Distinguishing Between Original and Translated Texts

Elad Tolochinsky, Ohad Mosafi

September 14, 2017

1 Introduction

This work aims to distinguish between original texts and translated texts. Given a document written in some language, we wish to accurately say if the document was originally written in that language of if it was translated to it. To distinguish such texts we use a linguistic concept called translationese. Translated texts, in any language, can be considered a dialect oft his language. We cal this dialect translationese Linguistic research has proposed several universal properties of translationese. We use this properties of translationese in conjunction with machine learning techniques to distinguish between translated texts and non-translated (original) texts. Previous work [1], on this subjects have tried to gauge which properties of translationese can be successfully utilized to classify original and translated texts. Our work reproduce the results achieved by [1] on a different corpus - the UN parallel corpus. The first part of this work is an automated derivation of the corpus. We derived five bilingual parallel corpora, from English to any other official UN language (French, Spanish, Russian, Arabic and Chinese).

2 Derivation of the Corpora

The base of this work is the UN parallel corpus which is described at length at [2]. The corpus is structured in a directory hierarchy, each language has a directory which holds the documents in that language. The documents are stored in a directory tree inside the appropriate language folder in a way that the relative path of a document inside a specific language directory is the same for all language directories, for example the file *add_1.xml* has a French version at the path *\fr\1990\trans\wp_29\1999\14\add_1.xml* and an English version at the path *\en\1990\trans\wp_29\1999\14\add_1.xml.* Every language pair has an additional directory which contains link files. The link files defines the translation direction of two documents and they reside at the same relative path as the documents, thus the link file of *add_1.xml* is located at *\fr_en\1990\trans\wp_29\1999\14\add_1.lnk.* To derive a bilingual parallel corpus, we pick a language and traverse it's directory along side the the English directory (Note that we must start the derivation from the non-English language, due to he fact that the link file matches non English sentences to their English translation but not the opposite). For each sentence, we use the link file to determine if the sentence is in original or translated. The outputs of the process are 3 files for each language: a file containing English sentences, a file in which each line is the translation of the appropriate line in the English file and a third file which specifies the original language of each line. While processing the documents, many documents were filtered out of the corpus due to various reasons, such as: documents that had no corresponding document in English, documents that did not specify their source language (either in English or in the source language), documents which source language did not match the current language, etc. After obtaining the valid documents we filtered out invalid sentences which include sentences which language tag was different then the language tag of the file and sentences which has no destination at the link file. Results of the derivation are summarized at 1

	French-English	Spanish-English	Russian-English	Arabic-English	Chinese-English
Initial number of protocols	185,800	130,275	138,168	116,751	95,946
Number of valid protocols in every language and percentage (from the initial amount)	4,567 / 51,807 (2.8% / 32.6%)	1,957 / 36,603 (1.5% / 28%)	706 / 25,218 (0.5% / 18%)	1,690 / 29,053 (1.4% / 24.8%)	10 / 8,290 (0.01% 8.6%)
Number of valid sentences	8,926,298	6,638,552	3,740,232	4,175,839	1,689,598
Number of valid sentences in every language ad their percentage	773,276 / 8,153,022 (6.5% / 68.8%)	447,445 / 6,191,107 (5% / 70%)	107,737 / 3,632,495 (2% / 73%)	88,263 / 4,087,576 (1.4% / 68.5)	4,768 / 1,684,830 (0.2% / 73%)

Table 1: Corpora derivation results

3 Classification

The methods we employ in this work are machine learning and specifically classification. Generally, given a set of labeled vectors $X \times Y$ where $X \subset \mathbb{R}^d$ and $Y = \{0, 1\}$ which are drawn from some distribution $\mathcal{D}$. A classification function is a function $f : \mathbb{R}^d \to Y$ such that with high probability $f(x) = y$ for every $(x, y) \in D$. The process of computing such a function is called *training* and is based upon feeding the learning algorithm with know examples from which it can "learn". In this work we employ well know learning algorithms such as SVM and logistic regression. We will use the implementation of Pyrhon's Scikit package. To sum up, in order for us to distinguish between original texts and translations, we transform a document to a vector, we label that vector according to the class of the document (translated or original) and we then proceed to train the appropriate machine until we obtain a function which distinguishes between translated and original.

4 From Documents to Vectors

In the previous section we described a method in which we can distinguish between different classes of vectors. We are left with the problem of representing a document by a multi dimensional numeric vector. There are many ways to represent a document as a vector, perhaps the simplest is called 'bag-of-words' in which every document is represented by a vector of counters, every entry in the vector represent the number of occurrences of a corpus word in this specific document. However such simple representation may not be helpful in distinguishing between original and translated texts. That is where we employ translationese. We use the hypothesized universal properties of translationese to derive a numeric representation of documents. As this properties represent the dialect of translated properties, we intuitively expect that will produce accurate classification results. The work at [1] have compared many of the universal properties of the translationese and have found the ones that are most effective for distinguishing original from translated. We used the following properties

- Function words
- POS trigrams
- POS bigrams

In order to obtain a dataset for training our learning algorithms we tokenize the text files and add part of speech tagging, we then break up a file to chunks of about 2000 tokens. Each chunk will be transformed into a vector according to the chosen property

	Function words	Trigrams	Bigrams
Classifying all languages Each chunk belong to one language	83%	85%	86.65%
French-English	86.5%	87%	88.28%
Spanish-English	86%	86.8%	88.5%
Russian-English	91.38%	92.48%	92.86%
Arabic-English	89.71%	93.6%	94.5%
All languages chunks randomly distributed across languages	90.53	92.05%	93.21%

Table 2: Classification Results

- Function words - Each chunk is transformed to a vector, where each entry in the vector represents the frequency of a function word in the chunk. We then normalize this quantity by multiplying it by $\frac{n}{2000}$ where n is the size of the chunk.
- POS trigrams - Each chunk is transformed to a vector, where each entry in the vector is the number of occurrences of a POS trigram in the chunk from a list of 400 top trigrams.
- POS bigrams - Each chunk is transformed to a vector, where each entry in the vector is the number of occurrences of a POS bigram in the chunk from a list of 400 top bigrams.

5 Experiments and Results

After obtaining the bilingual parallel corpora we divided the sentences at the English file of the corpus to tokens and add POS tags. The tokenizer we used is NLTK's tweet tokenizer and POS tagging was done with OpenNLP. After obtaining tokenized and tagged documents we start dividing each document to chunks of about 2000 tokens. Each chunk is comprised from sentences that are either all original or all translated. We then transform each chunk to a vector using one of the methods described above and label the vector accordingly. From this process we obtain a set of vectors that was derived from each language. We merge all of this vectors and feed them to a classifier. Afterwards we classify each language separately to see in which languages the distinction from original to speech is easier. Finally we reproduced the chunks, only this time we shuffled all of the sentences from all of the languages together and then divided them to chunks. The results are depicted at 2. In all tests we omitted the samples from the Chinese corpus since it was too small. All data sets were balanced, so the baseline is 50%.

The comparison between the classification results is depicted at figure 1.

We also tested the effects of changing the amount of tokens in a chunk. We produced various datasets from different sized chunks. Each dataset was balanced by taking all of the translated chunks and randomly choosing the same number of original chunks before classification. The results are depicted at2. The amount of chunks per chunk size can be seen at

6 Conclusions

The results of our work raise several conclusions: first, the corpora we derived are valid, for if they were not so we would not have been able to classify with high accuracy . Furthermore we can see that the different languages are classified more easily than others, this is somewhat intuitive, since French is much more similar to English then Russian. Last we see that building chunks out of sentences that were randomly chosen across all languages yield much better results, this is reasonable, since choosing sentences from all languages negates

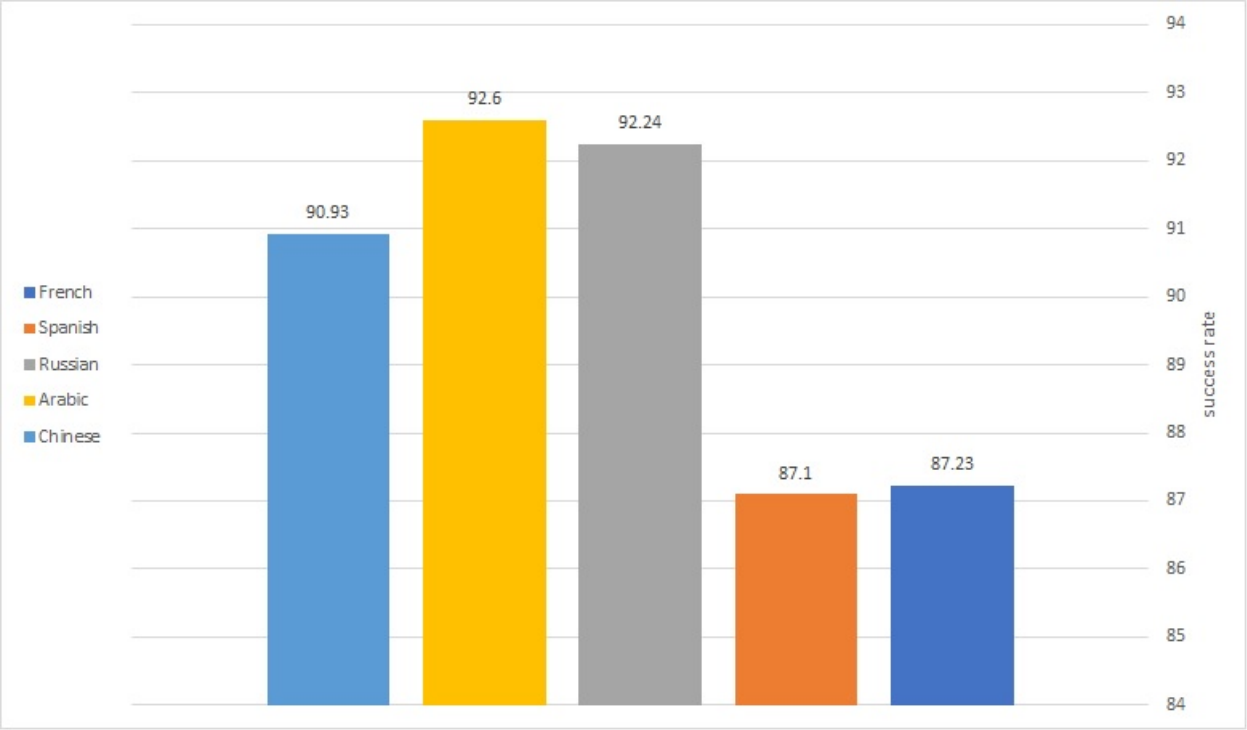

Figure 1: Comparing classification results between languages

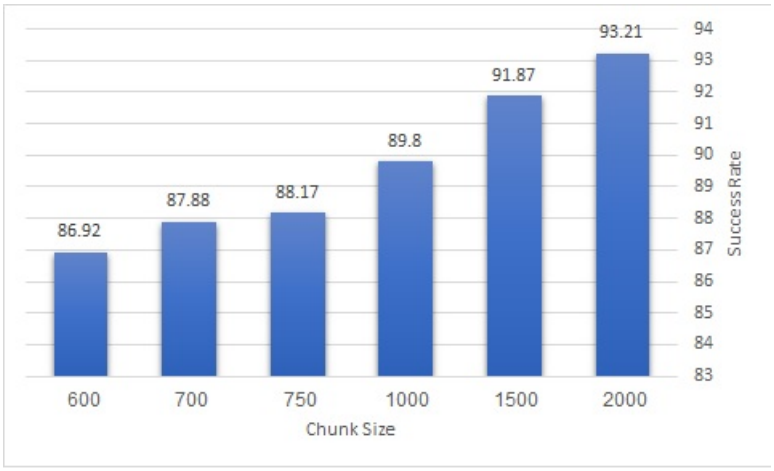

Figure 2: Classification results for various amounts of tokens in a chunk. Tests were performed using bigrams as features.

Chunk Size	Number of samples
2000	36,632
1500	48,422
1000	71,412
750	93,664
700	99,878
600	115,210

Table 3: Number of samples per chunk size

noises (personal speech style, subject, language of origin, etc.) and we are left with a "one dimensional" data - translate or origin.

References

[1] Vered Volansky, Noam Ordan, and Shuly Wintner. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, 2013.

[2] Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The united nations parallel corpus v1. 0. In *LREC*, 2016.