**Novel Approach to Detecting Manual and Machine Translation from an English Corpus**

**Sara Cantor, Martina Ducret, Lauren Kruse, Carlos A. Martinez Jr., Melaney E. Moffitt**

Montclair State University

Cantors2@montclair.edu ducretm1@montclair.edu krusel1@montclair.edu moffittm1@montclair.edu martinezc11@montclair.edu

**Abstract**

This paper presents an NLP model, the purpose of which is to distinguish academic papers written originally in English from French papers translated into English (either manually or by machine) with the intention of being passed off as original English. By comparing the linguistic features between essays written in French, essays translated into English, and essays written in English, the goal model was able to reliably classify a given text into two categories: Google Translated and original English. Using this model, along with the context with which a piece of text is submitted, a hypothetical professor can have a better idea as to whether a student is trying to pass off a foreign language work as their own. We intend for our novel research in cross-lingual detection to serve as a feature to be used in future advances in the field. Our goal is to contribute a small but meaningful piece to the intricate problem of cross-lingual plagiarism.

***Keywords****: cross-language plagiarism, machine translation, manual translation, deception detection, natural language processing, academia, academic plagiarism, neural networks, feature engineering*

# Introduction

Plagiarism is a widely used phenomenon in works of various disciplines and subject material. The concept is to deceive the reader into thinking that the work presented is their own when, in fact, it has been duplicated from another author. The infringement of work is now easy to detect with the help of internet applications such as Turnitin1. This program is often employed by universities to determine the authenticity of students’ submitted work.

Retrieved from: https://help.turnitin.com/feedback-studio/turnitin-website/administrator/customizing-account-settings/enabling-translated-matching.html

The material is typically plagiarized through monolingual translation, i.e., translation from one language into the target language.

We further investigate the detection of plagiarism in the forms of manually translated work versus machine translated work at a multilingual level.  It is more arduous to identify plagiarism in a multilingual form due to the inevitable shifts in syntax and semantics that occur during the process of translation. A translated text can be incredibly far removed from the original piece, depending on the goal of the translator. If the goal is to capture the sentiment of a text, then the translator may not be as concerned with word for word accuracy. This approach would be beneficial for prose and creative pieces, where many idioms tend to get lost in translation. It would not, however, be helpful in the scientific and legal fields, where literal translations are often preferred. We hypothesize that manual translators will opt for the former approach and strive to capture the sentiment of a phrase, while machine translations will not properly handle turns of phrases and colloquialisms. The goal of our research is to build a classifier that can accurately differentiate between original English texts and French to English translated texts. Our classifier will further sort the texts deemed to be translations by labeling them as manual or machine translations.

MT *(machine translation)* is defined as the process by which computer software is used to translate a text from one natural language to another.Given a source language, machine translation automatically produces a target-language. The convenience of machine translation has made it a frequently used tool for quick and informal tasks. The goal of our research is to detect if articles have been machine translated or manually translated by the plagiarizer. We chose to utilize Google Translate as our machine translator application.

In 2016, Google developers introduced the Neural Machine Translation System (GNMT). GNMT achieved a relative improvement of 83% on French to English translation and overall reduced translation errors across languages by over 60% compared to the PBMT model previously used (Wu et al., 2016). As of date, Google is the most widely used and known machine translation engine obtaining a G2 score of 4.2 / 5.0. This score is calculated based on reviews gathered from the G2 user community as well as data aggregated from online sources and social networks2. GNMT currently has over 500 million daily users and can recognize over 100 languages. The accessibility of the system coupled with its high performance makes Google Translate an unknowing facilitator of cross-lingual plagiarism. For this reason, the machine translated text further discussed in our data section has been converted using the most up to date version of Google Translate.

1. **Relevant Work**

Cross-language plagiarism has been analyzed across multiple languages, giving a firm basis on how to incorporate this research. Oftentimes plagiarism takes many techniques such as paraphrasing or change of use in vocabulary. Primarily sentiment analysis is used to detect plagiarism prior to classification. In Ratna’s paper, they utilize a latent sentiment analysis (LSA) on English and Bahasa-Indonesian papers and a linear vector quantizer (LVQ) classifier (2017). This research will explore the use of Valence Aware Dictionary and sEntiment Reasoner (VADER) sentiment analysis.

Turnitin is used widely in academic spaces, with papers submitted online automatically being assessed by Turnitin’s plagiarism detection3. This service works by comparing a submitted document against Turnitin’s large database, as

2Data gathered from [www.g2.com/categories/machine-translation](https://www.g2.com/categories/machine-translation)

3 Retrieved from: Promote Academic Integrity: Improve Student Outcomes. [www.turnitin.com](http://www.turnitin.com)

well as various sites around the Web, and informing the professor as to whether the document is similar to those found in its search (Similarity, 2020). Turnitin has basic features included in every package, but teachers can opt to include other features that help with specific cases. One option is to enable translated matching, which is meant for teachers who allow papers to be submitted in languages other than English (Enabling translated matching, 2020).

For a limited number of languages, Turnitin can identify the language of a submitted document, translate the document into English, and then compare its similarity to Turnitin’s database as it would regularly.

Scientists at IBM created the algorithm, BLEU *(bilingual evaluation understudy)* to evaluate the accuracy of a machine translation from one natural language to another one. The higher the accuracy rate of the machine translation compared to that of human translation indicates the higher the quality of work the MT completed. The goal for machine translation is to be as close to human translation as possible. Previous work done by a team of scientists at IBM, created the BLEU method to calculate the precision of machine translation to human translation. BLEU has recorded a high accuracy rate and corresponds closely with human translation. Despite its high accuracy rates and close correspondence, BLEU is listed to have its disadvantages as well. The algorithm has significant trouble detecting the accuracy of languages that do not have word boundaries embedded in its natural language.

In the paper, *“BLEU: a Method for Automatic Evaluation of Machine Translation”*, the authors discuss machine translation on multilingual corpora that have been investigated on by human evaluations of the machine translations accuracy. However, their research shows “Human evaluations of machine translation (MT) weigh many aspects of translation, including adequacy, fidelity , and fluency of the translation” *(Hovy, 1999; White and O’Connell, 1994).* The article considers factors of previous done translations including n-gram precision on blocks of texts as well as unigrams, the sentence length, and the sentence brevity. Their research concluded that BLEU apprehends human translation very highly.Bleu scores continue to increase as machine translators improve upon their techniques, making the differences between manual and machine translated text minimal. We strive to determine if there are differences between the two, primarily focusing on linguistic dissimilarities that may arise during our feature extraction.

1. **Data**

The dataset used to create the model consisted of approximately 198 text chunks from 2 sources. The first source was a dataset of computer science abstracts written in English, found in the ACM Digital Library. The second source was of abstracts written in French, along with English translations of these abstracted completed by their original authors. These can be found in Jeremy Ferrero’s Cross Language DataSet on Github. We then translated the French abstracts into English using Google Translate, taking the result as is without checking for accuracy. This resulted in 4 sets of data: French abstracts, manual translations, machine translations, and English abstracts. Each set of data was processed by removing lower case letters and all alphanumeric characters, as well as newlines and tabs. These were all used in either the feature engineering or for training the model, which will be elaborated on in the methods section.

1. **Methods**
   1. **Multilabel Classifier**

Our dataset consists of three classes: original French, French-to-English translation, and original English.  The French-to-English category was further divided into two categories: Machine Translation or Manual Translation, creating a total of four output labels for our classifier. The task was to create a multi-label classifier that can first classify Class and then further determine the Category for all French-to-English translated texts. Multilabel classification should allow us to consider the relationships between classes and categories. Figure 1 diagrams the overall goal of our Neural Network.

A close up of a device

Description automatically generated

Figure 1: Diagram of Neural Network Classes and Categories

* 1. **Features**

We began our multi-label classification by extracting Count Vectors, WordLevel TF-IDF, N-Gram Vectors, VADER and TF-IDF Vectors. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based tool that performs sentiment analysis on text from various domains. We gained an understanding of our database by looking additional features represented in Figure 2. We also used Word2vec trained on Wikipedia news in English to create word embeddings and built a containment feature to determine relations between the translations.

A close up of a piece of paper

Description automatically generated

Figure 2: Snapshot of our feature correlation matrix

Containment is defined as the intersection of the n-gram word count of the original English text with the n-gram word count of the translated English text, divided by the n-gram word count of all the original English essays. (Clough). It is generally represented through the equation below, with the Original Text being *class1* and the Translated Text being *class-1.*

A screenshot of a cell phone

Description automatically generated

The result is a number between 0 and 1, with a result of 0 signifying no n-grams in common and a result of 1 signifying complete intersection among n-grams. The closer a document’s score is to 1, the more similar the two documents are. In our classifier, each document is compared to original English abstracts. A containment score closer to 1, in this instance, would imply that the document was indeed written originally in English, while one closer to 0 could be a strong indication of a translated work. The downfall of the above equation is that it can only compare the n-grams in one *class1* at a time, so the equation was modified to be able to account for all of the source texts in our database.

(A picture containing knife

Description automatically generated

**4.3 Multi-class Classifier**

Multiclass classification makes the assumption that each sample is assigned to one and only one label, while multilabel classification allows for multiple labels to be assigned to each text. Multilabel classification allowed us to embed manually and machine translated texts into a broader term, “Translation”. Multiclass classification will look at these categories as two separate terms with no overarching commonalities. As a result, this classifier can differentiate between English, Manual, and Machine Translations using a Keras Sequential model trained on word2vec embeddings. For this classifier we decided to exclude the French translations since our focus here was primarily to determine if there was an improvement in the Manual and Machine Translation label classification.

**5. Results**

The multiclass model used was a Sequential model. The results are depicted in Table 1.

A screenshot of a cell phone

Description automatically generated

Table 1: Results of Mutli-class Classifier

Two multilabel methods were used: Binary relevance and Classifier chain. Both used RandForest as their base binary classifier. Accuracy is measured in terms of its Hamming loss, which is the fraction of labels incorrectly predicted.

A screenshot of a cell phone

Description automatically generated  
Table 2: Results of our Multilabel Classifier

The ROC curve below illustrates the Binary Relevance model’s predictive ability. The larger the space underneath each line, the more accurate the classification. As shown, the highest accuracy was categorizing French correctly.

**6. Conclusion**

In conclusion, we were able to do multiclass classification where we classified into 3 classes: Original English, Machine Translations, and Manual Translations.

The created features include VADER Sentiment Analysis, POS tag counts, Char count, TF-IDF,and Containment. Although the features were not implemented for deep learning classification, they are expected to contribute strongly in both the multiclass class classification model and the multilabel classification model.

When the multilabel classification model was created, it was able to classify into main classes French, English, and Translations, from there it is also able to distinguish a subgroup/Category in Translations and divide between Manual English Translation and Machine English Translation. Our classifier was able to do this task with 71% accuracy for Binary Relevance with Random Forest base and 67% accuracy for Classifier chain with Random Forest base.

In future research, we would add more features to improve the accuracy of the classifiers due to the fact the two languages share some syntactic structure. We would create a feature for tracking the language structure, like subject-verb-object and object-verb-subject. This would allow us to track syntax errors as well as syntax similarity, giving us insight on distinguishing which ones used machine translation.

# 7. Future Work

In order to improve the level of accuracy we could include more data in the dataset because it's small and unbalanced. If we added to the dataset this would contribute to the size of the training set and further retrieve better results on the test set. After further research, we could balance the dataset by using Synthetic Minority Oversampling Technique (SMOTE). In addition, multilingual word embeddings would be useful since the dataset was for French and English. We only used Doc2Vec and Word2Vec, which only made the words vectors on a shallow level.

Features were created but not implemented for the Deep learning classification. The created features were VADER Sentiment Analysis, POS tag counts, Char count, TF-IDF,and Containment. These can be found in our correlation matrix in Figure 2.

A feature that may prove helpful in future reworkings of this model is the use of collocations. These are simply words that go together, in the sense that native speakers know when to pair certain words and when to not. For example, the phrase “keep a secret” is common in English. A synonym of “keep” is “save,” but native speakers know never to say “save a secret,” even if it is semantically correct at an individual word level. Google Translate depends on entries from its users in order to learn the rights and wrongs of these collocations, and sometimes it misses the mark, likely because of lack of user input. Providing the model with a corpus of English collocations could help point to these collocational errors and signify that a document has been created with Google Translate.

This model has the potential to be used as a node in a larger model meant for plagiarism detection. Along with features included in existing monolingual plagiarism detection models, whether a document has been translated or not can be an indicator as to whether it was plagiarized. If a document is determined to have been translated at some point, this could be a strong red flag, indicating plagiarism or at least some level of unoriginality.  We would be able to create a classifier chain where the ultimate goal is to determine whether something has been plagiarized or not. An idea of what this chain might look like is depicted in Figure 3.

# A close up of text on a white background Description automatically generated

# Figure 3: Classifier chain

This model could also be used to assist in Native Language Identification. Does the way in which someone writes in English tell us anything about their native language? This would serve greatly in forensic linguistics in determining authorship of a piece of text or in creating programs tailored for speakers/writers of English as a second language.

# References

Clough, P., Stevenson, M. Developing a corpus of plagiarized short answers. Lang Resources & Evaluation 45, 5–24 (2011). <https://doi.org/10.1007/s10579-009-9112-1>

Florence Reeder. 2001. Additional mt-eval references. Technical report, International Standards for Language Engineering, Evaluation Working Group.

http://isscowww.unige.ch/projects/isle/taxonomy2/

Jeremy Ferrero’s Cross-Language Dataset: Conference

https://github.com/FerreroJeremy/Cross-Language-Dataset/tree/master/dataset/documents/Conference\_papers

Kishore Papineni, Salim Roukos, Todd Ward, & Wei-Jing Zhu. (2002). “BLEU: a Method for Automatic Evaluation of Machine Translation”. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics , 311–318. Retrieved from https://www.aclweb.org/anthology/P02-1040.pdf

Multi-Class Plagiarism Features: Google Colab

<https://colab.research.google.com/drive/1opShhIwaxpT-UbfN3DHwA6FZqKfuY4DT>

Ratna, A. A. P., Purnamasari, P. D., Adhi, B. A., Ekadiyanto, F. A., Salman, M., Mardiyah, M., & Winata, D. J. (2017). Cross-Language Plagiarism Detection System Using Latent Semantic Analysis and Learning Vector Quantization. *Algorithms*, *10*(2), 69–83. doi: 10.3390/a10020069

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G.S., Hughes, M., & Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv, abs/1609.08144.