Language as Data
Final Project: Comparative Data Analysis Report

Mohammed Majeed

December 2, 2023

Introduction

In a world where language and cultural nuances are vital aspects of communication, the assertion that "translated talks sound different" often incites debate. On one hand, proponents of translation assert that when executed with skill, the essence and emotional tone of the original conversation can be preserved. They advocate the view that adept translators are capable of effectively bridging linguistic gaps, thereby reducing significant disparities in the translated version's resonance. Conversely, critics maintain that irrespective of a translator's proficiency, subtle linguistic nuances and cultural references are invariably lost, thus diluting the talk's authenticity. They argue that these variances in tone and meaning may result in misinterpretations, a risk that bears particular weight in sensitive contexts such as diplomatic exchanges or legal matters.

This study, leveraging datasets from the official IWSLT 2018 website[2] under the Creative Commons BY-NC-ND license[1], investigates whether translated TED Talks retain their original resonance in English when translated to Arabic. With a carefully curated selection of 300 talks per language, our analysis utilizes 240 talks for an exploratory examination, focusing on potential disparities in linguistic features between the two languages. Each talk is associated with comprehensive metadata to ground the analysis in a comparative context. The outcome aims to shed light on the empirical validity of the claim that "translated talks sound different" and to explore the linguistic dimensions in which these differences may be observed.

General Descriptive Statistics

This section presents the general descriptive statistics for the English and Arabic datasets used in our analysis. The datasets consist of talks, each uniquely identified by a TalkId, along with associated speakers, keywords, content entries, and URLs. The descriptive statistics are summarized in Table 1.

Both datasets exhibit a high degree of consistency in terms of the number of records, keywords, content entries, and URLs, all totaling 240 for each language. This indicates a comprehensive collection of talks, providing a substantial basis for comparative linguistic analysis. The TalkId range for the English dataset spans from 2591 to 12908, while the Arabic dataset spans a slightly narrower range from 2491 to 12908. This difference in TalkId range may reflect variations in the collection period or the volume of talks in each language.

One notable difference between the datasets is in the number of unique speakers. The English dataset comprises talks from 240 unique speakers, whereas the Arabic dataset includes talks from 239 speakers. This slight variation suggests a diverse representation of speakers in both languages, albeit with a marginally higher diversity in the English dataset.

The consistency in the number of keywords, content entries, and URLs across both datasets indicates a uniform structure in data collection and categorization. Such uniformity is crucial for ensuring the comparability of the datasets in subsequent analyses.

Statistic	**English Dataset**	**Arabic Dataset**
Number of Records	240	240
TalkId Range	2591 to 12908	2491 to 12908
Speakers	240	239
Keywords	240	240
Content Entries	240	240
URLs	240	240

Table 1: Descriptive Statistics of English and Arabic Datasets

Preprocessing

Upon further examination of the data, it has come to our attention that the Arabic dataset contains outliers, as indicated in Table 1a. The abnormal values in both standard deviation and the maximum average words per sentence are evident.

After the removal of outliers, approximately 13 instances were identified and addressed, resulting in an improved dataset,as depicted in Table 1b. The impact of outliers on the mean, standard deviation, and maximum values is observable. Additionally, to maintain a fair comparison, 13 instances were also removed from the English dataset.

Figure 1: Comparison of Statistics for Arabic Dataset with and without Outliers.

(a) Statistics before removing outliers.

	Word Count	Sentence Count	Avg Word Per Sentence
Number of talks	240	240	240
Mean	1622.850	93.508	38.266
Standard Deviation	850.487	53.038	121.586
Minimum	16.000	1.000	10.231
25th Percentile	1094.751	61.750	14.930
50th Percentile	1596.000	92.000	16.780
75th Percentile	1971.250	119.500	19.239
Maximum	7858.000	390.000	1216.000

(b) Statistics after removing outliers.

Word Count	Sentence Count	Avg Word Per Sentence
227	227	227
1634.348	98.445	17.074
850.959	50.192	3.610
16.000	1.000	10.231
1089.500	68.000	14.835
1617.000	93.000	16.483
1973.000	121.000	18.643
7858.000	390.000	36.583

Exploratory Analysis for Both Languages

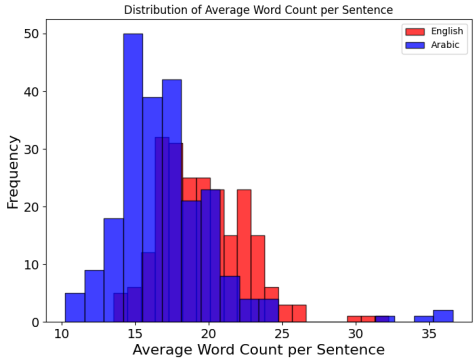
In our exploratory analysis, we investigated various linguistic features within the English and Arabic datasets. The objective was to discern patterns and characteristics unique to each language and to understand how they compare when analyzed through computational methods.

Average Word Count per Sentence

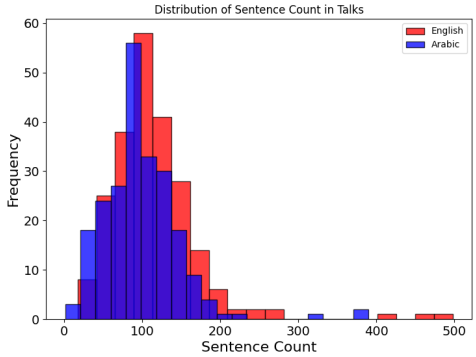
The first aspect we explored was the average word count per sentence for each talk. Figure 2a illustrates the distribution of average word count per sentence for both datasets. It is evident that there are notable differences in sentence length between the two languages, which could be attributed to linguistic structure and usage in different contexts.

Sentence Count in Talks

Next, we assessed the number of sentences in each talk, as shown in Figure 2b. The frequency distribution reveals that English talks tend to have a higher sentence count, which may suggest a difference in content delivery or structuring between the two languages.



(a) Distribution of Average Word per Sentence



(b) Distribution of Sentence Count in Talks

Figure 2: Distribution Plots

Word Count in Talks

The total word count in talks was also compared, depicted in Figure 3. The graph shows the distribution of word count in talks by language. The most common word count for talks in English is 2,000 to 3,000 words, while the most common word count for talks in Arabic is 1,000 to 2,000 words. The graph also shows that there is a wider range of word counts for talks in English than for talks in Arabic.

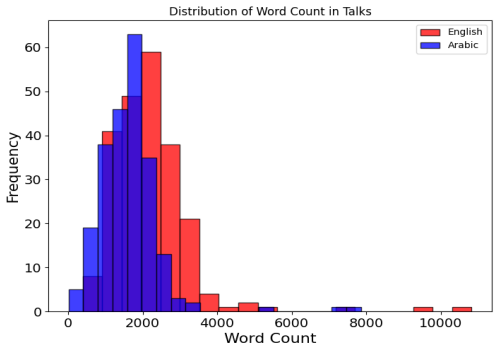


Figure 3: Distribution of Word Count in Talks

Description of the Initial Findings

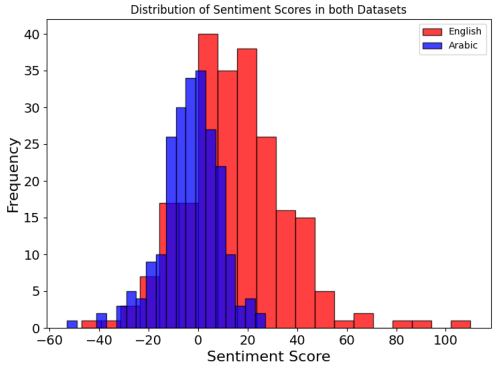
In this section, we provide an overview of the initial findings from the analysis of TED Talks datasets, focusing on the sentiment of both Arabic and English TED Talks, as well as the distribution of sentiment scores in each dataset and a comparison between the two. In conducting sentiment analysis for the English language, we used the opinion lexicon provided by the NLTK library. Conversely, for Arabic sentiment analysis, we used the Googletrans library to translate the English opinion lexicon from the NLTK library to Arabic.

Sentiment Analysis: Original English vs Translated Arabic Texts

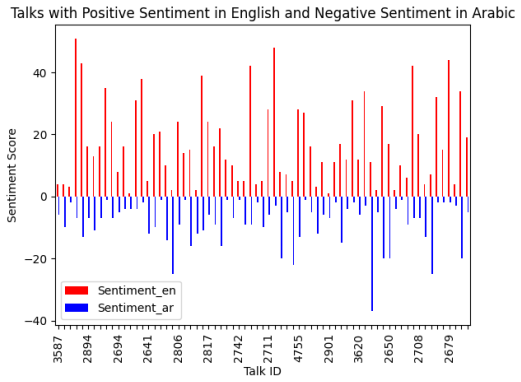
The histogram depicted in Figure 4a illustrates the distribution of sentiment scores across two datasets, one in English and the other in Arabic. The x-axis represents the sentiment score, which quantifies the emotional content ranging from highly negative to highly positive sentiments. The y-axis denotes the frequency of each sentiment score within the datasets.

Observations from the histogram indicate a notable contrast between the two languages. The English dataset exhibits a bimodal distribution, with peaks in both the positive and neutral sentiment regions. This suggests a diverse range of expressed sentiments, with a substantial presence of both positive and neutral emotions. Conversely, the Arabic dataset displays a unimodal distribution with a peak centered around neutral sentiment scores. This pattern implies that the Arabic texts tend to convey a more neutral emotional tone, with less frequent occurrences of extreme positive or negative sentiments. The overlay of the two distributions also reveals that the range of sentiment scores in the English dataset is wider than that in the Arabic dataset, indicating a broader spectrum of emotional expression in English-language texts.

The graph 4b display the sentiment scores for a series of talks, comparing English and Arabic. The sentiment scores for English talks are shown in red and are all above the zero line, indicating a positive sentiment. In contrast, the sentiment scores for the Arabic translations are shown in blue and are all below the zero line, indicating a negative sentiment. The graph highlight the challenges in maintaining consistent sentiment across translations in different languages.



(a) Histogram showcasing the frequency of sentiment scores in English and Arabic datasets



(b) Talks with Positive Sentiment in English and Negative Sentiment in Arabic

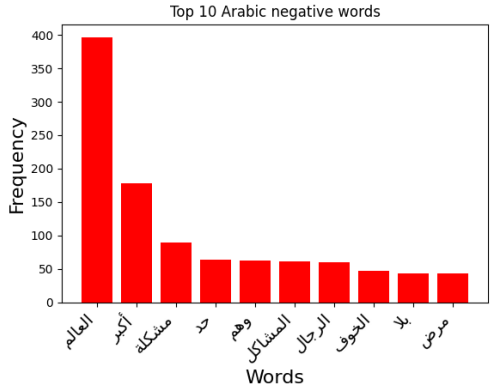
Figure 4: Comparison of sentiment-related figures.

Exploring Sentiment in Arabic Translations

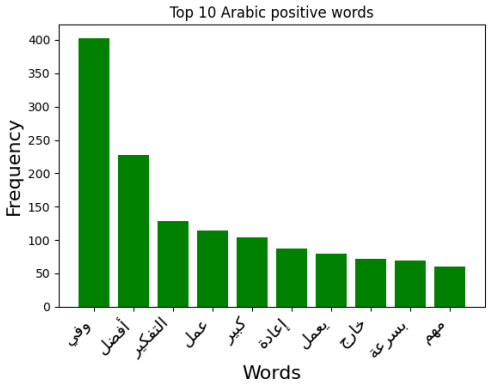
The first plot (see Figure 5a) illustrates the frequency of the top 10 negative words found in the Arabic translation of the dataset. The x-axis lists the words, while the y-axis indicates the frequency of each word's occurrence. The highest frequency word dominates the chart, suggesting a prevalent negative sentiment or topic within the Arabic texts. In contrast, the second plot (see Figure 5b) displays the frequency of the top 10 positive words in the Arabic dataset. The distribution is more uniform, yet one word still significantly leads in frequency, indicating particular positive themes or sentiments that resonate within the Arabic version.

Exploring Sentiment in English Texts

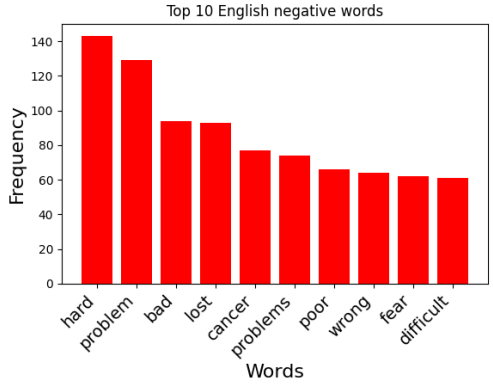
The third plot (see Figure 5c) shows the top 10 negative words in the original English texts. The words are more evenly distributed across the frequency spectrum, suggesting a variety of negative concepts are present in the English dataset. Lastly, the fourth plot (see Figure 5d) presents the top 10 positive words from the original English dataset. The word "like" significantly outnumbers the others, which might indicate a strong preference for this term in expressing positive sentiments or could be a linguistic filler commonly found in spoken language.



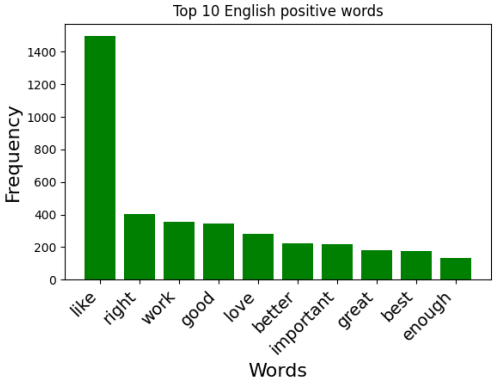
(a) Negative words in the Arabic dataset. The translation of these words, in the same order, is *[World, Greater, A problem, Limit, Delusion, The problems, Man, Afraidness, without, sickness]*



(b) Positive words in the Arabic dataset. The translation of these words, in the same order, is *[Loyal, Better, Thinking, Work, Bigger, Again, Works, Outside, Fast, Important]*



(c) Top 10 negative words in the English dataset, by frequency.



(d) Top 10 positive words in the English dataset, by frequency.

Figure 5: Comparison of top 10 positive and negative words in the Arabic and English datasets.

Conclusion

The exploratory analysis focused on sentiment scores derived from TED Talks in both English and Arabic, aiming to identify if the sentiment conveyed in translated TED Talks retains fidelity to the original English versions. The initial findings suggest that while there is an overlap in the range of sentiments between the two languages, the translated Arabic talks tend to lean towards a more neutral sentiment, potentially implying a loss of emotional intensity in the translation process. This comparative sentiment analysis serves as a stepping stone for further investigation into the nuances of emotional expression across languages and the impact of translation on sentiment portrayal.

References

[1] *IWSLT 2018 Commons BY-NC-ND license.* https://www.ted.com/about/our-organization/our-policies-terms/ted-talks-usage-policy. Accessed: 2023-04-05.

[2] *IWSLT 2018 Datasets.* https://wit3.fbk.eu/2018-01-b.