An Exploration of Models That Differentiate Translationese from Original English Texts

Pearl Hwang and Cynthia Lin

Abstract

In this project, we seek to build machine learning models that can differentiate between translated texts and non-translated texts in English using previously hypothesized characteristics of translated texts, which have been coined "translationese." We succeeded in building seven models, which are combinations of different classifiers and features that seem promising in their ability to distinguish between these two types of texts. Our best model is a simple neural network trained on TF-IDF features and reaches 86.7% accuracy, 82.4% precision, and 93.3% recall performance on our test set. Our success in building these models affirms the hypothesis that distinct features exist within translated texts, and also supports the fact that these features are apparent enough for simple machine learning models to distinguish between translated and non-translated texts.

1 Introduction

Texts in translation have often been thought of as significantly different from original texts, both on a stylistic level and a lexical level. Used by Gellerstam (1986) in a study into Swedish novels translated from English, the term "translationese" is now used in linguistics to describe the marks that the source language leaves on the target languages. For Gellerstam, these marks were often more lexical; for example, he found that translations tended use fewer colloquialisms and more foreign words.

Most of the work done in this area during the 90s built models from large corpora of original text in a non-English language and the translation of those texts into English (Baroni & Bernadini, 2006). These studies found that the most significant differences between translated texts and non-translated texts lie in lexical variety, lexical density, and sentence length, wherein translated texts had less lexical variety, tended to have a lower ratio of content to non-content words, and a higher mean sentence length (Lembersky et al., 2012). In our project, we first start differentiating through these characteristics, but we also approach the difference in translated texts and original texts in a few other ways.

More recent attempts to establish the existence of this so-called translationese have shifted to corpora that consist of bodies of translated texts and bodies of original texts that are not necessarily direct translations from each other (Baroni & Bernadini, 2006). This project also uses a dataset consisting of translated texts and original texts, which is described in greater detail in section 2.

After Gellerstam's investigation of Swedish translations of English novels, linguists have also experimented with the idea that all translated texts share a set of universal attributes, regardless of source language, an idea that was first presented through by Baker in 1993 and further elaborated on by Koppel & Ordan in 2011. Since our dataset includes articles translated from a mix of different languages (including Chinese, Arabic, French, Russian, and Spanish, among others), we will attempt to make use of these universal attributes.

Outside of the lexical features, some sources also suggest that translated texts have different syntactic features. For example, Baroni & Bernadini (2006) identified translations as containing an overrepresentation of pronouns, adverbs, infinitives, and other parts of speech.

In our project, we wish to explore various combinations of classifiers and features to see if we can produce a model that can reliably differentiate between translated and non-translated English texts. Through the process of using simple lexical and grammatical features in combination with different classifiers, we hope to be able to affirm the existence of translationese and bring to light some of the more apparent markers of this “language.”

2 Corpus Construction

Our corpus consists of news articles, mostly analysis or opinion as opposed to strictly factual reporting. Our body of non-translated texts was pulled from American news sources, such as the *New York Times*, *CNN*, the *Washington Post*, and also foreign news outlets written in English, such as *Arab News* and the *South China Morning Post*. The body of translated texts was taken mainly from *WatchingAmerica.com*, which is a website that provides manual English translations of articles published by various foreign sources in a mix of source languages, which include Chinese, Spanish, French, and Arabic, among other languages; we also sourced a few translated articles from *Taipei Times*.

Our training corpus includes 100 articles in translation and 83 articles written originally in English. We refined our models by performing tests on a validation set consisting of 25% of articles taken randomly from this corpus, and the final results shown in this paper are the results of running our models on a test set of 30 new articles, half of which are translations.

In order to make the articles more comparable and to attempt to limit the lexicon, we focused on articles that involved discussion of International Relations, especially as pertaining to the U.S. and another country. To avoid any confounding factors that may arise from things such as changes in writing conventions or differences in prominent events or figures of discussion, the texts in our corpus were all published within the past year. Keeping the topic and genre of our corpus uniform was important, since it allows us to run our models on a relatively smaller dataset with a variety of source languages, but it also has some disadvantages. In particular, a smaller, more focused corpus might mean that some of the models we develop only work well for texts that also fit under the topics that we have chosen.

Due to time constraints and our interest in the presences of universal markers within translated texts, regardless of source language, we will not be taking the different source languages into consideration in this project.

3 Model Components

In order to build the most effective model and find features that best reflect the differences between original English and translationese English, we experimented with a combination of features and classifier types. The features and classifiers we used to build each of our seven final models are detailed below.

3.1 Features

The first features that we experimented with were n-grams. However, we found that our data was generally not large enough or extensive enough to be able to support bigrams or trigrams, and a large fraction of unique n-grams would only occur once or not at all, giving rise to model parameters that are not readily generalizable. Thus, we decided to simply use unigrams, under the assumption that individual words within the text are independent of the other words—the bag-of-words assumption, henceforth referred to as BOW in this paper. Though it is unfortunate that bigrams and trigrams appeared to be less effective with our corpus, using BOW as a feature allows us to explore the hypothesis that translated texts have significantly lower lexical variety than texts written originally in English.

The next features we used was the term frequency-inverse document frequency (TF-IDF) measure, which creates document vectors with metrics containing an entry for each word within the document, representing how relevant each word is to the content of each document. In theory, this feature could help bring out the lexical density, or ratio of content words to non-content words. In our implementation, we used the 5000 most common word bigrams or trigrams for this feature.

The last features that we implemented were a custom set of features that includes average sentence length, average word length, and percentage of stop words within a document. These features each correspond with previous research regarding sentence length, lexical diversity, and lexical density. For our custom features, we tokenized sentences and words using NLTK’s `sent_tokenize()` and `word_tokenize()`.

In engineering our features (average sentence length, average word length, unigrams under the BOW assumption, and bi-/trigrams for the TF-

IDF), we did not remove any stop words, since we found that there was a significant difference between the percentage of stop words between translated texts and non-translated texts, and believed that accounting for the stop words in our features would help capture that difference. This decision is also supported by research done previously that posits that translated texts have a lower lexical density than non-translated texts.

In some of the models, we also experimented with using these features with different word tokens. For example, we used lemmatized words as tokens as the features for one of our models to help mitigate the sparse data problem. In another model, we used POS tokens with TF-IDF to gain further insight on the grammatical characteristics of translationese. A more specific overview of the different tokens used with each model are explained further on in the paper.

3.2 Machine Learning Classifiers

To implement the above features, we tried a series of machine-learning models: naïve Bayes, logistic regression, support vector machine, and simple neural networks.

We implemented our naïve Bayes (NB) models using scikit-learn's Multinomial NB model, with the default smoothing value of 1.0. In the results, we show two models that use the NB classifier, one with the custom and BOW features, and the other using bi-/trigram TF-IDF features.

We implemented three logistic regression (LR) models (using scikit-learn), with three sets of features: custom and BOW features, BOW features with spaCy POS tagger, and unigram TF-IDF features with spaCy lemmatizer.

We also explored the usage of support vector machines (SVM), which were used successfully by Baroni & Bernadini to differentiate between translated and non-translated texts in an Italian corpus. (We also used scikit-learn to implement this model.)

Lastly, we experimented briefly with different neural networks, including convolutional, recurrent, and a simple three-layer network The most promising model we found among these network architectures using our set of features was with the simple neural network (SNN).

We explore results from seven of our best models, which were constructed using a combination of one of these classifiers and one or more of the feature sets described in section 3.1.

4 Implementation of Models

Through experimenting with different combinations of features and models, as well as different tokenizers for some of the features, we came up with seven models that were relatively successful in differentiating between the two types of texts, shown in Table 1.

Model ID	Classifier	Features	Tokens/ Tokenizers
1	NB	custom, BOW	token_pattern =r'\w{1,}'
2	NB	2-/3-gram TF-IDF	token_pattern =r'\w{1,}'
3	LR	custom, BOW	token_pattern =r'\w{1,}'
4	LR	BOW (spaCy POS tags)	spaCy
5	LR	unigram TF-IDF (spaCy word lemmas)	spaCy
6	SVM	custom, BOW	token_pattern =r'\w{1,}'
7	SNN	2-/3-gram TF-IDF	token_pattern =r'\w{1,}'

Table 1: Classifiers and features of our final models.

We first started by exploring different combinations of features with NB classifiers, and ended up with two models, one with a combination of our custom features and the BOW feature, and the other using bi-/trigram TF-IDF vectors as features. As seen in Table 2, the first NB model was more effective than the first with regards to accuracy and precision, but not recall.

Next, we experimented with LR models, and used three different combinations of features to see if we could improve upon the earlier models. The first LR model uses custom and BOW features. The second model uses spaCy's tokenizer and POS tagger to create BOW features using POS tags (instead of lexical tokens). The third model uses spaCy's tokenizer and word lemmatizer to create the unigram TF-IDF feature vectors. In Table 2, we can see that the model using custom and BOW features performed best out of the three LR models, while the performance of the other two LR models are about the same.

The next model is SVM, which works by constructing a hyperplane that maximizes the difference between a positive output and a negative output. Depending on which side of the hyperplane a document falls on, the model classifies it as translated or original English. We used a combination of our custom features and BOW unigrams to train this model. The model was not significantly worse than any of the previous models, but also did not perform much better.

The final model that we ended up with uses a simple neural network trained on bi-/trigram TF-IDF feature vectors. The NN that we implemented contains three layers: an input layer (size = length of feature vectors), a hidden layer (size = 180), and an output layer (size = 1). We trained the model for 150 epochs. For our predictions, we use 0.5 as a threshold in determining whether a text is classified as a translation or an original English document. This model yielded the highest precision and accuracy and relatively high recall, making it the most effective model.

5 Results and Discussion

5.1 Results

After refining our models on a validation set, which was a subset of or corpus (100 translations and 83 original English texts), we ran the seven models on a test set of 15 translated articles and 15 non-translated articles. The metrics from running the models on the test set are shown in Table 2.

ID	Accuracy	Precision	Recall	F1
1	0.767	0.722	0.867	0.788
2	0.733	0.652	1.000	0.789
3	0.800	0.765	0.867	0.813
4	0.767	0.700	0.933	0.800
5	0.767	0.681	1.000	0.811
6	0.733	0.705	0.800	0.750
7	0.867	0.824	0.933	0.875

Table 2: Metrics for each of the seven models, when run on the test set.

Most notably, the model that performed best is the SNN using TF-IDF features. We were able to achieve 86.7% accuracy, 82.4% precision and 93.0% recall on the test set, as shown in row 7 of Table 2. The success of this model supports the idea that lexical variety and lexical density are two markers of translationese, regardless of source language.

There are a few other points from our results that are worthy of note.

The first is our attempt to use POS tags as BOW features for model 4, as supported by the results of Baroni & Bernadini, where they found that pronouns and adverbs served as cues for texts in translation. While the model was not the worst of the seven models, it does not support the hypothesis that markers in translation all come from syntactic and grammatical properties of the text. Perhaps further experimentation on building models that use both POS tags and lexical features would yield better results. The confusion plot of results from model 4 (LR using BOW with POS tags) are shown in Figure 1.

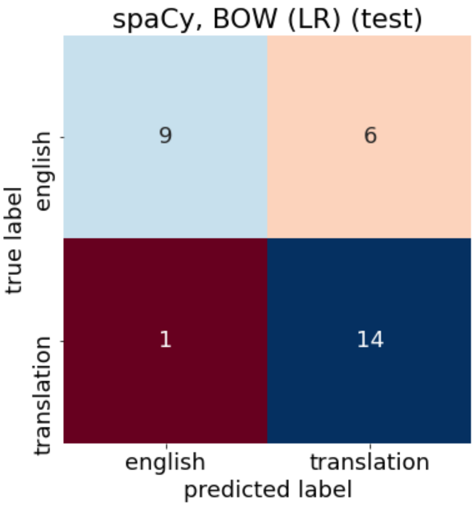

Figure 2: Confusion plot for model 4 (LR using spaCy POS tags as BOW features).

The second point of note is that our best model (model 7) uses TF-IDF as a feature, yet the other two models that also use TF-IDF as features appeared less successful in differentiating between the two types of texts. These other models (2 and 5) use NB classifier and LR

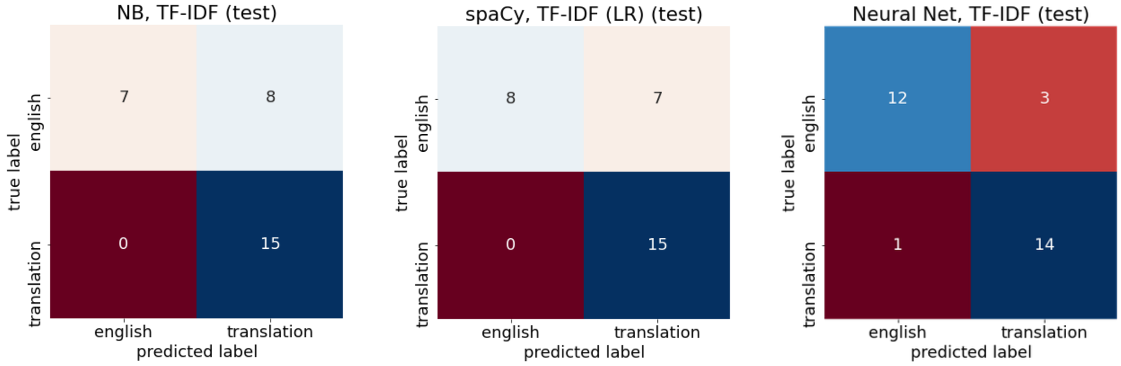

Figure 3: Confusion plots for the test set from models 2, 5, and 7, respectively.

regression classifiers. The confusion plots for models 2, 5, and 7 are shown in Figure 3.

While the NB and LR models using TF-IDF features do not misclassify translations as original English texts, they also do not seem to be able to reliably identify original English texts as English, classifying a considerable proportion of original English texts as translations. (We also note that during our experimentation, we observed that models (not shown here) using unigram TF-IDF features tokenized under token_pattern= r'\w{1,}' performed significantly worse than the models we describe in this paper, leading us to choose bi- and trigrams TF-IDF features instead; but the preprocessing step of lemmatizing words using spaCy prior to constructing TF-IDF features in the case of model 5 resulted in better performance, likely due to a mitigation of the sparse data problem.)

In contrast to models 2 and 5, while the SNN (model 7) also tended to assign texts as translations, it was much better at differentiating between the two types of texts.

This discrepancy can likely be attributed to the different ways in which the models perform supervised learning. While NB and LR classifiers are linear models, the SNN can capture non-linear relationships among the inputs.

5.2 Other Approaches

Other models that we briefly explored on the validation set include a convolutional neural network (CNN) and a recurrent neural network (an LSTM) on GloVe embeddings.

However, we did not explore either of these models in depth due to time constraints as well as initial non-so-promising results on the validation set. We also did not experiment much with network parameters such as size, number, and types of layers or activation functions. Confusion matrices of these two models when run on the validation set are shown in figures 4 and 5. These results should not be taken to mean that these models do not perform well on the task at hand; we have simply included them here for the sake of completeness, and as potential areas for further exploration.

5.3 Limitations & Further Directions

Several factors may pose limitations on the results from our project. For one, our corpus is small in size, and narrow in regard to the themes of the texts, which limits the generalizability of our models to texts on other topics. Although we did try to source articles that discuss topics that are generally similar to reduce the likelihood that our models learns distinctions based on content/theme as opposed to linguistic features, we did not perform any sort of normalization to account for any possible discrepancies that may have been present between the translations and the English originals. One possible thing future projects may do is to remove named entities such as names of people and places from the corpus, or somehow normalize them so that we can be sure that the model is not simply distinguishing between the content of translated and non-translated texts.

Additionally, as mentioned in section 5.2, we did not explore more complicated types of neural networks in any depth. It is possible that these networks are in fact capable of performing the task at hand much more proficiently than any of the seven models we explored in this project.

6 Conclusions

In this project, we explored implementations of models that use supervised machine learning techniques to differentiate between translated and non-translated texts.

By experimenting with various classifiers and features, we were able to create seven models that were relatively successful in this differentiation, and our best model using a simple neural network and TF-IDF features achieved an accuracy of 86.7%.

These results are in line with past work that have affirmed the existence of translationese. In particular, the existence of some kind of universal translationese that exists regardless of what the source language is.

While we confirmed that the usage of different POS are in fact a marker of translationese, we found that within simple models, properties such as lexical variety and lexical density are often sufficient to distinguish between translated texts and non-translated texts. One of the more striking results from our project is the fact that model 4, which we trained using POS tags, performed on par with the other models that use lexical tokens: Because this model uses POS tags, we know that it is not classifying the texts based on content, but that it is making distinctions based on some other linguistic characteristics of translated versus non-translated texts.

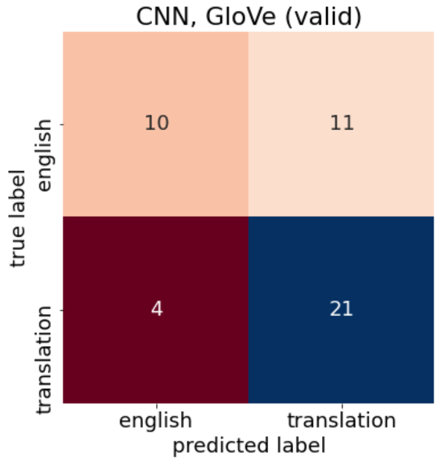

Figure 4: Confusion plot for a CNN on the validation set.

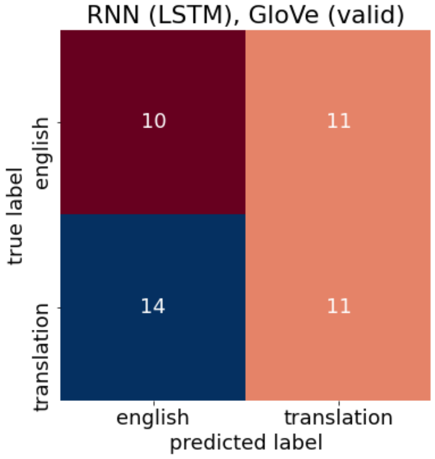

Figure 5: Confusion plot for an LSTM on the validation set.

References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In Gill Francis Mona Baker and Elena Tognini Bonelli, editors, *Text and technology: in honour of John Sinclair*, pages 233-252. John Benjamins, Amsterdam.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing,* 21(3):259-274.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English, in Lars Wollin & Hans Lindquist (eds.), *Translation Studies in Scandinavia* (88-95). Lund: CWK Gleerup

Moshe Koppel and Noam Ordan. 2011. Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pages 1318–1326.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language Models for Machine Translation. *Association for Computation Linguistics (Volume 38, No. 4),* pages 799-825.