Did subtitles ruin your movie? Correlating Movie Ratings with Cross-Lingual Subtitle Translation Quality

Vishnu Sashank Dorbala
vdorbala@umd.edu
University of Maryland, College Park

Shreelekha Revankar
revankar@umd.edu
University of Maryland, College Park

Abstract

We look at how the quality of subtitles influences audience perception of foreign language movies, and ask if current state-of-the-art Neural Machine Translation (NMT) models can serve as viable aide or alternative to professional movie translation. Our results show that NMT is indeed a viable option as an aide or alternative to human translation. Our code can be found here - https://github.com/vdorbala/Movie-Subtitle-Quality

1 Introduction

Cinema is a powerful medium for visual-language communication. Movies have the potential to deeply influence our lives in conveying ideas and abstract concepts in a grand manner (Madsen, 1973). With the advent of globalization, the cross-cultural exchange of movies across various regions has become commonplace. As such, there is a necessity for accurate screen translation for foreign language films so that they may convey the right contextual information to audiences (Dwyer, 2017).

As the cost of remaking an entire film in multiple languages is high, movie producers often resort to either *dubbing* (*dubs*) or *professional translation* (*subtitling* or *subs*). For dubbing, the script is translated and voice actors are hired to re-record the audio for the entire film. Subtitling is the more efficient of the two, requiring only the translation of the script and can even be extended to reach audiences with auditory disabilities with the inclusion of a few descriptive phrases.

It is important to note that poorly translated subtitles can result in a *defamiliarizing* effect on audiences for whom film was not originally written for (Kapsaskis, 2008). To better understand this, one can look at the ratings of a film in its origin country compared to its reception in a country using a different language.

For this reason, in this work, we focus on estimating the quality of "*subs*" rather than "*dubs*". In particular, we look at how the quality of *subtitles* influences audience perception of movies, and ask if current state-of-the-art Neural Machine Translation (NMT) models can serve as viable aide or alternative to professional movie translation.

A lot of previous work in this area deals with Automated Subtitle Generation (Bywood et al., 2014; Gupta et al., 2019). In (Gupta and Sharma, 2020; Gupta and Nelakanti, 2020), the authors present approaches to perform multilingual translation quality estimation. While these are certainly useful to independently analyze translation quality, they fail to consider its real world impact. Other independent research studies cross-lingual sentiment preservation (Kajava et al., 2018, 2020) on a large-scale movie subtitles dataset (Creutz, 2018). Their results suggest that sentiment information is sufficiently preserved; however they also observe that translated sentiment data is likely to contain samples which are not representative of their assigned sentiment class in the translated language. We take inspiration from the outcome of these studies to assess the real world impact of cross-lingual subtitle quality.

We hypothesize that films which perform similarly or better in their translated scripts have similar qualities to that of the original script. Whereas, those films that perform worse, are missing qualities that the original script contained. In this direction, we first look at identifying and evaluating various script qualities including sentiment analysis, type-to-token ratio, lexical density,and unique words. We then utilize these qualities to retrieve *Key Scenes of Disagreement (KSD)* between original, machine-translated, and human-translated scripts. These retrieved scenes are representative of which parts of the movie failed or achieved to get accurately translated. Finally, we look at correlating *KSD* with audience ratings of movies in different countries, and comment on when NMT outperforms human-translated scripts. In summary, our work has the following contributions:

- We tackle the novel problem of correlating multilingual movie ratings with script translation quality. For simplicity and proof of concept, we limit ourselves to *English-French* and *French-English* scripts.
- Our approach seeks to identify *Key Scenes of Disagreement (KSD)* across the bilingual movie scripts. These scenes give us information about instances where subtitles failed to convey proper meaning.
- We provide an analysis of the correlation between features of movie scripts before and after translation with audience scores in both languages.
- Finally, we perform experiments with using SOTA *multilingual Neural Machine Translation (NMT)* models on the bilingual script data, and use these results to infer if these models could be used as a substitute of human-translation.

2 Related Works

Script writing and their approval process has been an integral part of the film making industry since its inception (Ivarsson, 2009). The quality of the subtitles severely impacts the viewers comprehension of the content(Kapsaskis, 2008), with viewers heavily relying on subtitles when the language being spoken is unknown (bis).

The quality of a translation is characterized by many features. Sixel explored the features that created a good translation, describing how it is important for a translator to know which audience they are translating for; for the translation to be a correct translation, it must be true. Sixel highlights that for a true translation, the emotional contexts as well as formality and word choices are maintained (Sixel, 1994). Our goal is to utilize language models and language processing metrics to characterize the features of our translated film scripts.

To analyze how well emotional context is maintained, we utilize sentiment analysis. Sentiment analysis is a natural language process by which we can quantify which sentiment is being expressed in a set of words (Feldman, 2013).

To analyze how well linguistic features were maintained when scripts are translated we look to Translationese. The objective of Translationese is to recognize the features of a translation which allow us to differentiate between translations and their reference texts (Volansky et al., 2015). Volansky et al. utilized various linguistic features to classify whether a text was translated or not. The findings of their studies showed several linguistic features as fairly accurate indicators of translated text. From this research, we chose several of the effective linguistic features from to analyze our scripts, to quantify the similarity of the lexicons across scripts.

Much previous work has been done Machine Generation of subtitles. Gupta et. al present issues within machine translation for subtitle generation (Gupta et al., 2019). However, they failed to capture human sentiments towards the subtitles before and after translations, as well as the real world impacts of the poor translations. Etchegoyhen et. al present approaches to perform multilingual translation quality estimation (Bywood et al., 2014). However they do not take into account an important feature, sentiment analysis. Other have looked specifically into preservation of sentiment; Kajava has performed an analysis on the preservation of sentiments after annotation project(Kajava et al., 2018, 2020) using a large-scale movie subtitles dataset (Creutz, 2018). They found that sentiments were generally preserved well. However, this sentiment may be the result of an inaccurate translation.

From these studies we seek to analyze both the sentiments as well as statistical linguistic features in our source and translated scripts to see how well these features are preserved when translated for foreign audiences. We then hope to utilize real world human ratings for the films in both their source and translated state, to examine our findings. Ultimately, we hope to provide a holistic review of where we may proceed with the machine translations of film scripts.

3 Method

3.1 Data Curation

To analyze the cross-lingual subtitles we chose French and English films. This is largely due to the large availability of human audience ratings in both languages. These real-world audience ratings were retrieved from Internet Movie Database(imdb) AlloCiné (allo) for Ratings. They are both online film databases, which collate user ratings for films (all; imd). For our film subtitles we used Opensubtitles, a large free searchable subtitle database (Creutz, 2018) To find our experiment dataset we began by selecting films for which ratings were available in both allo and imdb. Following that we filtered down to films by their country of origin, selecting films made in France, and The United States. Although there are many other French and English speaking countries, and both France and the USA have multilingual films, the demographics of film database users were mainly French and American for Imdb and AlloCiné respectively. We then normalized all of the film ratings on a 10 point scale and removed all of the films which had equivalent ratings as our goal was to see what may have caused the disparities in the reception of the films in foreign audiences. We then calculated the absolute difference in ratings between imdb and allo for each film, using this as metric for selecting our films later on.

Once we had narrowed down this subset of films, we had many duplicate titles for different films created sometimes decades apart, to reduce confusion we removed duplicate titles. For example ”Gabrielle, 2003” and ”Gabrielle, 2013”.

We used the OpenSubtitle API to retrieve matched subtitles, and ended up with 666 films. We cut these down to 58 films using genre, and rating disparity, to alleviate computational issues (takes a month to train) and to have an even distribution of films across these factors. Figure 1 showcases the distribution of rating differences across all films.

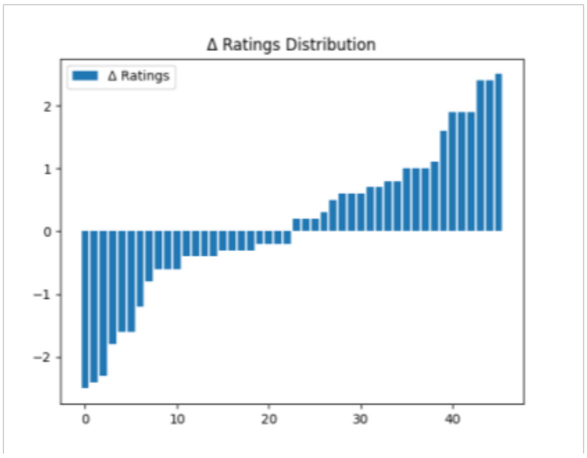


Figure 1: The distribution of ratings across all our movies.

Finally, we are left with our groundtruth subtitles in English and French **en(GT)** and **fr(GT)**.

3.2 Scene Segmentation

Movies comprise of a set of scenes having unique qualities. The nature of storytelling is such that each scene flows from one to the next. We look at computing script qualities across these scenes, rather than a line by line analysis. This gives us a more logical and interpretable understanding of the scene flow properties, which we hypothesize as one of the key factors affecting moviegoer experiences.

Since obtaining the actual lines of the script that correspond to scenes is cumbersome, and sometimes impractical we look at averaging out script qualities on a per-scene basis. This does not affect our research question, as we look at comparing different corresponding sets of lines across original, machine translated and human-translated scripts.

As a narrative planning principle, movies on average must contain an adequate number of scenes (Riedl and Young, 2010). We approximate this number as a hyperparameter that we set to 50. We then divide the number of lines in each movie by 50, to give us the average number of lines per scene. Table 1 showcases these stats.

Once we have these scenes extracted for *fr* and *en*, we perform NMT on them, to get *en-fr(MT)* and *fr-en(MT)* scripts. We use the OPUS MT (Tiedemann et al., 2020) for this purpose, which provides an interface for utilizing the latest SOTA transformer-based translation models.

Finally, we have a curated list of 58 movies for our analysis, each having - **en(GT)**, **en-fr(MT)**, **fr-en(MT)**, **fr(GT)** subtitles.

Movies	Avg. Total Lines	Avg. Lines Per Scene
58	1567	32.64

Table 1: Stats about movies and scene selection. We heuristically select 50 as the average number of scenes in a movie, and use this to compute the average number of lines per scene.

3.3 Cross-Script Sentiment Analysis

We use the multilingual BERT NMT model (Zhu et al., 2020) to perform the downstream sentiment analysis task to find English and French movie sentiments. Sentiments are computed line by line, and averaged out per scene. We observe that the pretrained ML BERT model gives higher sentiment scores in general for French scripts, and this pattern is similar across all movies. To further verify this, we compared sentiment performances of (**en**,**fr-en**) and (**fr**, **en-fr**) with (**en**, **fr**). This result can be seen for Paranormal Activity in figure 3, and holds true across all other films as well.

3.4 Statistical Linguistic Features

To quantify the lexical features of our subtitles, we decided to utilize two measures of Lexical Density to produce a Statistical Linguistic Features Score (Stat.Feat). These are type-to-token ratio (Halliday, 1989) and Herdan's or the Logarithmic type-to-token ratio (TTR) (Popescu, 2009)

Type-to-token ratio ($\mathcal{TTR}$) is calculated as follows: Where V is the number of types and N is the number of tokens in the given chunk of text.

$$\mathcal{TTR} = V/N \tag{1}$$

And, Herdan's TTR ($\mathcal{HTR}$) is calculated as:

$$\mathcal{HTR} = V/N \tag{2}$$

3.5 Key Scenes of Disagreement (KSD) Score

Once we calculate Statistical Linguistic features (LS) and Sentiment Transfer scores (ST) we utilize these scores to obtain Key Scenes of Disagreement scores for each movie.

For Sentiment Transfer, we follow the steps described in section 3.3. We an additional thresholding on the scene-level disparity between two scripts (say en, and fr-en), to extract sentiment differences for top 'm' worst scenes, and top 'n' best scenes. Rather than just computing the scene average which tends to normalize the sentiments across the whole movie, this would give us a better idea of the scene flow.

We take the maximum value of this and the average sentiment '$\mathcal{SSA}$', to give us our sentiment transfer score. Lower the sentiment score, the better the scenes correspond to teach other.

Sentiment-transfer score can be defined as,

$$\mathcal{ST} = \max(\mathcal{SSA}, \frac{\sum_{i}^{n} \mathcal{SSD}[i] + \sum_{i}^{m} \mathcal{SSD}[j]}{m+n}) \tag{3}$$

Here, $\mathcal{SSD}$ is scene sentiment difference.

We get statistical linguistic data on both scripts, using the process outlined in section 3.4. Using a similar process used for obtaining the *ST* score, we get lexical feature quality using the following equations.

$$\mathcal{LS} = \frac{\sum^{t} \mathcal{L}_t}{t} \tag{4}$$

where,

$$\mathcal{L}_t = \max(\mathcal{LA}_t, \frac{\sum_{i}^{m} \mathcal{LD}_t[i] + \sum_{j}^{n} \mathcal{LD}_t[j]}{m+n})$$

$\mathcal{TA}_t$ is the average statistical difference across all scenes and all lexical qualities, while $\mathcal{TD}_t[i]$ is the statistical difference for the t^{th} lexical quality and the i^{th} scene.

Finally, our *KSD* score is a weighted sum of the sentiment transfer and the statistical linguistic score. It can be written as,

$$\mathcal{KSD} = \lambda_1 \mathcal{ST} + \lambda_2 \mathcal{LS} \tag{5}$$

Hyperparameters are m, n, λ_1 and λ_2.

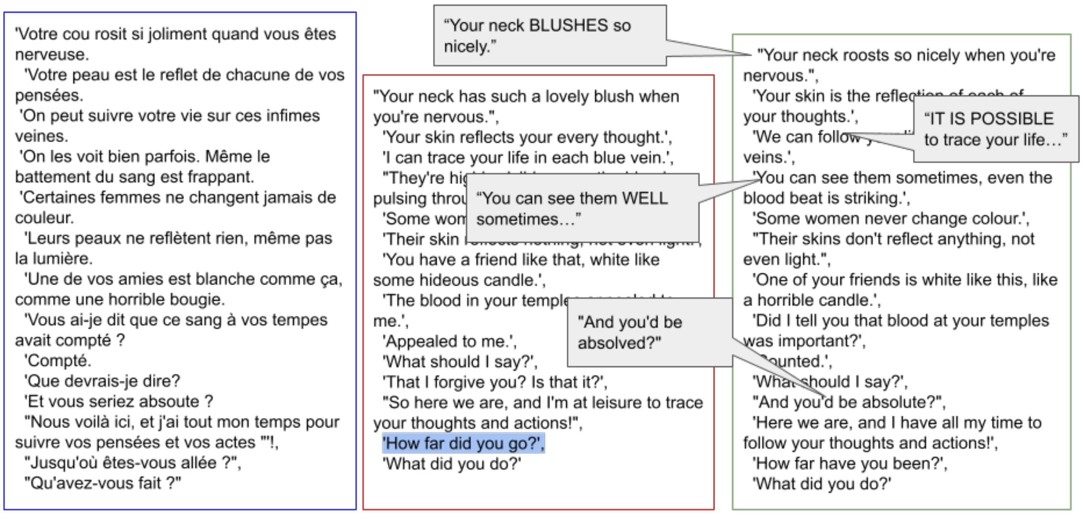
4 Experimentation and Results

We trained all our models on an NVIDIA TITAN X GPU. For evaluating the KSD score, we set our hyperparameters as $m, n = 5$, and λ_1 and λ_2 as 0.5 each. Our dataset of 58 movies contains 34 original English films, and 24 original French ones. We chose these by filtering out the dataset using *genre*, and *rating difference*. Tables 3 and 2 showcase our results.

4.1 Sentiment Transfer

ML BERT gives us a star rating between 1-5 for sentiment labels, and a corresponding *score* value for magnitude. To get a net mean sentiment around 0, we take the negative of the *score* value of any label with $\leq$ 2 stars. This is slightly biased towards a positive score, as there is not 2.5 star rating available.

As mentioned in section 3.3, the ML BERT model shows a higher overall sentiment across French movies in comparison with English ones indicating a bias on the downstream sentiment analysis task. This is clearly reflected on our quantitative results. in table 3. Notice the negative trend in sentiment score values, when translated English, which shows higher sentiment scores in French.



'Votre cou rosit si joliment quand vous êtes nerveuse.
'Votre peau est le reflet de chacune de vos pensées.
'On peut suivre votre vie sur ces infimes veines.
'On les voit bien parfois. Même le battement du sang est frappant.
'Certaines femmes ne changent jamais de couleur.
'Leurs peaux ne reflètent rien, même pas la lumière.
'Une de vos amies est blanche comme ça, comme une horrible bougie.
'Vous ai-je dit que ce sang à vos tempes avait compté ?
'Compté.
'Que devrais-je dire?
'Et vous seriez absoute ?
"Nous voilà ici, et j'ai tout mon temps pour suivre vos pensées et vos actes '"!,
"Jusqu'où êtes-vous allée ?",
"Qu'avez-vous fait ?"

"Your neck has such a lovely blush when you're nervous.",
'Your skin reflects your every thought.',
'I can trace your life in each blue vein.',
"They're hi
pulsing throu
'Some wom
'Their skin
'You have a friend like that, white like some hideous candle.',
'The blood in your temple
me.',
'Appealed to me.',
'What should I say?',
'That I forgive you? Is that it?',
"So here we are, and I'm at leisure to trace your thoughts and actions!",
'How far did you go?',
'What did you do?'

"Your neck roosts so nicely when you're nervous.",
'Your skin is the reflec
your thoughts.',
'We can follow
veins.',
'You can see them sometimes, even the blood beat is striking.',
'Some women never change colour.',
"Their skins don't reflect anything, not even light.",
'One of your friends is white like this, like a horrible candle.',
'Did I tell you that blood at your temples was important?',
ounted.',
'What should I say?',
"And you'd be absolute?",
'Here we are, and I have all my time to follow your thoughts and actions!',
'How far have you been?',
'What did you do?'

"Your neck BLUSHES so nicely."

"IT IS POSSIBLE to trace your life…"

"You can see them WELL sometimes…"

"And you'd be absolved?"

Figure 2: Human Evaluation of Human vs. Machine Translation: We perform a single user study to evaluate our approach. Our results here show that the machine translated version on the right gave better results in some cases as opposed to human-translation. This confirms our hypothesis and gives credibility to our research question of using NMT for generating better subtitles.

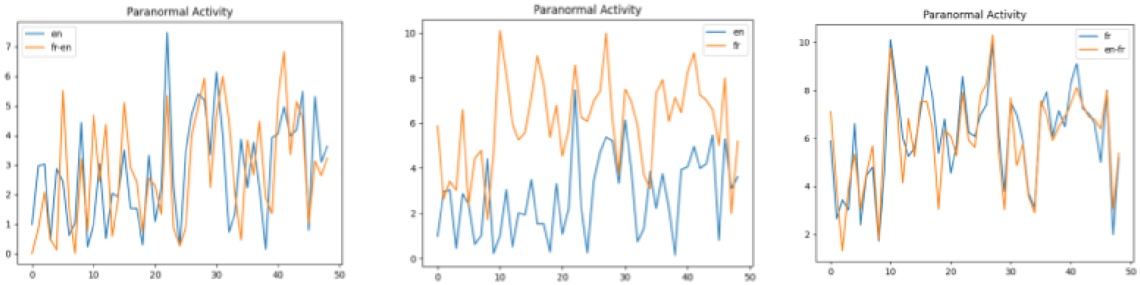


Figure 3: We compare the sentiments of (**en**,**fr-en**) and (**fr**, **en-fr**) with (**en**, **fr**). Observe the more optimistic French sentiments on all the plots. A plausible explanation for this is a bias in the multilingual BERT model.

4.2 Statistical Linguistic Features

Post-translation we found that for some film the human translation had more similar features to the ground-truth script while for others the machine translation achieved that. To compare how these scores characterize the different scripts fig.4 displays the scene-wise $\mathcal{HTR}$ scores for two films, showing disparities in the statistical linguistic features across the different versions of the scripts. We see that for Paranormal Activity, the human French translation is more akin to the ground-truth English subtitles, while for Gabrielle the statistical linguistic features of all of the scripts are more similar to one another, with the machine translated English script being the closest across most scenes.

4.3 Human Evaluation

To get a qualitative analysis of our experiments as well as a way to validate our findings we decided to conduct a human evaluation. A student at a large public university who is fluent in both French and English was the participant of our study. Using our KSD scores we computed which scene from Gabrielle performed the worst in the original ground-truth translation from French to English. This is the scene which we presented to the participant. They were given 3 sets of parallel corpora, named *French 1*, *English 1* and *English 2*, and were given the prompt "After reading *French 1* we would like you to pick which English translation between *English 1* and *English 2* you feel is most accurate, please include any comments or notes you had while making this decision" *French 1* was the French ground-truth, *English 1* was the English human translation while *English 2* was our NMT translation. The response, seen in fig. 2 we received was that *English 2* offered the most accurate translation of *French 1* with some comments of words

Δ Audience Reception	Avg. Δ KSD		Avg. Δ Sent. Transfer		Avg. Δ Stat.Feat.	
	GT_{EN}	NMT_{EN}	GT_{EN}	NMT_{EN}	GT_{EN}	NMT_{EN}
Positive ($\Delta > 0$)	4.321	-3.617	2.770	-1.732	0.083	0.155
Neutral ($\Delta \quad = 0$)	-6.812	-7.135	-3.368	-3.483	0.075	0.168
Negative ($\Delta < 0$)	-6.960	-7.314	-3.447	-3.588	0.0685	0.138

Table 2: Average Δ in Sentiment Transfer for French Films, Statistical Linguistic Features, and KSD compared with groundtruth French subtitles

Δ Audience Reception	Avg. Δ KSD		Avg. Δ Sent. Transfer		Avg. Δ Stat.Feat.	
	GT_{FR}	NMT_{FR}	GT_{FR}	NMT_{FR}	GT_{FR}	NMT_{FR}
Positive ($\Delta > 0$)	7.511	7.154	3.789	3.603	0.067	0.050
Neutral ($\Delta \quad = 0$)	5.08	5.022	2.581	2.543	0.081	0.063
Negative ($\Delta < 0$)	3.015	2.649	5.935	5.237	0.095	0.061

Table 3: Average Δ in Sentiment Transfer for English Films, Statistical Linguistic Features, and KSD compared with groundtruth English subtitles

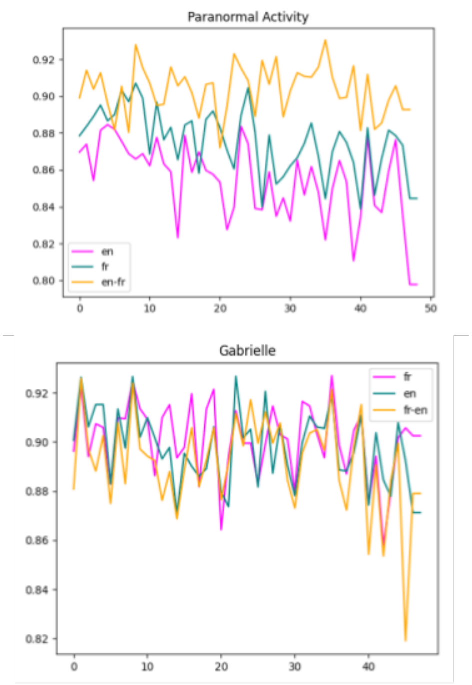


Figure 4: We compare the performance of NMT and GT translations for both originally English (Paranormal Activity, bottom), and originally French (Gabrielle, top) films. These are quanitfied using $\mathcal{HTR}$ scores. In the case of Gabrielle, observe that the NMT performance is on par with the GT, but with Paranormal Activity, the NMT translation to french seems to perform worse than the ground truth French

they wished were different and a single sentence for which they felt *English 1* was preferred.

5 Discussion

5.1 Multilingual BERT Sentiments

While raw sentiment values are useful, we observe that it might not give a good estimate of the emotional flow of events in the movie, following the principle of the Freytag Pyramid (Jago, 2004). Events in a movie build up from previous scenes, and as such, the sentiments are cumulative. Defining a metric to cumulatively evaluate sentiments across scenes might help capture the flow of the movie better, and provde a better Freytag Pyramid - leading to better understanding of audience responses.

Also, we have seen that the Multilingual BERT model does not normalize sentiments across languages. As such, we require systems that can provide such normalized sentiments for easier comparison, and less bias across languages.

5.2 Social Impact

Cinema is a powerful medium, with subtitles being the most efficient way of reaching foreign audiences as well as those with auditory disabilities this is why making the process of getting subtitles more efficient is extremely important. We hope with the shown improvement in subtitle generation allows for ease of access to films and other visual media.

6 Conclusion

Our results, and the subsequent human evaluation seem to suggest that NMT can sometimes give better quality translations, which show NMT may be a viable alternative or aide for generating subtitles.

References

Allociné. Accessed: 2022-05-7.

Internet movie database. Accessed: 2022-05-7.

L Bywood, T Etchegoyhen, Panayota Georgakopoulou, M Fishel, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, A Turner, M Volk, and M Maucec. 2014. Machine translation for subtitling: A large-scale evaluation. In *LREC 2014, Ninth International Conference on Language Resources and Evaluation*, pages 46–53.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. *arXiv preprint arXiv:1809.06142*.

Tessa Dwyer. 2017. *Speaking in subtitles: revaluing screen translation*. Edinburgh University Press.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Prabhakar Gupta and Anil Nelakanti. 2020. Deepsubqe: Quality estimation for subtitle translations. *arXiv preprint arXiv:2004.13828*.

Prabhakar Gupta and Mayank Sharma. 2020. Unsupervised translation quality estimation for digital entertainment content subtitles. *International Journal of Semantic Computing*, 14(01):137–151.

Prabhakar Gupta, Mayank Sharma, Kartik Pitale, and Keshav Kumar. 2019. Problems with automating translation of movie/tv show subtitles. *arXiv preprint arXiv:1909.05362*.

Michael Alexander Kirkwood Halliday. 1989. *Spoken and written language*. Oxford University Press, USA.

Jan Ivarsson. 2009. The history of subtitles in europe. *Dubbing and subtitling in a world context*, pages 3–12.

Carol Jago. 2004. Stop pretending and think about plot. *Voices from the Middle*, 11(4):50.

Kaisla Kajava, Emily Öhman, Piao Hui, Jörg Tiedemann, et al. 2020. Emotion preservation in translation: Evaluating datasets for annotation projection. *Proceedings of Digital Humanities in Nordic Countries (DHN 2020)*.

Kaisla Kajava et al. 2018. Cross-lingual sentiment preservation and transfer learning in binary and multi-class classification.

Dionysis Kapsaskis. 2008. Translation and film: On the defamiliarizing effect of subtitles. *New voices in translation studies*, 4:42–52.

Roy Paul Madsen. 1973. The impact of film. how ideas are communicated through cinema and television.

Ioan-Iovitz Popescu. 2009. Word frequency studies. In *Word Frequency Studies*. De Gruyter Mouton.

Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.

Friedrich Sixel. 1994. What is a good translation? some theoretical considerations plus a few examples. *Meta: journal des traducteurs/Meta: Translators' Journal*, 39(2):342–361.

Jörg Tiedemann, Santhosh Thottingal, et al. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.