
# RESEARCH PAPER ON FOREIGN LANGUAGE DOCUMENT PRIORITIZATION METHODOLOGY USING MACHINE TRANSLATED TEXT

Author: Patrick Walsh
Data Scientist, Arabic Language Analyst
SYSCOM, Inc.

# Contents

# Introduction

This paper will describe research that SYSCOM is currently conducting in the area of Natural Language Processing and Machine Translation. The section immediately following the introduction identifies the business problem along with the solution, as well as a scenario use-case to illustrate the solution in more detail. After that, the paper will introduce the hypothesis upon which the business solution is based, and describe the process that is currently underway to substantiate that hypothesis.

## *Problem*

The organization needs to be able to analyze a high volume of text using time and resources. Much or all the text is in a language other than English, and the organization only has a few qualified linguists who can understand the foreign language text. Currently, the linguists must identify documents of interest by manually scanning the text, i.e. visually reading segments of text to characterize the documents and flag documents that they deem to be of interest, while discarding documents that are not of interest. This manual scanning process is labor-intensive, time-consuming, prone to error, and exhausting for the human linguists who spend a significant amount of time and energy doing it. Once the scanning process is complete and the documents of interest have been identified, the linguists' job has only just begun, because now they must undergo the even more labor-intensive process of translating the foreign language documents into English.

## *Solution*

The organization can save precious time and resources by automating part of the scanning process that is currently being done by humans. Using a combination of Machine Translation and Natural Language Processing (NLP) annotators, computers will characterize and prioritize documents of interest, freeing up the linguists' time and energy to focus on the more fruitful task of translating the documents. To accomplish this, the entire corpus must undergo an automated Machine Translation so that all texts are in English. Next, an NLP annotator analyzes the translated text, identifying key terms of interest and assigning a score to each document to prioritize which documents are likely to yield the most value. To illustrate this, here is a hypothetical scenario:

# Scenario Use-Case

A team of 10 people are working for an organization that is tasked with analyzing socioeconomic relations between Country A and Country B. They must evaluate thousands or even 10s of thousands of documents per day to identify information that is of value. Only five of the team members are qualified Arabic linguists who can understand the foreign language text that they work in. One of the linguists also understands French and a little bit of Persian Farsi. The majority of the documents they come across are in Arabic, with some French, Farsi, and occasionally Turkish documents. Since no one on the team speaks Turkish, these documents are immediately discarded, creating a blind spot in their collection. It is the job of these five linguists to visually scan thousands of documents to identify content that pertains to socioeconomic relations between Country A and Country B. Because of time constraints of the large volume of content, it is impossible for the linguists to do an in-depth analysis of all the documents, so

they must use their best judgement to determine which documents are likely to yield the best results and scan those documents first. Each linguist may come across 3,000 documents per day, scan 500 of them, and of that 500, identify 25 documents that need to be translated. By the time the linguist is done scanning and has identified the 25 documents of interest, he or she is tired, having spent the first 4 hours of the day reading segments of foreign text, making a series of judgement calls, and making characterizations about hundreds of documents and thousands of lines of text to decide whether or not what they are looking at is of value. Once the scanning process is complete and the 25 documents of interest have been identified, the linguist still has to translate these documents so that other analysts who speak only English can then assess the information and decide next steps. Not only is this process time-consuming and mentally taxing for the linguists, it almost guarantees that the organization is missing key insights due to the limited time and resources they possess and the high volume of content with which they are faced. Of the 3,000 documents that the linguist had access to, he or she only had time to scan 500 of them, which means that the remaining 2,500 documents do not even get a cursory glance to determine their potential value. On top of that, some of the documents were in Farsi, French, and Turkish, and only one of the linguists on the team speaks Farsi and French and none of them speak Turkish. This means that the Turkish documents had to be immediately discarded and the one linguist who speaks French and Farsi did not have time to scan all the French and Farsi documents. Consequently, the organization is only visualizing tiny fragments of the available information at its fingertips while most of the data remains in the dark.

Using SYSCOM's foreign language document prioritization method, linguists will spend less time scanning and more time translating. Using the same scenario as before, the linguist will Machine Translate all 3,000 documents into English, including the ones in Farsi, French, and Turkish. Then, an NLP annotator specifically designed to work in the team's area of interest will analyze the documents, creating annotations of socioeconomic terms. The annotator does this by analyzing the content and tagging key words and phrases. In this context, this could be terms such as "economy," "international relations," "domestic exports," "economic policy," "trade deficit," "production," "legislature," "minister of finance," "inflation," and so on. The annotator then counts the frequency and density of these key terms and assigns each document a score based on how closely it aligns with the desired category. The higher a document scores in the socioeconomic category the higher a priority it will be given. In this scenario, the 3,000 documents will be prioritized in terms of how their content scored on socioeconomic issues. The documents that scored the highest will be given top priority for the linguist to scan first. Of the 3,000 documents, perhaps only 600 of them are given a marginal score in socioeconomic content, and of those 600, 75 of them score exceptionally high in that category. The linguist will start scanning the 75 first, moving on to the other 600 as needed and as time permits. Before ever setting eyes on any of the documents, the linguist will have a sense of what they are looking at and will be able to prioritize their scanning efforts. This helps reduce wasted time and effort that would otherwise be spent scanning junk content that gives no value to the organization's efforts. Since the annotator analyzes content that has been Machine Translated into English, this document prioritization method applies universally to all documents regardless of which language they were originally written in. As long as a Machine Translator exists for that language and can translate from that language into English, the annotator will be able to create annotations and assign scores to the documents. This gives the team greater visibility of documents in languages that they have limited or no visibility of currently. The annotator will characterize and prioritize not only the Arabic documents, but also the Farsi, French, and even the Turkish documents. If a document in Turkish scores particularly high in a certain category, the team

could pass it off to another team that has Turkish language capabilities, or perhaps bring in Turkish language support to perform a manual translation if necessary.

Now that the problem has been identified and the business solution has been outlined, it is time to consider the methodology in more detail. The following section will explain how the method was developed and the research used to validate its effectiveness.

# The Hypothesis

The method outlined in the above scenario is based on the question of Machine Translation and its effectiveness in NLP. While Machine Translation may fall short of rendering perfect translations that correctly identify and interpret the subtle nuance and cultural and societal characteristics of text in its native language with 100% accuracy, SYSCOM wanted to demonstrate that Machine Translated text can still provide valuable insights for NLP annotators. This is predicated on the fact that annotators deal with natural language primarily at the *atomic level,* that is, with individual words and phrases. Machine Translation is able to achieve a high degree of accuracy at the atomic level. Where Machine Translation lacks (and where good human translators excel) is in *comprehension* of the foreign language text. Having said that, even when human translators are able to achieve superior comprehension, the subtlety and complexity of language is oftentimes lost when translating from one language into another. To fully appreciate Beowulf, one must read it in its original Old English. To fully appreciate the Qu'ran, no translation can compare to reading it in its original Arabic. Fortunately for our purposes, it is not necessary to achieve that level of comprehension. NLP annotators are operating at the atomic level, the same level at which Machine Translation excels. Based on this idea, it follows that NLP annotators should be able to achieve similar results analyzing Machine Translated documents as they can analyzing documents in their native language. That is the premise behind this research paper, but it is not enough to merely develop a hypothesis, there must be evidence to back it up. The proceeding sections describe how SYSCOM developed a Machine Translation/NLP annotator use-case and built the case for using this method in real-world business solutions.

# Testing the Hypothesis

To test the hypothesis, SYSCOM gathered documents and designed annotators that could analyze the documents. The documents consisted of reports published by the World Health Organization (WHO), research papers written by Georgetown University's Center for International and Regional Studies (CIRS), and a Health and a Safety requirements manual from the U.S. Army Corps of Engineers. All the documents were written in both English and Arabic. The English corpus added up to about 590,000 words, and the Arabic equivalent added up to around 589,000 words. Each corpus contained the same documents written in each language. Two NLP annotators were created in IBM Watson Explorer Content Analytics Studio, and the annotators were deployed to IBM Watson Explorer Content Analytics where the corpora were analyzed in Watson Explorer's Content Miner. The first annotator was designed to

analyze content in English, and consisted of custom-made dictionaries and parsing rules to create annotations in the documents. Dictionaries are collections of words and phrases, also referred to as lexical databases. Examples of simple dictionaries are days of the week, months of the year, or a list of car manufacturers. The annotator references the dictionary to check for instances of those words or phrases and then creates annotations based off the dictionary entries. Parsing rules are syntactic rules designed to look for patterns in text and create annotations based off those patterns. An example of a parsing rule would be to find two digits, followed by a forward slash "/" or period ".", two more digits, another forward slash or period, and finally two or four more digits. This pattern will find dates in text in the mm/dd/yy or mm/dd/yyyy format. Dictionaries can be used in conjunction with parsing rules. For example, to find dates in the format of the dictionary month, followed by the day, a comma "," and then the year, you could create a parsing rule that references a months dictionary, then looks for the pattern of 1-2 digits, followed by a comma, followed by four digits. This combination of dictionary and parsing rule would find dates in the format of "January 5, 2013."

## *Initial Results*

For this test, SYSCOM created a series of dictionaries, parsing rules, and dictionary + parsing rule combinations to annotate the documents. The dictionaries and parsing rules were written in English for the English annotator, and Arabic for the Arabic annotator. For the initial phase of the test, SYSCOM compared how each annotator performed when tasked with creating annotations for specific categories. The categories are as follows:

-Cities

-College Degrees

-Countries

-Dates

-Newspapers

-Political Parties

-Universities

-United Nations Organizations

-Websites

-World Leaders

The results were analyzed in Watson Explorer's Content Miner, and the number of annotations created for each category were counted. Next, the annotations were examined in more detail to determine false positives, i.e. how many of the annotations were falsely attributed to a given category. Using these numbers, both the English and Arabic annotators were assigned accuracy scores and compared. Overall, the English annotator achieved an average accuracy rating of 95.63%, that is, 95.63% of the annotations created by the annotator were accurate, while the remaining 4.37% were incorrectly annotated. The Arabic annotator performed slightly better, with an average 97.77% accuracy rating.

## *Machine Translation*

The next step was to add Machine Translated documents into the test to see how they performed. For this example, the Arabic documents were Machine Translated using Google Translate. The translated documents were then analyzed by the English annotator, creating the same annotations for the same categories as outlined above. The results were analyzed and the English annotator achieved an average accuracy rating of 96.51%, slightly below the Arabic annotator's results, but slightly better than the English annotator's performance against the original English documents.

## *Document Prioritization*

The initial results looked promising, but SYSCOM wanted to take the test to the next level to see if NLP annotators could be used to characterize corpora and create subsets of documents by assigning density scores to documents based on category. For this phase of the test, three new categories were created: Economic Terms, Political Terms, and Healthcare Terms. Dictionaries were compiled containing lists of words and phrases from each of these three categories. The terms in each of these categories were meant to classify the content of the documents based on the frequency with which they occurred within a given document. If a document had a high number of words and phrases pertaining to Healthcare, such as "Health," "universal coverage," "doctor," "hospital," "patient," "disease," and so on, then that document would receive a higher score in the Healthcare category. The raw number of annotations created from each of these categories were recorded and measured against the total number of words in each document. This gave us a density score, i.e. how many terms out of the total number of words in this document were classified according to one of these categories.

During this phase of the test, one additional corpus was added to the list, that being the Arabic documents that had been Machine Translated into English using Watson Language Translator. This gave us an additional dataset to work with to compare results. So now, SYSCOM was using two annotators (one English and one Arabic) and four corpora. (one in English, one in Arabic, one Machine Translated from Arabic into English using Google Translate, and one Machine Translated from Arabic into English using Watson Language Translator.)

The Arabic annotator was deployed to the Arabic corpus, and the English annotator was deployed separately to each of the three corpora in English. The annotations were recorded and analyzed. There is ongoing analysis of the results, but the initial results look good. Each of the four corpora produced a similar number of annotations across a similar proportion of the documents. The annotations from each category, (Economic Terms, Political Terms, and Healthcare Terms) were counted and the results for each document, across every corpus was analyzed. This image shows the total density scores that each corpus produced across the three categories:

| | Economy | Political | Healthcare |
|---|---|---|---|
| English | 0.17 | 0.34 | 0.18 |
| Arabic | 0.14 | 0.28 | 0.2 |
| Google | 0.16 | 0.31 | 0.17 |
| Watson | 0.15 | 0.31 | 0.15 |

The first column lists the corpus, and the proceeding three columns give a density score for each corpus according to category. As can be seen, the corpus of Arabic documents translated into English using Google Translate (Google) had a density score of 0.16 for Economic Terms. What this means is that for the entire corpus, 0.16% of the words (1,029 out of 645,722) consisted of Economic Terms. The density scores are fairly consistent across all four corpora, as we had hoped to find. To get a better idea of what these numbers mean, one additional formula was used that took into account the density score of each category relative to the other two categories. This provides a breakdown of the content by percentage, showing that out of all the annotations created by the annotator, a certain percentage belonged to each of the three categories. This can be seen in the following graphic:

| | Economy | Political | Healthcare |
|---|---|---|---|
| English | 24.64% | 49.28% | 26.09% |
| Arabic | 22.58% | 45.16% | 32.26% |
| Google | 25.00% | 48.44% | 26.56% |
| Watson | 24.59% | 50.82% | 24.59% |

This makes it easy to visualize the results quickly. Roughly half of the annotated content of the total corpus has been classified as Political, while the remaining half is fairly evenly split between Economic and Healthcare Terms. This pattern holds true across each of the four corpora. There is some variance when evaluating the individual documents, but the general breakdown is similar.

To find the variance that each annotator had when annotating each of the corpora, a Standard Deviation formula was applied to the results, giving the Mean and Standard Deviation score that each corpus had for each of the three categories. We used a traditional Standard Deviation formula:

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

*where S = the standard deviation of a sample,*
*Σ means "sum of,"*
*X = each value in the data set,*
*$\overline{X}$ = mean of all values in the data set,*
*N = number of values in the data set.*

Using this formula we found the average or mean of the density scores for each corpus in each category, then calculated the Standard Deviation or variance of each density score. The mean was calculated by adding the sum total of each of the density scores for each document, then dividing that sum by the number of documents. This gave the English corpus a mean density of 0.30% for Economic Terms, the Arabic corpus a mean density of 0.25% for Economic Terms, and so on. Next, we used the Standard

Deviation formula to calculate the variance of each density score. As can be seen, the English corpus had a mostly higher variance overall, with +/- 0.18% for Economic Terms, +/- 0.31% for Political Terms, and +/- 0.09% for Healthcare Terms. The Watson corpus had the lowest variance overall, with +/- 0.13%, 0.25%, and 0.07%, respectively:

| | Economic | | Political | | Healthcare | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| English | 0.3 | 0.18 | 0.61 | 0.31 | 0.26 | 0.09 |
| Arabic | 0.25 | 0.14 | 0.54 | 0.27 | 0.34 | 0.13 |
| Google | 0.26 | 0.14 | 0.54 | 0.28 | 0.24 | 0.08 |
| Watson | 0.26 | 0.13 | 0.52 | 0.25 | 0.22 | 0.07 |

# Conclusion

To summarize, SYSCOM wanted to demonstrate that NLP annotators could achieve similar results against Machine Translated text as text written in its original language. Preliminary analysis of the results show at least marginal success in the annotator's ability to create annotations and classify documents according to their content. There is ongoing evaluation of the results and continued effort to substantiate the hypothesis set forth in this paper, but we are optimistic about our endeavor.