Machine Translated Text Detection Through Text Similarity with Round-Trip Translation

Hoang-Quoc Nguyen-Son[1], **Tran Phuong Thao**[2], **Seira Hidano**[1], **Ishita Gupta**[3], and **Shinsaku Kiyomoto**[1]

[1]KDDI Research, Inc., Japan

[2]The University of Tokyo, Japan

[3]Indian Institute of Technology, India

[1]{ho-nguyen,se-hidano,kiyomoto}@kddi-research.jp

[2]tpthao@yamagula.ic.i.u-tokyo.ac.jp

[3]ishita.gupta.ee317@ee.iitd.ac.in

Abstract

Translated texts have been used for malicious purposes, i.e., plagiarism or fake reviews. Existing detectors have been built around a specific translator (e.g., Google) but fail to detect a translated text from a strange translator. If we use the same translator, the translated text is similar to its round-trip translation, which is when text is translated into another language and translated back into the original language. However, a round-trip translated text is significantly different from the original text or a translated text using a strange translator. Hence, we propose a detector using text similarity with round-trip translation (TSRT). TSRT achieves 86.9% accuracy in detecting a translated text from a strange translator. It outperforms existing detectors (77.9%) and human recognition (53.3%).

1 Introduction

A reader may misunderstand the original meaning of a translated text[1]. For example, Facebook translated "*good morning*" into "*attack them*," leading to an arrest[2]. Adversaries can use a translator for malicious tasks such as *round-trip translation* used in plagiarism (Jones and Sheridan, 2015) to avoid human recognition or in adversarial text (Iyyer et al., 2018) to fool AI.

Existing work has investigated the detection of translated texts in various approaches. The parse tree approach (Chae and Nenkova, 2009; Li et al., 2015) exploits text structure. The N-gram approach (Aharoni et al., 2014; Arase and Zhou, 2013) estimates text fluency. The text complexity approach uses complex words (Nguyen-Son and Echizen, 2017) and phrases (Nguyen-Son et al., 2017). The text coherence approach is based on matching similar words on a paragraph level (Nguyen-Son et al., 2018, 2019b). A three-layer CNN (Riley et al., 2020) is trained on either one-way or round-trip translated texts. Our previous work (Nguyen-Son et al., 2019a) combined round-trip translation with $BLEU$ scores. All these approaches fail to detect a text translated by another translator or from a different language.

Motivation The first translation round induces a low similarity between the translated and original texts, whereas the extent of similarity increases in later rounds (Vanmassenhove et al., 2019). Let us consider an example in Fig. 1. We randomly selected an English text t from an English-Russian pair[3]; the Russian text was translated into English by Google, called $t'^{(Go,RU\rightarrow EN)}$. We measured the similarity between a text and its round-trip translation using the minimum edit distance (MED) (Levenshtein, 1966). The translated text t' is the result of using the translator once, and the similarity between t' and its round-trip translation $t'^{(Go,RU\rightarrow EN\rightarrow RU}$) is high ($MED = 1$). Otherwise, the similarity between the original text t with $t^{(Go,RU\rightarrow EN\rightarrow RU)}$ is low ($MED = 5$). Based on the difference in similarity, we can distinguish the original from the translated text.

In reality, a translator's source language is often unknown. The similarity decreases when using another language. For example, the similarity between $t'^{(Go,RU\rightarrow EN)}$ translated from Russian and its round-trip translation $t'^{(Go,RU\rightarrow EN\rightarrow DE\rightarrow EN)}$ from German is low ($MED = 6$). It is close to the similarity in the original pair

[1]When we mention a translated text, translation, translator, and Google, all are related to machine translation systems

[2]www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

[3]This pair belongs to a Commentary News corpus (Barrault et al., 2019)

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5792–5797
June 6–11, 2021.

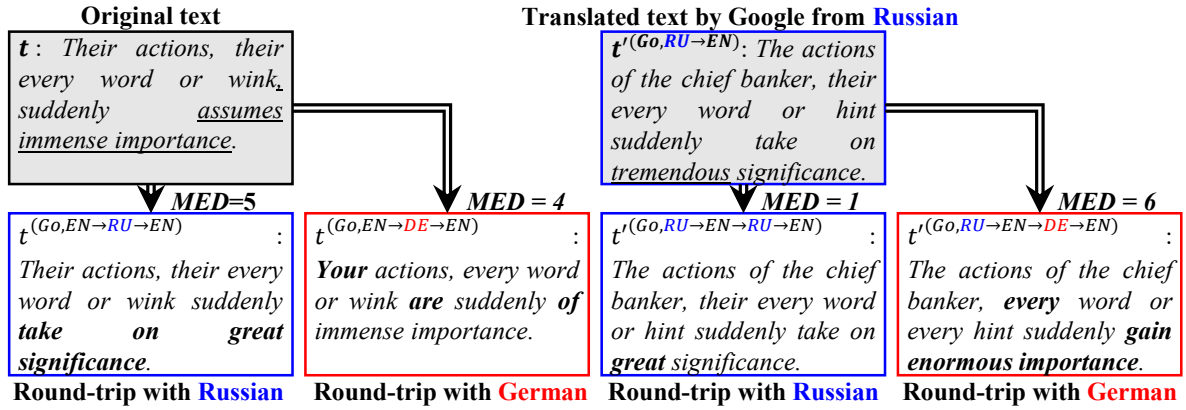

Figure 1: Round-trip translations from an original text t and a translated text t'. The superscripts indicate the translator—Google(Go)—and the language—Russian(RU) or German(DE)—that are used to generate the round-trip translations.

$\{t, t^{(Go,EN\rightarrow DE\rightarrow EN)}\}$ $(MED = 4)$. A change in a translator induces a similar phenomenon. We thus detected the translator and the language before detecting the translated text.

Contributions We propose a novel translation detector that utilizes text similarity with round-trip translation (named TSRT). This detector can be used as a warning to prevent the risk of translated texts in a certain region where people are familiar with few languages and translators. First, we create round-trip translations from multiple configuration translator and language tuples. Second, we use each tuple's round-trip translations to train individual subclassifiers. Then, we use the tuple with the highest similarity between a suspicious text and its round-trip translation to choose a suitable subclassifier. Finally, we use the subclassifer to determine if the text is an original or translated text. Experiments demonstrate that TSRT efficiently detects different kinds of translated texts (round-trip and one-way) when the translation translator and language is changed.

2 Text Similarity with Round-Trip Translation

Training Phase First, we collect original texts T_i and translated texts T'_i, which are translated with a configuration tuple π_i = {language λ_i, translator τ_i} (see Fig. 2). Second, we generate round-trip translations $T_i^{\pi_i}$ and $T_i'^{\pi_i}$ for T_i and T'_i, respectively. Finally, T_i and T'_i are combined with $T_i^{\pi_i}$ and $T_i'^{\pi_i}$ to train a subclassifier χ^{π_i} by fine-tuning the BERT model (Devlin et al., 2019). We repeat the procedure with other subclassifiers. In Fig. 1, t, t', $t^{(Go,RU)}$, and $t'^{(Go,RU)}$ belong to T, T', $T^{(Go,RU)}$, and $T'^{(Go,RU)}$, respectively, with $\pi = (Go, RU)$.

Testing Phase For a suspicious text s, we aim to determine if s is an original or a translated text. First, we generate round-trip translated texts s^{π_i} with all configuration tuples in the training phase. Next, we calculate the similarity σ^{π_i} between t and all s^{π_i} using the minimum edit distance (MED). Finally, we process s with the subclassifier associated with the best similarity σ_b corresponding to the lowest MED. In the case of t' in Fig. 1, two round-trip translations $t'^{(Go,RU)}$ and $t'^{(Go,DE)}$ are generated with respect to $\sigma^{(Go,RU)} = 1$ and $\sigma^{(Go,DE)} = 6$. The subclassifier $\chi^{(Go,RU)}$ associated with the lower MED is chosen for classifying t'.

3 Evaluation

3.1 Unchanged Translator and Language

Round-trip translation detection: We collected 11,748 distinct movie reviews from the Sentiment Treebank (Socher et al., 2013) (19.1 words/review). We chose 9,000/1,000 reviews for training/developing and used the remaining pairs for testing. This ratio is reused in further experiments. We used the original reviews to generate round-trip translations by using configuration tuples of two translators and three languages (Table 1). In addition to Google, we chose Fairseq[4] (Ng et al., 2019), the winner in the WMT'19 shared task. We compare TSRT[5] with

[4]Fairseq is only supported for Russian and German, so we cannot use it for Japanese.

[5]The source code is available at https://github.com/quocnsh/machine_translation_

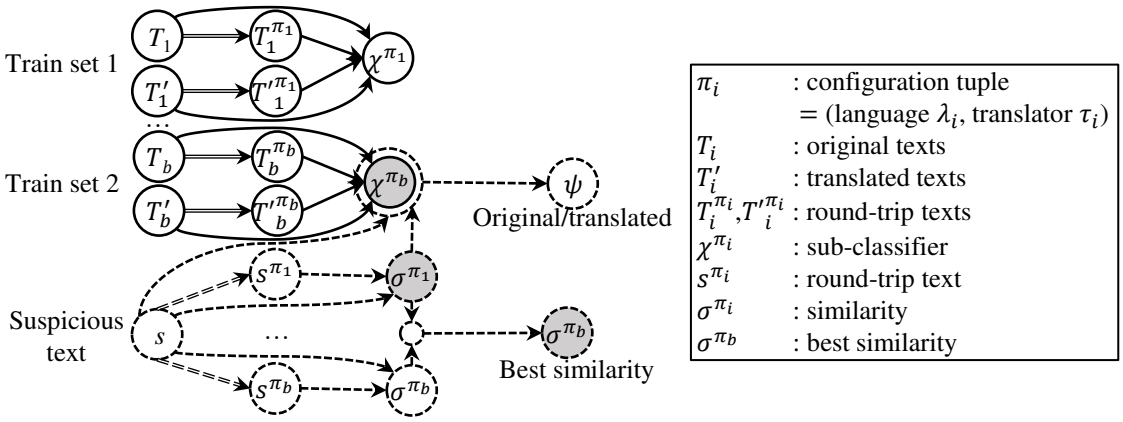

Figure 2: Text similarity with round-trip translation process (training phase: solid lines, testing phase: dashed lines).

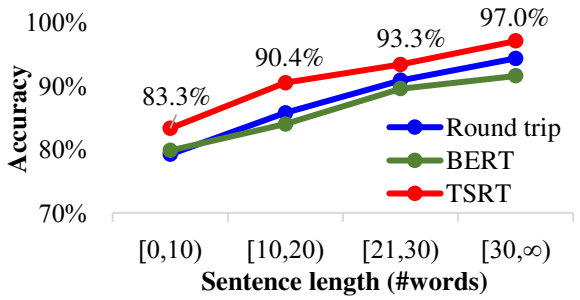

Figure 3: Top three round-trip translation detectors.

existing methods using the accuracy metric (accuracy and F-score are equivalent in this balanced corpus). BERT and TSRT have the same optimized hyperparameters[6]. The first four methods do not work well with this parallel corpus. The round-trip translation (Nguyen-Son et al., 2019a) based on $BLEU$ and BERT (Devlin et al., 2019) improves by approximately 10%. TSRT provides the highest performance, as it captures round-trip information using deep learning.

We analyzed the text lengths of the top three detectors on the whole (Go,RU) test set (Fig. 3). BERT surpasses round trips in only short length ranges, while TSRT outperforms the others in all ranges.

Human recognition: We selected 100 random reviews from the test set for human recognition[7]. We sent them to 14 raters (6 were native English speakers), who decided whether each review was an original or a translated text. The average accuracy was 53.3% (55.0% for the native speakers and 52.0% for the nonnative speakers), which was close to random. The low Fleiss' $\kappa = 0.13$ implied slight agreement in the native speakers' ratings. For nonnative speakers, κ was even lower ($\kappa = -0.07$). This indicates that the translated texts were indistinguishable by humans.

One-way translation detection: We collected parallel sentences from the Commentary News corpus (Barrault et al., 2019). We randomly selected $11,748$ pairs with 21.9 words on average per sentence (same as the movie reviews). We experimented with two languages (Russian and German) and two translators (Google and Fairseq) (see Fig. 4). Since one-way translation is more challenging to detect, the accuracy is decreased for all methods. In the top three detectors, while BERT and round-trip translation yield unstable results, TSRT remains consistent.

3.2 Changed Translator and Language

Comparison: Humans are familiar with limited languages and translators. Normally, they use their mother tongue and English (international language) and translate by choosing a popular translator such as Google or an open-source translator such as Fairseq. Table 2 presents the translation detection with translator and language changes. While the existing methods are trained with (Go,DE) or (Fa,RU), TSRT is trained on (Go,DE)+(Go,RU) or (Fa,RU)+(Go,RU), respectively. We tested all of them in (Go,RU). Our results showed that the existing methods were significantly downgraded in terms of accuracy, but TSRT remained stable.

detection

[6]We optimize hyperparameters with recommended values from BERT (maximum size of 128, batch size of 32, learning rate of $2e$-5, and epoch of 3). Since the development accuracy is equivalent to the test accuracy, we use the test accuracy for further experiments.

[7]The survey is available at https://forms.gle/L8EkZxXuEH9Co3UB7.

Method	(Go,RU)	(Fa,RU)	(Go,DE)	(Fa,DE)	(Go,JA)
Complexity (Nguyen-Son et al., 2017)	52.7	54.9	52.2	51.5	53.6
Parse tree (Li et al., 2015)	58.3	55.5	56.0	53.6	58.1
Coherence (Nguyen-Son et al., 2019b)	60.7	60.1	57.7	55.0	62.4
N-gram (Aharoni et al., 2014)	74.7	69.0	68.0	64.9	72.6
Round trip (Nguyen-Son et al., 2019a)	86.4	82.2	82.9	83.8	80.3
BERT (Devlin et al., 2019)	85.2	80.4	77.7	72.9	86.8
TSRT	**90.2**	**87.6**	**85.5**	**85.2**	**89.8**

Table 1: Round-trip translation detection with a combination of a translator—Google(Go) or Fairseq(Fa)—and a language—Russian(RU), German(DE), or Japanese(JA).

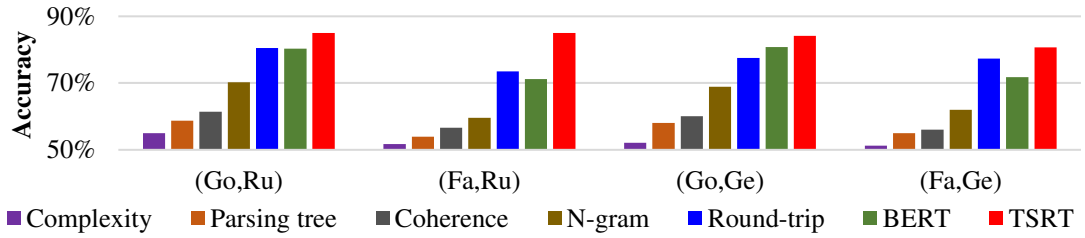

Figure 4: Detecting one-way translation with various translators and languages.

Method	Round-trip		One-way	
	(Go,DE)	(Fa,RU)	(Go,DE)	(Fa,RU)
Complexity	52.2	54.2	55.3	54.0
Parse tree	57.1	56.4	57.1	54.9
Coherence	59.4	58.9	58.5	59.3
N-gram	67.8	68.1	61.8	63.9
Round trip	61.8	56.2	61.7	60.3
BERT	77.9	75.5	67.1	75.4
TSRT	**86.9**	**86.6**	**81.9**	**82.2**

Table 2: Translation detection with translator and language changes.

Training data	Acc(Red.)
(Go,RU)	**90.2(-00.0)**
(Fa,RU)	70.2(-20.0)
(Fa,RU)	70.2(-20.0)
(Go,DE)	73.4(-16.8)
(Fa,DE)	66.6(-23.6)
(Go,RU)+(Fa,RU)	**86.9(-03.3)**
(Go,RU)+(Go,DE)	86.6(-03.6)
(Go,RU)+(Fa,RU)+(Go,DE)+(Fa,DE)	81.5(-08.7)

Table 3: TSRT's results with individuals and combinations of configuration tuples of translators and languages.

Ablation Studies: We trained TSRT on various configuration tuples and tested it on (Go,RU) (Table 3). Training TSRT on the combination with the correct configuration tuple (Go,RU) boosts the performance.

Configuration identification: We identify the translator and language on round-trip translation detection while the one-way approach obtains similar results. For translator change (Table 4's second column), we used (Go,RU) and (Fa,RU). For the language change (the third column), we used (Go,RU) and (Go,DE). All were tested on (Go,RU). We used BERT as the identification baseline. We replaced MED with $BLEU$ in TSRT. All the metric-based approaches outperformed the baseline. The translator detection outperformed language detection. While a specific translator often uses the same architecture for all languages, various translators have different architectures. Therefore, a translator change was more apparent than a language change. MED (designed for structure similarity) was better than $BLEU$ (designed for corpus levels).

4 Conclusion

This paper proposed a one-way and round-trip translation detection mechanism using text similarity with round-trip translation (TSRT), which is robust to language and translator changes. First, we trained subclassifiers on specific lan-

Method	Translator	Language
BERT	63.4%	70.0%
$BLEU$-1	92.4%	84.5%
$BLEU$-2	92.3%	84.6%
$BLEU$-3	92.3%	84.4%
$BLEU$-4	92.2%	85.0%
MED	**93.3%**	**85.6%**

Table 4: Translator and language identification.

guages/translators using round-trip translation. Then, we identified the language and translator using the highest similarity between the suspicious and round-trip translation texts. Finally, we chose the corresponding subclassifier for translation detection. The evaluation results show that TSRT outperforms other methods, with an accuracy of up to 90.2%. Moreover, TSRT could also identify the original translator and translation language with 93.3% and 85.6% of accuracy, respectively. In future work, we will exploit saturation after repeatedly using the same AI system to detect other artificial texts such as fake COVID-19 news.

Acknowledgments

We would like to thank you very much for the anonymous reviewers to provide useful comments.

References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 289–295.

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1597–1607.

Loïc Barrault, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation. In *Proceedings of the 4th Conference on Machine Translation (WMT)*, pages 1–61.

Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 139–147.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1875–1885.

Michael Jones and Lynnaire Sheridan. 2015. Back translation: an emerging sophisticated cyber strategy to subvert advances in 'digital age'plagiarism detection and prevention. *Assessment & Evaluation in Higher Education*, 40(5):712–724.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Yitong Li, Rui Wang, and Hai Zhao. 2015. A machine learning method to distinguish machine translation from human translation. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 354–360.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. In *Proceedings of the 4th Conference on Machine Translation (WMT)*, pages 314–319.

Hoang-Quoc Nguyen-Son and Isao Echizen. 2017. Detecting computer-generated text using fluency and noise features. In *Proceedings of the International Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 288–300.

Hoang-Quoc Nguyen-Son, Huy H Nguyen, Ngoc-Dung T Tieu, Junichi Yamagishi, and Isao Echizen. 2018. Identifying computer-translated paragraphs using coherence features. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC)*.

Hoang-Quoc Nguyen-Son, Thao Tran Phuong, Seira Hidano, and Shinsaku Kiyomoto. 2019a. Detecting machine-translated text using back translation. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, pages 189–197.

Hoang-Quoc Nguyen-Son, Tran Phuong Thao, Seira Hidano, and Shinsaku Kiyomoto. 2019b. Detecting machine-translated paragraphs by matching similar words. In *ArXiv Preprint arXiv:1904.10641*.

Hoang-Quoc Nguyen-Son, Ngoc-Dung T Tieu, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2017. Identifying computer-generated text using statistical

analysis. In *Proceedings of the 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1504–1511.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" nmt. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7737–7746.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of the 17th Machine Translation Summit (MT Summit)*, pages 222–232.