	8 July & 1 October	Completion ceremony and opening ceremony of the gas connection with Greece
	2 October	Early parliamentary elections again
2023	2 April	Parliamentary elections and a rotating prime minister
	1 May	Assassination attempt against Geshev
	24 May	The book "Time Refuge" by Georgi Gospodinov was awarded the prestigious international Booker Prize
	June	The Debora Case
	6 July	Zelensky's visit to Bulgaria
	16 August	The murder of Alexei Petrov
	August	Vezhdi Rashidov resigned
	September	The floods along the Southern Black Sea coast

Continued on next page

[15] https://www.crowdtangle.com/

Table 1: Topics and time periods of the Bulgarian source texts (continued)

Year	Date	Topics (translated into English)
	29 October	Local elections
	6 December	The protest in Culture - cultural figures came out in a national protest against low incomes
2024	5 January	The fail of one of the assemblies of different Bulgarian parties
	11 January	The prosecutor's office rejected the ideas of legal processing of industrial hemp again
	12 February	Rumen Radev attacks the changes in the Constitution
	14 March	Patriarch Neophite died
	24 March	Emergency elections, rotation of the Bulgarian government
	31 March	Bulgaria partially entered Schengen
	April	The scandal in Bulgarian customs

Table 2: Topics and Search Strings of the Romanian source texts

Topics	Search String in Romanian
NATO	NATO OR "Alianța Nord-Atlantică" OR "Tratatul Nord-Atlantic"
Ukraine War and Russia	Ucraina OR "război" OR Rusia OR "conflict militar" OR "invazie rusească"
Direct attacks near Romanian border	atacuri OR "atacuri directe" OR "granița României" OR "incidente frontieră"
Moldova (Republic of Moldova)	"Republica Moldova" OR "Chișinău"
Vaccination	vaccinare OR vaccinuri OR "campanie de vaccinare" OR "imunizare COVID"
Schengen	Schengen OR "spațiu Schengen" OR "aderare Schengen"
Different protests and strikes	proteste OR greve OR "manifestații publice" OR "conflict social"
Taxation (and other government policies)	taxare OR impozite OR "politici fiscale" OR "legislație fiscală" OR guvernamentale OR "cota unica"
Energy prices (help from the government)	energie OR "prețuri energie" OR "ajutor stat" OR subvenții OR "scutiri fiscale"

Continued on next page

Table 2: Topics and Search Strings of the Romanian source texts (continued)

Topics	**Search String in Romanian**
Justice (and a lot of particular cases - drugs, corruption, incidents)	justiție OR droguri OR corupție OR incidente OR "cazuri judiciare" OR tribunal
Relationship with the Hungarian minority	"minoritatea maghiară" OR "relații interetnice" OR UDMR OR "drepturi minoritare" OR "Viktor Orban"
Anti-globalist messages	anti-globalizare OR "mesaje anti-globaliste" OR "suveranitate națională" OR "opoziție globalizare" OR globalisti
Romanian diaspora	diaspora OR "români în străinătate" OR emigranți OR expatriați OR "comunități românești"

Table 3. Detailed text lengths in number of words.

Language Pair	**Source Texts**			**Translations**			
	Min	**Average**	**Max**	**MT tools**	**Min**	**Average**	**Max**
BG-EN	9	101.29	264	OPUS	8	112.36	308
				eTrans.	8	110.91	309
RO-EN	61	135.73	236	OPUS	59	138.98	252
				eTrans.	55	138.70	245
				UEDIN	60	135.46	265

4.3 Evaluation Procedure

The manual evaluation was done by using the simple web-based annotation tool Datasaur[16] by 2 human evaluators, one per language. Both evaluators were native in the respective language, and had experience writing and reading news articles and social media messages, translating for their respective language pairs in both directions, and using, post-editing, and evaluating MT translations. While more evaluators would reduce human subjectivity, we did not have the time and resources to include more of them. The evaluators followed specially prepared Evaluation Guidelines and knew that they were evaluating MT translations.

When evaluating, they were shown a pair of source text, followed by its translation by one of the MT engines. We selected Datasaur, as it was user-friendly, and allowed both to answer questions about each source text-translation shown (called "*document labeling*") and mark sequences of words and assign them categories inside the texts (called "*span labeling*").

[16] https://app.datasaur.ai/

We created two detailed and clear Evaluation Guidelines documents: 1) Guidelines explaining how to use the Datasaur[17]; and 2) Guidelines which explain how to evaluate the MT-produced translations[18].

Specifically, for each pair of source text - MT translation, the evaluator had to read the texts, answer 5 general multiple-choice questions, and if any errors were present - mark the respective words either in the source text or in the translation, with a specific color (corresponding to an error category). The evaluator was also introduced to the general requirements explained in Section 3.

After the manual evaluation was completed, we automatically obtained a unique final score for each MT engine for the respective language pair, along with scores, reflecting the answers to the questions. These scores indicated which MT engine to select for each language pair. To obtain the scores for the questions, we assigned a numerical value to each of their answers. The Question score for the respective pair was then obtained by multiplying some values by a numerical weight. The numerical weights can be modified, to reflect what is possible at the next stage (for example if words in the source language can be still processed), and any specific requirements. For these experiments, we used the weights below:

1. **Question 1: Is the original text translated into English?**
   (a) Yes, fully. - $Q1Score_{pair} = 1$
   (b) Partially, some parts are missing or left in the original language. - $Q1Score_{pair} = 2$
   (c) Most of the text is left in the original language. - $Q1Score_{pair} = 3$
   (d) No, the text is all left in the original language, or no English text is produced. - $Q1Score_{pair} = 4$
2. **Question 2: Is the translation text written in fluent English?** (lower weight)
   (a) The English translation text is overall correct, but it is clear that it is an automatic translation. - $Q2Score_{pair} = 1x0.5$
   (b) The English translation text is overall correct, but it is clear that it is an automatic translation. - $Q2Score_{pair} = 2x0.5$
   (c) The English text is too much wrong or completely incorrect. - $Q2Score_{pair} = 3x0.5$
3. **Question 3: Is the English translation text cohesive with logically linked sentences?**
   (a) Completely cohesive, all sentences are logically linked and the whole text sounds natural. - $Q3Score_{pair} = 1$
   (b) Somewhat cohesive. Some sentences appear unconnected. - $Q3Score_{pair} = 2$
   (c) Most sentences appear unconnected, but some are connected. - $Q3Score_{pair} = 3$
   (d) All sentences appear unconnected. - $Q3Score_{pair} = 4$
4. **Question 4: Are the Named Entities (NE) correctly translated?** (higher weight)

[17] Can be accessed at: https://shorturl.at/th44W.
[18] Can be accessed at: https://shorturl.at/xXt36.

(a) Yes, all and correctly. - $Q4Score_{pair}$ = *1x1.5*
(b) Partially - some are not in English or wrongly translated. - $Q4Score_{pair}$ = *2x1.5*
(c) Most are left in the original language or are wrongly translated. - $Q4Score_{pair}$ = *3x1.5*
(d) No, all are untranslated, missing, or wrongly translated. - $Q4Score_{pair}$ = *4x1.5*

5. **Question 5: How confident are you about evaluating the quality of this translation?**
   (a) Confident. - $Q5Score_{pair}$ = *1*
   (b) Somewhat confident/at least a bit unsure. - $Q5Score_{pair}$ = *2*
   (c) Not confident. - $Q5Score_{pair}$ = *3*

Question 5 was assessing the confidence of the evaluator, if the evaluator was less confident, this would result in a lower score.

We obtained the scores for the error span categories by dividing (normalizing) the number of words, marked with this category in each pair by the total number of words in the source text. The result is then multiplied by each weight. Please note, that we call "original" the source text (in Bulgarian or Romanian language).

– **untranslated_words_orig** - words or expressions from the source text left in the source language in the translation.

$$UnTransScore_{pair} = \frac{\text{Num}_{\text{UnTransTokensPair}}}{\text{Num}_{\text{AllTokensSourceTextPair}}} \tag{1}$$

– **words_translated_wrong_meaning** - words or expressions translated, but with a wrong meaning - corresponds to modified meaning, and does not include synonyms, paraphrases, and style changes

$$TransWrongMeanScore_{pair} = \frac{\text{Num}_{\text{TransWrongMeanTokensPair}}}{\text{Num}_{\text{AllTokensSourceTextPair}}} \tag{2}$$

– **words_in_original_missing_in_translation** - words or expressions present in the source text, but completely missing (does not include if left in the source language) in the translation text - correspond to omitted meaning

$$ExtraTransMissOrigScore_{pair} = \frac{\text{Num}_{\text{ExtraTransMissOrigTokensPair}}}{\text{Num}_{\text{AllTokensSourceTextPair}}} \tag{3}$$

– **extra_words_in_translation_missing_in_orig** - correspond to added meaning, expressions, or words that are present in the translation, but missing in the source text

$$OrigMissTransScore_{pair} = \frac{\text{Num}_{\text{OrigMissTransTokensPair}}}{\text{Num}_{\text{AllTokensSourceTextPair}}} \tag{4}$$

– **translated_words_wrong_English** - correspond to wrongly written expressions in the English translation, not fluent English (lower weight)

$$WrongEngScore_{pair} = (\frac{\text{Num}_{\text{WrongEngTokensPair}}}{\text{Num}_{\text{AllTokensSourceTextPair}}}) * 0.5 \tag{5}$$

Each source text-translation pair gets a unique score, obtained by summing all the questions and error span scores for this pair:

$$\begin{aligned} TotalScore_{pair} = \\ & Q1Score_{pair} + Q2Score_{pair} + Q3Score_{pair} + Q4Score_{pair} + Q5Score_{pair} \\ & + UnTransScore_{pair} + TransWrongMeanScore_{pair} \\ & + ExtraTransMissOrigScore_{pair} + OrigMissTransScore_{pair} \\ & + WrongEngScore_{pair} \end{aligned} \tag{6}$$

Finally, we obtain a total score for each MT engine, by summing the total scores for each pair, and dividing the sum by the number of pairs (100):

$$\boldsymbol{MTtoolScore_{LangEn}} = \frac{\sum TotalScore_{pair}}{Num_{LangEnPairs}} \tag{7}$$

To more meaningfully compare the results of different engines, we convert the total score to a value between 0.0 (the best) and 1.0 (the worst), by using the min-max normalization, and define numerical ranges with specific meanings:

$$\boldsymbol{MTtoolScoreNorm_{LangEn}} = \frac{MTtoolScore_{LangEn} - TotalScore_{pairMin}}{TotalScore_{pairMax} - TotalScore_{pairMin}} \tag{8}$$

1. 0.00 to 0.19, corresponding to "excellent"
2. 0.20 to 0.39, corresponding to "good"
3. 0.40 to 0.59, corresponding to "average"
4. 0.60 to 0.79, corresponding to "poor"
5. 0.80 to 1.00, corresponding to "very poor"

We compare the $\boldsymbol{MTtoolScoreNorm_{LangEn}}$ for all the MT tools separately for Bulgarian-English and Romanian-English. We also obtain the values for the specific aspects.

Evaluation Experiments Results and Discussion Table 4 shows the final normalized scores for each MT tool.

Table 4. Normalized Total MT Evaluation Results.

Lang. Pair	MT Tool	Total Norm. Score	Range
BG-EN	**HelsinkiNLP - OPUS-MT**	**0.3169**	0.20 - 0.39 (**good**)
BG-EN	eTranslation	0.4373	0.40 - 0.59 (average)
RO-EN	HelsinkiNLP - OPUS-MT	0.3762	0.20 - 0.39 (good)
RO-EN	**eTranslation**	**0.3570**	0.20 - 0.39 (**good**)
RO-EN	UEDIN-MT	0.4638	0.40 - 0.59 (average)

The higher the score - the worse the MT tool performance, and the lower the score - the better it is. The results show that for Bulgarian-English the best

MT tool for manual and automatic fact-checking is the current (as per June 2024) HuggingFace model HelsinkiNLP - OPUS-MT and for Romanian-English - eTranslation. The aspects-specific scores, overall also reflect this finding (see Table 5). The Python script, used for calculating the scores is available by contacting the first author by e-mail.

Table 5. MT Evaluation Results for Specific Aspects.

Aspects	OPUS BG	eTransl. BG	OPUS RO	eTransl. RO	UEDIN RO
Q1(Translated?)	**1.28**	1.39	1.31	**1.25**	1.58
Q2(Fluent En?)	**0.94**	1.03	0.98	**0.94**	1.09
Q3(Cohesive?)	**1.46**	1.82	1.84	**1.67**	2.14
Q4(NEs?)	**2.20**	2.44	2.79	**2.64**	2.83
wrong mean.	**0.04**	0.05	0.02	**0.013**	0.014
untrans. words	**0.0006**	0.0011	**0.0056**	0.0057	0.0068
missing words	**0.0080**	0.0122	0.0072	**0.0031**	0.0263
extra words	**0.0005**	0.0007	0.0026	**0.0006**	0.0090
wrong English	**0.0102**	0.0137	0.0040	**0.0039**	0.0042

5 Conclusions, Ethical Aspects, and Future Work

In this article we presented the first version of an approach for evaluating the suitability of MT engines to translate texts, which will be used for manual and automatic claim matching and fact-checking. The approach includes an error classification with categories and weights, evaluation guidelines, formulas for obtaining the final scores, and a Python script for calculating these scores. We also present the results of two experiments in which we apply this evaluation approach to determine the most suitable free-access MT tools for translating short news articles and social media texts from Bulgarian and Romanian into English. While our work contains some limitations, we plan to improve it in the near future, by including the detection of temporal expressions, adding more evaluators, and detecting incorrectly translated NEs in a more precise way. However, we are confident that this approach will be useful to researchers planning to apply MT for cross-lingual fact-checking. The manual evaluation was performed by the first author for Bulgarian (who is hired by GATE Institute and this work is considered as part of her duties). The Romanian evaluation was manually done by a specially hired native Romanian translator from a Romanian translation agency (not included as a co-author). The Romanian evaluation work was paid 600 USD for 300 translation pairs, which is considered a well-paid manual annotation job.

6 Acknowledgements

This work has received funding from the European Union under Contract number: 101083730 — BROD. This document reflects the views only of its co-authors,

and the Commission cannot be held responsible for any use which may be made of the information contained herein. The results presented in this paper are also part of the GATE project. This project has been funded by the European Union's Horizon 2020 WIDESPREAD 2018–2020 TEAMING Phase 2 programme under Grant Agreement No. 857155 and in part by the BROD Project, funded by the European Union under Grant Agreement No. 101083730.

The authors' contributions to this work are: Irina Temnikova designed the MT evaluation approach, prepared the Evaluation Guidelines, trained the Romanian evaluator, manually evaluated the Bulgarian-English translations, read the related work, added some Bulgarian topics, collected part of the Bulgarian texts, did automatic filtering of the Helsinki OPUS MT texts, and wrote the article (most from scratch, but also by taking a bit from the BROD 2.4 Deliverable). Silvia Gargova participated in the manual collection of Bulgarian texts, run the MT tools, wrote the used bits of the BROD 2.4 Deliverable (e.g. the tables with Bulgarian and Romanian topics), provided comments on the article, and added Tables 1, 2, and 3. Tsvetelina Stefanova run the MT tools, participated in the initial collection of Bulgarian texts, and gave feedback on this article. Iva Marinova prepared a Python script for splitting the text into sentences and running it through the HuggingFace Helsinki Opus MT models. Ruslana Margova took part in the collection of Bulgarian texts, added bits from the Deliverable to this article, and participated in its final revision. Nevena Grigorova collected almost all Bulgarian topics and participated in manually filtering the Romanian and part of the Bulgarian texts. Alexander Komarov collected part of the Bulgarian texts, Dan Sultanescu prepared the list of Romanian topics and search keywords, as well as shared the Romanian and Bulgarian BROD datasets. Kalina Bontcheva acted as a senior author and provided insights at different stages of the work.

Finally, we would like to express our gratitude to Institute GATE, and Keith Kiely – for giving us the opportunity to work on MT evaluation during project BROD; to the NETTT conference organizers; and to the anonymous reviewer – for supplying very useful comments, which significantly improved our paper.

References

1. Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Míguez, R., Caselli, T., Kutlu, M., et al.: Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In: Conference and Labs of the Evaluation Forum, CLEF 2022, CEUR Workshop Proceedings, vol. 368-392, pp. 368-392. CEUR-WS.org (2022).
2. Panchendrarajan, R., Zubiaga, A.: Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. Journal of Information Processing and Management, vol. 60, pp. 102527. Elsevier, Amsterdam (2023). https://doi.org/10.1016/j.ipm.2023.102527
3. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311-318 (2002)
4. Chatzikoumi, E.: How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering **26**(2), 137-161 (2020)

5. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pp. 65-72 (2005)
6. Tiedemann, J., Thottingal, S.: *OPUS-MT* — Building open translation services for the World. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT), Lisbon, Portugal (2020)
7. Snover, M., 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, pp. 223-231 (2006)
8. Koehn, P., Monz, C.: Manual and automatic evaluation of machine translation between European languages. In: Proceedings of the Workshop on Statistical Machine Translation, pp. 102-121. Association for Computational Linguistics (2006)
9. Menezes, L. M. C.: Named Entities Recognition for Machine Translation: A Case Study on the Importance of Named Entities for Customer Support. Doctoral dissertation (2021)
10. Aïmeur, E., Amri, S., Brassard, G.: Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining 13(1), 30 (2023)
11. Babych, B., Hartley, A.: Comparative evaluation of automatic named entity recognition in machine translation output. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP) (2004, March)
12. Babych, B.: Information Extraction Technology in Machine Translation. Doctoral dissertation, University of Leeds (2005)
13. Babych, B., Hartley, A.: Sensitivity of Automated MT Evaluation Metrics on Higher Quality MT Output: BLEU vs Task-Based Evaluation Methods. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), p. 6th (2008, May)
14. Gekhman, Z., Aharoni, R., Beryozkin, G., Freitag, M., Macherey, W.: KoBE: Knowledge-based machine translation evaluation. arXiv preprint arXiv:2009.11027 (2020)
15. Aïmeur, E., Amri, S., Brassard, G.: Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining 13(1), 30 (2023)
16. Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.): Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020.
17. Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., Al-Homaid, A., Zaghouani, W., Caselli, T., Danoe, G., Stolk, F., Bruntink, B., Nakov, P.: Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
18. Mori, M., Papotti, P., Bellomarini, L., Giudice, O.: Neural machine translation for fact-checking temporal claims. In: Aly, R., Christodoulopoulos, C., Cocarascu, O., Guo, Z., Mittal, A., Schlichtkrull, M., Thorne, J., Vlachos, A. (eds.) Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER), pp. 78–82. Association for Computational Linguistics, Dublin (2022).
19. Nakov, P., Alam, F., Shaar, S., Da San Martino, G., Zhang, Y.: A second pandemic? Analysis of fake news about COVID-19 vaccines in Qatar. arXiv preprint arXiv:2109.11372 (2021).

20. Huang, K.H., Zhai, C., Ji, H.: CONCRETE: Improving Cross-lingual Fact-checking with Cross-lingual Retrieval. arXiv preprint arXiv:2209.02071 (2022).
21. Nikoulina, V., Sandor, A., Dymetman, M.: Hybrid adaptation of named entity recognition for statistical machine translation. In: Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT, pp. 1-16 (2012, December)
22. Menezes, L. M. C.: Named Entities Recognition for Machine Translation: A Case Study on the Importance of Named Entities for Customer Support. Doctoral dissertation (2021)
23. Şahin, B. U., Karagoz, P.: Cross-Lingual Learning vs. Low-Resource Fine-Tuning: A Case Study with Fact-Checking in Turkish. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1054-1065. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.80
24. Zhou, X., Karou, M., Papadopoulos, S.: CONCRETE: Improving Cross-Lingual Fact-Checking with Cross-Lingual Retrieval. In: Proceedings of the 2022 Conference on Computational Linguistics (COLING), pp. 2448-2458. ACL, Online (2022). https://doi.org/10.18653/v1/2022.coling-main.212
25. Celikyilmaz, A., Bosselut, A.: T3L: Translate and Test Transfer Learning for Cross-Lingual Fact Verification. In: Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 317-326. ACL, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.35
26. Ruiter, D., Dong, L.: A Resource-Light Method for Cross-Lingual Semantic Textual Similarity. In: Proceedings of the 2018 Conference on Neural Information Processing Systems (NeurIPS), pp. 2915-2925. NIPS, Montréal (2018). https://doi.org/10.5555/3327757.3327812
27. Thakur, P., Yuan, M., Zervas, P.: Learning Cross-Lingual IR from an English Retriever. In: Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 1123-1133. ACL, Online (2022). https://doi.org/10.18653/v1/2022.naacl-main.89
28. Nakov, P., Barrón-Cedeno, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., Da San Martino, G.: Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings, pp. 372–387. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_33