Translated Texts Under the Lens: From Machine Translation Detection to Source Language Identification

Anonymous ACL submission

Abstract

In this work, we tackle the problem of the detection of translated texts from different angles. On top of addressing the classic task of machine translation detection, we investigate and find the presence of common patterns across different machine translation systems as well as different source languages. Then, we show that it is possible to identify the translation systems used to produce a translated text (F1-score 88.5%) as well as the source language of the original text (F1-score 79%). We assess our tasks using Books, a new dataset we built from scratch based on excerpts of novels and the well-known Europarl dataset.

1 Introduction

Today, commercial machine translation systems (MTS) are used worldwide by hundreds of thousands of people for personal or working purposes. They help bridge the gap in language barriers, especially on the Web, by facilitating communication between people. However, bad actors also use these systems to massively target potential victims of email-phishing (Parmar and Jahankhani, 2021) or fake reviews of products to trick recommendation systems (Juuti et al., 2018). Thus, machine translation detectors are actively used to infer spam emails or to detect poor quality web pages (Google, Dec. 2021).

In this work we put automatic translated texts under the lens. We study the impact of the MTSs and the source language on the Machine Translation Detection (MTD) task leveraging Books, a novel dataset built from excerpt novels that we plan to release publicly. (Anonymized). We find that MTSs have common patterns that can be learned using a single MTS. Similarly we discovered that the source language of the text does not significantly impact MTD performances.

We then investigate the possibility of identifying the MTS used to produce the translation and its source language. To explore these points, we introduce, to the best of our knowledge, two new tasks: Machine Translation Identification (MTI) and Source Language Identification (SLI). For the first task, MTI, we built a classifier that shows promising results, with an average F1-score of 88.5%. In the second task, SLI, we propose a stacked classifier able to identify the source language with an F1-score of over 79%. We also believe that this task could be helpful in forensic analysis, where malicious actors attempt to obfuscate their writing style using MTSs (Kacmarcik and Gamon, 2006; Mahmood et al., 2019).

2 Dataset

To assess our experiments under different settings and topic domains, we perform our study using two datasets: one extracted from novels and the other based on speech transcriptions.

The first dataset we use is *Books*, a novel dataset we introduce. To build Books, we collect 100 books originally written in 4 different languages by 100 different established writers of the XX century. In particular, we select 25 books for each of the following source languages: Italian, French, Spanish, and German. The selected books belong to several different domains and authors. Thus they have very different writing styles. From each book, we select an excerpt of approximately 10,000 characters (on average 1642.67 words per novel) and their corresponding translation from the English edition. Finally, we produce 3 more English translations for each original excerpt using the APIs of 3 state-of-the-art commercial Machine Translation Systems: Google Translate(*GT*), Microsoft Translation(*MT*), and DeepL(*DL*). At the end of the process the Books dataset is made of 400 different samples.

The second dataset we use for our experiments is Europarl (Koehn, 2005). It is a parallel corpus extracted from the proceedings of the European Parliament containing *speech transcripts* of European parliamentarians and the corresponding professional translations into each of the 20 European languages. The texts on this dataset include many speech-distinctive elements such as hesitations, broken sentences, and repetition (Bizzoni et al., 2020). Consistently with Books, we obtain 100 seed samples by extracting from Europarl 25 samples for each of the 4 languages we consider. Every sample is made using transcripts of speakers of the same source language and contains about 10,000 characters (on average 1512.81 words per sample). We pre-process the dataset using the tools provided by Moses (Koehn et al., 2007). Then, we collect the parallel English translation of each seed sample. Finally, we translate each seed sample using the selected MTSs. Final datasets contain 400 (100 human-translated and 300 machine-translated) English samples.

3 Experimental Settings

For all the experiments, we use 60% of the dataset as train and 40% as test. We use Python Scikit-learn (Pedregosa et al., 2011) to implement all the models and the feature selection techniques. We use default parameters unless specified.

Feature Description. Tab. 3 shows the features we use for our tasks. ***Words avg*** is the average number of words for each sentence of the text, while ***Adjectives avg*** is the average number of adjectives for each text. We use the notation *Char-gram (i-k)* (resp. *Word-gram (i-k)*) to indicate all the char n-gram (resp. word n-gram) with $n \in \{i, \dots, k\}$. ***Dist Char-gram (i-k)*** are char-grams computed over the distortion text —text where ascii characters are replaced with a special character (Stamatatos, 2017). ***POS Word-gram(i-k)*** are word-grams computed over Part of Speech (POS) tagged text. Finally, the ***Type Token Ratio*** (*TTR*) is the ratio between the number of unique words and the total number of words for a given text. We use the Bag of Words to weight the char-gram and word-gram, while we use Tf-Idf to weight distortion text.

4 Machine Translation Detection

The *Machine Translation Detection* (*MTD*) task aims to automatically detect whether a text has been translated by a machine translation system or is human-generated. This task was broadly studied

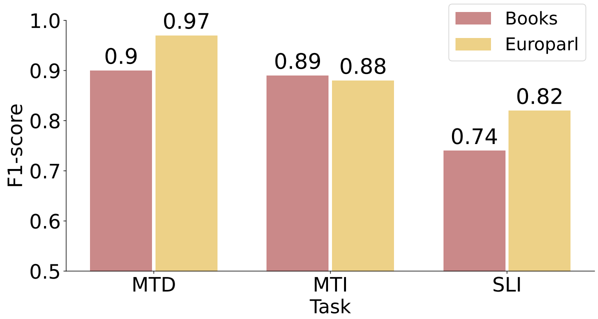

Figure 1: F1-score for the Machine Translation Detection (MTD), Machine Translation Identification (MTI) and Source Language Identification (SLI) tasks on the Books and Europarl datasets.

in the literature with different approaches such as using fixed features (Aharoni et al., 2014; Li et al., 2015), n-gram (Arase and Zhou, 2013; Popescu, 2011), coherence score (Nguyen-Son et al., 2018) and similarity with round-trip translation (Nguyen-Son et al., 2021). In this Section, we first want to replicate results similar to the state-of-the-art on our datasets. Then, we perform two experiments to further explore the underlying patterns of machine-translated texts.

For all the experiments in this section, we use the following model. We train a Multilayer Perceptron (Hornik et al., 1989) with a single hidden layer made of 10 neurons and a BFGS optimizer (Battiti and Masulli, 1990) for weights optimization. Regarding the features, we compute all the char n-gram with $n \in \{1, \dots, 6\}$ and then select the 2,500 more relevant n-gram according to the chi-square metric (Forman et al., 2003) and normalized with the SkLearn StandardScaler.

Figure 1 shows the results on Books and Europarl datasets. We obtain a high F1-score on both corpora (0.9 on Books and 0.97 on Europarl), showing that our model can achieve state-of-the-art comparable results in distinguishing machine-translated and human-translated texts.

Learning from a single MTS. The next interesting point we want to explore is if exists some common pattern among the different MTSs that allow us to detect machine-translated texts. We use only samples translated by a single MTS, and the human translated samples as the training set. Thus, we repeat the experiment 3 times, once for each MTS. Tab. 1 shows the results of this experiment for the different datasets. As we can see, the models achieve good results when tested on samples produced by machine translators not present on

Train	Books	Europarl
GT	0.85	0.82
MT	0.89	0.95
DL	0.84	0.94

Table 1: F1-score for Task 1 training on a single MTS and testing on the others.

Train	Books	Europarl
IT	0.91	0.93
FR	0.85	0.74
ES	0.88	0.78
DE	0.73	0.81

Table 2: F1-score for Task 1 training on single language and testing on the others.

Feature Type	MTI		SLI	
	Books	Euro	Books	Euro
Char-gram (1-6)	$318,250$	$220,593$	$261,895$	$175,247$
Words avg	1	1	1	1
Sentence Length	1	1	1	1
Adjectives avg	1	1	1	1
Dist. Char-gram(5-8)	$15,134$	$12,080$	-	-
Dist. Char-gram(2-8)	-	-	$13,897$	$9,522$
POS Word-gram(1-6)	-	-	$187,481$	$145,473$
TTR	1	1	-	-
All	$333,389$	$232,678$	$463,277$	$330,246$

Table 3: Features used for the MTI and SLI tasks.

the training set. Interestingly, the model trained on MT achieves similar results to those obtained by training the model with the whole dataset. These results suggest that there is some common pattern among the MTSs, that the model can learn from a single MTS.

Learning from a single language. Since we have 4 different source languages in our dataset, we want to understand the impact they might have on the MTD task. In this experiment, we train our model using only the translation from one source language and test it against the sample produced by the other source languages and the human translated samples. Tab. 2 shows the F1-score using the different source languages. Results show that the model can learn machine translation patterns even when training only on one language, suggesting that these patterns are unrelated to the source language.

5 Machine Translator Identification

Results from the previous section show that there is a common pattern among the different MTSs that allow us to differentiate machine-translated text from humans-translated. In this section, we investigate if MTS translations differ enough from each other to be able to identify which one has been used to translate a sample. Other works show that there could be potential differences between MTSs without trying to attempt to detect them. (Bhardwaj et al., 2020; Aharoni et al., 2014; Bizzoni et al., 2020) Thus, given a machine-translated text T', our goal is to identify the MTS M that generated the text T'. We call this task *Machine Translator Identification* (*MTI*). In particular, we focus on the identification of the 3 MTSs used to build the Books and Europarl datasets: *Google Translate*, *Microsoft Translation*, and *DeepL*. Given the goal of the task, for the following experiments, we use a sub-set of Europarl and Books datasets, removing from each dataset the 100 samples representing the class of human translations. For this task, we build an ensemble classifier. The first level comprises three different classifiers: a Support Vector Machine, a Logistic Regression, and a Random Tree. Then, the outputs of the classifiers are used as input to feed a hard voting layer (SkLearn VotingClassifier) for the final prediction. Tab. 3 shows the type and the number of features we use to train the three classifiers at the first level of our architecture. For all the n-gram type features, we select only the 85% most significant using SelectPercentile of SkLearn, and we standardize them with the SkLearn StandardScaler. Fig 1 reports the F1-score for the two datasets. As we can notice, our classifier performs similarly on both datasets, with an F1-score of 0.89 and 0.88 for Books and Europarl, respectively. To better understand the results, we analyze the confusion matrices of the two classifications. The confusion matrix of Books (Tab. 4) shows that GT is the hardest MTS to identify, and its misclassified samples are mostly assigned to the MT class. We found a possible explanation for these errors analyzing the BLUE score (Papineni et al., 2002) for each pair of MTS, obtaining a value of 69 for the pair GT-ML, 63 for GT-DL, and 62.9 for DL-ML. The high BLEU score between GT and MT shows that they have similar translations, leading to an erroneous classification of the GT samples. Conversely, the low similarity between the MT and DL classes could lead to the high accuracy we observe in our experiment. Finally, we obtain similar results analyzing the confusion matrix and the BLUE score for the Europarl dataset.

Actual \ Predicted	GT	MT	DL
GT	30	6	4
MT	1	39	0
DL	0	2	38

Table 4: Confusion Matrix on Books for the MTI task.

Actual \ Predicted	DE	ES	FR	IT
DE	25	0	5	0
ES	2	23	4	1
FR	0	2	27	1
IT	1	5	9	15

Table 5: Confusion Matrix on Books for the SLI task.

Actual \ Predicted	DE	ES	FR	IT
DE	27	3	0	0
ES	0	24	3	3
FR	0	3	24	3
IT	0	9	1	20

Table 6: Confusion Matrix on EuroParl for the SLI task.

6 Source Language Identification

As a final task, we propose the *Source Language Identification* (SLI). The goal of the task is, given a machine-translated text T' in a language $L2$, identify the source language $L1$ of the text T. For our experiments we consider English as $L2$ and the possible $L1$ languages are: Italian, French, Spanish or German. This task could be considered a variation of other tasks already studied in the literature. Indeed, a similar task is the Native Language Identification (NLI), where the goal is to identify the native language *L1* of a person that writes a text in a second language *L2* (La Morgia et al., 2019; Tetreault et al., 2013). Another similar task is to determine the source language of a human-translated text (van Halteren, 2008; Koppel and Ordan, 2011). Unlike the previous study, our task focuses on identifying the source language of a text that is not written or translated by a human but by a MTS. For this task, we use the stacking ensemble technique. In particular, we stacked an AdaBoost (Freund and Schapire, 1997) model with 50 LinearSVC (Cortes and Vapnik, 1995) and a Logistic Regression (Wright, 1995) model as base estimators. Tab. 3 shows the type and the number of features we use to train the stacking classifier. For all the n-gram features, we select the top 70% according to their F-value, computed with the variance analysis (ANOVA) (St et al., 1989). Then we standardize them with a StrandardScaler.

Fig 1 shows the F1-score of the model trained and tested on both our datasets. The results suggest that identifying the source language detection is easier on Europarl than in Books. As noted in (van Halteren, 2008) a possible reason could be that the Europarl dataset may contain some distinctive patterns for the source language of the speaker. Instead, the Books dataset covers a wide area of topics and contains fewer clues about the speaker's source language. Tab. 5 and 6 show the confusion matrices on the Books and Europarl dataset. The most challenging source language to detect on both datasets is Italian, frequently misclassified as Spanish or French. German is generally better identified than the other languages except for French on the Books dataset, with five classification errors. Indeed, German has the highest F1-score among all the classes, with a value of 0.86 in Books and 0.94 in Europarl. This is intuitive since German is a West Germanic language while the other 3 are Romance languages and have more features in common (Padró and Padró, 2004).

7 Conclusion and future work

In this work, we put translated text under the lens. We start evaluating the impact of MTSs and source languages on the Machine Translation Detection task. We find that MTSs have a common pattern that can be learned by a machine learning model trained with a single MTS. Moreover, the source language of the text does not significantly affect the performance of the task. Then, we introduce two new tasks: Machine Translator Identification and Source Language Identification. The goal of the first task is to identify the MTS that produced a translated text, while the second aims to identify the source language of a machine-translated text. The models we propose for both the tasks achieve an average F1-score of 88.5% and 78% respectively for MTI and SLI. Finally, we introduce Books, a novel dataset built for these tasks. Our results represent a first attempt to tackle the newly presented tasks. While we achieve good performance, we believe they could be further improved by using more advanced machine learning techniques. In our study, we perform all the analyses at the document level. In the future, it would be interesting to face the same problem at a more challenging limit, attempting to solve the tasks at the sentence level.

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 (Volume 2: Short Papers)*, pages 289–295, Baltimore, Maryland. Association for Computational Linguistics.

Anonymized. The resource will be available after acceptance.

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 (Volume 1: Long Papers)*, pages 1597–1607, Sofia, Bulgaria. Association for Computational Linguistics.

Roberto Battiti and Francesco Masulli. 1990. Bfgs optimization for faster and automated supervised learning. In *International neural network conference*, pages 757–760. Springer.

Shivendra Bhardwaj, David Alfonso Hermelo, Phillippe Langlais, Gabriel Bernier-Colborne, Cyril Goutte, and Michel Simard. 2020. Human or neural translation? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6553–6564, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

George Forman et al. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3(Mar):1289–1305.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Google. Dec. 2021. Managing multi-regional and multilingual sites.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Mika Juuti, Bo Sun, Tatsuya Mori, and N Asokan. 2018. Stay on-topic: Generating context-specific fake restaurant reviews. In *European Symposium on Research in Computer Security*, pages 132–151. Springer.

Gary Kacmarcik and Michael Gamon. 2006. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 444–451, Sydney, Australia. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Massimo La Morgia, Alessandro Mei, Eugenio Nemmi, Simone Raponi, and Julinda Stefa. 2019. Nationality and geolocation-based profiling in the dark (web). *IEEE Transactions on Services Computing*.

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: Posters*, pages 354–360, Shanghai, China.

Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using mutant-x. *Proc. Priv. Enhancing Technol.*, 2019(4):54–71.

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*, Hong Kong. Association for Computational Linguistics.

Hoang-Quoc Nguyen-Son, Tran Thao, Seira Hidano, Ishita Gupta, and Shinsaku Kiyomoto. 2021. Machine translated text detection through text similarity with round-trip translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5792–5797, Online. Association for Computational Linguistics.

Muntsa Padró and Lluıs Padró. 2004. Comparing methods for language identification. *Procesamiento del lenguaje natural*, 33.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yogeshvar Singh Parmar and Hamid Jahankhani. 2021. Utilising machine learning against email phishing to detect malicious emails. In *Artificial Intelligence in Cyber Security: Impact and Implications*, pages 73–102. Springer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Marius Popescu. 2011. Studying translationese at the character level. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 634–639, Hissar, Bulgaria. Association for Computational Linguistics.

Lars St, Svante Wold, et al. 1989. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272.

Efstathios Stamatatos. 2017. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1138–1149.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Hans van Halteren. 2008. Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 937–944, Manchester, UK. Coling 2008 Organizing Committee.

Raymond E Wright. 1995. Logistic regression.