Similarity Score Analysis of Language Translations Dataset using Meta's "Seamless_M4t_v2_large" Model

Rahul Kumar
Department of Electronics and Communication
Delhi Technological University
Delhi, India
rahulkumar_ec20b13_38@dtu.ac.in

Naveen Bemad
Department of Electronics and Communication
Delhi Technological University
Delhi, India
naveenbemad_ec20b13_14@dtu.ac.in

Sachin Taran
Department of Electronics and Communication
Delhi Technological University
Delhi, India
sachintaran@dtu.ac.in

***Abstract*— The advanced approach to multilingual translation by the M4T communication model is based on modern techniques that make possible smooth and effective communication across languages in today's world. It is a good system for doing real-time translations (interpretations) at international conferences, personal travels, or business meetings as it reduces the amount of time taken during translation which means there will be no interruptions while interpreting. This work presents the similarity score analysis of language translations using Meta's "Seamless M4t v2 large" Model. The " Seamless_M4t_v2_large" model is applied to determine translations between different languages and assess similarity scores between the output of the model and reference language translations from the Opus books dataset. It is composed of text-to-text translation data that provides a basis for categorizing, based on areas that can be improved through additional training using this method. Next, similarity scores were calculated between original and translated strings using spacy similarity scoring functionality. The similarity scores of different language translations provide an idea of which areas the model needs to improve.**

Keywords— SeamlessM4T, Similarity score, EMMA, Opus books, Text to text translation, Spacy.

I. INTRODUCTION

Meta introduced a new model of communication model where users can simultaneously translate from speech to speech and text, text to text, and speech with a very low latency score but what is latency score and what is its significance? Have a look at it here. The model's need for partial input information to generate a portion of the translation is what alignment latency refers to. regularization term derived from the estimated alignment leads to reduced latency. It helps to determine how fast data can be translated so that every person can communicate in an instant. There are main three models associated with this new technology that is expressive for speech-to-speech, streaming for speech-to-speech(S2S), speech-to-text(S2T), and automatic speech recognition (ASR) [1].

Monotonic attention models have a learnable policy that is based on aligning estimates during training [2]. This means that although translations can be done in such a way that certain words will retain their meaning, some intrinsic aspects of them cannot be translated [3]. Even before source sequences have been read or encoded, simultaneous machine translation models begin generating target sequences. The recent approaches to this task either involve a fixed policy over a state-of-the-art Transformer model or a weak recurrent neural network-based structure with a learnable monotonic attention. This work extends the monotonic attention mechanism to multihead attention (MMA) [3]. Additionally, two new novels have introduced interpretable and dedicated methods for multiple attention heads latency control. For the simultaneous machine translation system. This analyses how the latency controls influence the length of an attention span to motivate our model by investigating how quality and latency are affected by decoder layers and the number of decoders per head [4].

EMMA: state-of-the-art concurrent translator based on monotonic multi-head attention Idea. Moreover, we propose more effective methods for training and inference in these models by performing joint fine-tuning from an offline translation model while minimizing the alignment shifts that are monotonic. This approach is demonstrated by experimental results to outperform the existing models on Spanish-English simultaneous speech-to-text translation systems [5]. Another thing to notice is that human speech and translation take into consideration a range of pragmatic nuances such as turn-taking and timing controls [6]. Besides learning languages through teaching and instruction, many learners seek alternative approaches to bridge communication gaps [7].

By analyzing the similarity scores of different language translations this paper identified in which areas where model needed to improve. When the similarity score between translations in different languages is significantly low, it indicates discrepancies or inconsistencies in the model's performance across languages. Therefore, by observing these instances of low similarity scores, we can improve training processes in that significant area of the model. "Seamless_M4t_v2_large" model for testing text-to-text translation between different languages from the "opus_books" dataset which we got from hugging face community given by Abhishek Kumar thakur. This dataset consists of translations of 16 languages. This dataset includes Catalan, Greek, English, Esperanto, Spanish, French, Hungarian, Russian, Dutch, Italian, German, Finnish, Norwegian, Polish, Portuguese, and Swedish translations which help to determine model output and compare with output to analyze similarity scores [8].

979-8-3503-8719-3/24/$31.00 ©2024 IEEE

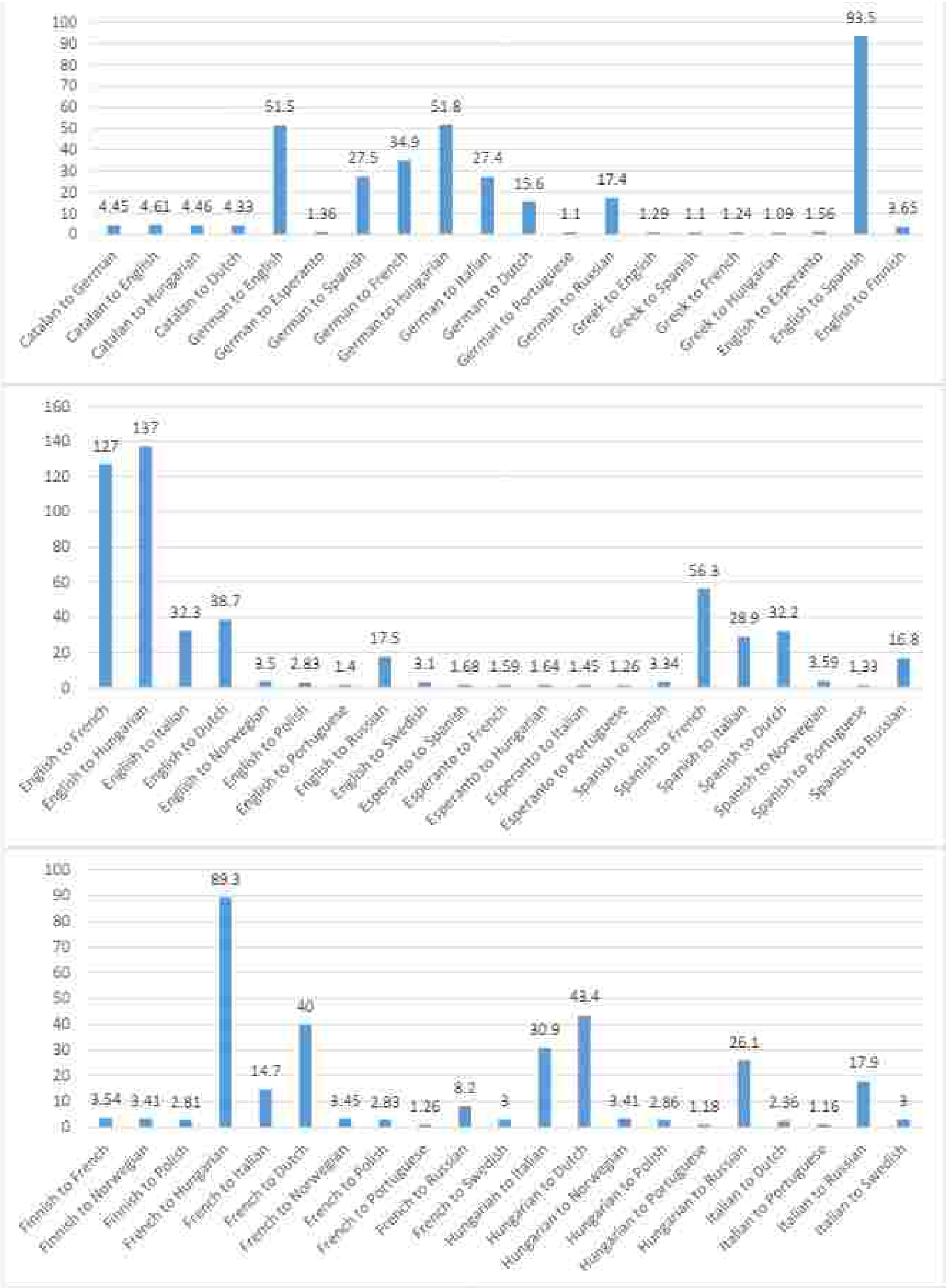

Figure 1 . Number of rows in thousands (K) of different language translations of the opus books dataset

II. METHODS

To make a better version of the M4T model, which is massively multilingual and multimodal, developed the M4Tv2 model. It is the most up-to-date version of this model that utilizes the revamped unity2 framework; it was also trained using data from many more low-resource languages. In terms of language coverage, align has been expanded to include 76 languages in total, adding 114,800 hours of automatically aligned data. It is on this basis that our two latest models are launched: expressive and streaming emanate from M4T v2. vocal style preserving translation is made possible by expression. As opposed to former forms of expressive speech research, focused on previously unexplored areas of prosody such as speed of speech and length of pauses while maintaining one's idiosyncratic speaking manner. Concerning the latter, our solution entitled 'streaming' exploits an efficient monotonic multi-head attention (EMMA) mechanism that enables the generation of low-latency target translations without having to wait for complete source utterances. A graph in another graph is called "inset" and not "insert." If you want to mean that there are alterations, the word "alternatively" should be used rather than different things alternately. Three things are important about EMMA numerical stability in estimation, alignment design, and adaptive fine-tuning. Ma et al. (2019b) [9], estimated α in the same way as before the app. It is important to note that only the infinite lookback monotonic attention model is used here [10].

A. Dataset

"Opus Books" dataset from the hugging face community given by Abhishek Kumar thakur [4]. dataset shown in Figure 1 contains over languages translations up to 17 languages, having 1,168,328 rows and the size of data is 192 mb used for giving input in the "m4t_v2_large" model checking similarity against training output [8].

B. Spacy Library

Usage of the vast corpus of examples in the sentence repository of the spacy library has been continuing. This resourceful library has enabled us to always improve our natural language processing capabilities. By using this repository, consistently fine-tuned approaches in creating NLP objects, which serve as a foundational framework for rigorous evaluations of the performance of the model [11].

Throughout time, this NLP object systematically compares output produced by the model against target examples stored in our comprehensive training database. Such a comparative process has been essential in providing insights into how well and accurately the model translates. By careful analysis that resulted in similarity scores indicating how much alike or different the outputs from the model compared to the target translations desired.

The iterative nature of this evaluation process has allowed us to continually refine and optimize the performance of the model. Each iteration comparison helps us understand better our strengths and areas for improvement within the model. By using the spacy library together with the NLP framework [12].

C. Steps of Methodology

C.1 Data Collection

Testing data has been acquired from the dataset titled "opus_books", which contains various textual collections in many languages' translation. Dataset input has been collected by this function below function named sentence_extract() which takes input as raw_datasets single row of dataset and language which decides which language sentence returned having string length greater than 300.

```
raw_datasets = load_dataset("opus_books","en-fr")
def sentence_extract(raw_datasets,lang):
  for i in range(0,1000):
    if(len(str(raw_datasets["train"][i]["translation"][lang]))>300):
     return i
```

C.2 Using *M4T_v2_large_Model*

The testing data was translated using the M4t_v2_large model developed by Meta for text-to-text translation, resulting in output strings [13]. In below code snippet "m4t_v2_large" model has been used as its translator object which predicts language translation aka text_output with target language aka tgt_lang.

```
text_output, speech_output = translator.predict(
   input = input,
   task_str = "t2tt",
   tgt_lang = tgt_lang,
   src_lang = input_lang_code3,
)
```

C.3 Similarity Score calculation

To access linguistics resources for each language involved in the translation, spacy's language datasets were used. Next, similarity scores were calculated between target (original) and output (translated) strings using spacy similarity scoring functionality calculated for each pair of target/output strings across all languages involved in the translation task. Function below lang_download() function download language by spacy library which against similarity score has been checked and nlp object has been created.

```
def lang_download(output_languages):
 for ele in output_languages:
  lang = str(ele + "_core_news_md")
```

Below code snippet for loading the nlp object from the declared language module and after that similarity score with an output of the model and target string from the dataset has been checked.

```
module = importlib.import_module(lang)
nlp = module.load()
target = nlp(target)
output = nlp(output)
```

C.4 Similarity Score calculation

The obtained similarity scores have been evaluated and analyzed to evaluate the quality and fidelity of translations, identify patterns, trends, and potential areas for improvement as well as compare performance variations across different language pairs. In addition, where possible this methodology has been authenticated by comparing these similarity scores with human evaluations or reference translations followed by interpretations of findings so that we can conclude how effective is M4T_v2_large model in text-to-text translation [14].

In the below function plot_line_graph takes input of similarity scores corresponding to each language translation and language translation gives line graph similarity score versus languages_translation in code2 format.

```
def plot_line_graph(languages, similarity_scores):
   plt.figure(figsize=(4, 3))
   plt.plot(languages, similarity_scores, marker='o', color='c',
linestyle='-')

   plt.title('Similarity Score vs. Language Translations')
   plt.xlabel('Language Translation')
   plt.ylabel('Similarity Score')
   plt.xticks(rotation=45)
   plt.grid(True)
   plt.tight_layout()
   plt.show()
```

III. RESULTS AND DISCUSSION

Using "M4t_v2_large" Meta's sequence-to-sequence model, "opus_book" language translation dataset, and spacy's python library language datasets we computed a similarity score with different combinations of language translation codes [15] as follows. Table I and Figure 2 contain similarity scores from English to Swedish, Russian, Portuguese, Polish, and Dutch; Flemish, Italian, French, Finnish, and Spanish; and Castilian translations using spacy's nlp function. Notably for example the translations to Dutch; Flemish (en-nl) Spanish; and Castilian (en-es) show extremely high similarity scores which indicates that these language pairs are well translated.

Table II and Figure 3 contain similarity scores from German to Spanish; Castilian, Flemish, French, Hungarian, Dutch; Portuguese, Italian, and Russian using spacy's nlp function with translations to Spanish; Castilian (de-es) and French (de-fr) having relatively higher scores among other language pairs. Also translated into Portuguese (de-pt), it has a lower similarity score which suggests some issues or specificities in translating from German into Portuguese.

Table III and Figure 4 contain similarity scores from French to Italian, Dutch; Flemish, Polish, Portuguese, Russian, and Swedish via Spacy's nlp function. Translations where for instance the Italian translations(fr-it) and Polish ones(fr-pl) register higher grades as an indication of good translation. Table IV and Figure 5 contain similarity scores calculated from Spanish to Finnish, French, Italian, Dutch; Flemish, Portuguese, and Russian using spacy's nlp function, with translations to French (es-fr) and Italian (es-it) showing relatively high scores indicating accurate translations. A lower similarity score in translations to Russian (es-ru) suggests possible problems while translating Spanish into Russian.

TABLE I. OBSERVATION TABLE OF LANGUAGE TRANSLATION FROM ENGLISH TO OTHER LANGUAGES

Language Translation (English to)	Language in Code2	Similarity Score
Swedish	en-sv	0.7251
Russian	en-ru	0.9573
Portuguese	en-pt	0.9695
Polish	en-pl	0.8237
Dutch; Flemish	en-nl	0.9847
Italian	en-it	0.9574
French	en-fr	0.9495
Finnish	en-fi	0.9349
Spanish; Castilian	en-es	0.9859

TABLE II. OBSERVATION TABLE OF LANGUAGE TRANSLATION FROM GERMAN TO OTHER LANGUAGES

Language Translation (German to)	Language in Code2	Similarity Score
Spanish; Castilian	de-es	0.9505
French	de-fr	0.9393
Italian	de-it	0.9085
Dutch; Flemish	de-nl	0.8929
Portuguese	de-pt	0.8050
Russian	de-ru	0.9225

TABLE III. OBSERVATION TABLE OF LANGUAGE TRANSLATION FROM GERMAN TO OTHER LANGUAGES

Language Translation (French to)	Language in Code2	Similarity Score
Italian	fr-it	0.9650
Dutch; Flemish	fr-nl	0.8810
Polish	fr-pl	0.9309
Portuguese	fr-pt	0.9162
Russian	fr-ru	0.6896
Swedish	fr-sv	0.2654

TABLE IV. OBSERVATION TABLE OF LANGUAGE TRANSLATION FROM SPANISH TO OTHER LANGUAGES

Language Translation (Spanish to)	Language in Code2	Similarity Score
Finnish	es-fi	0.9188
French	es-fr	0.9601
Italian	es-it	0.9530
Dutch; Flemish	es-nl	0.9219
Portuguese	es-pt	0.9166
Russian	es-ru	0.9217

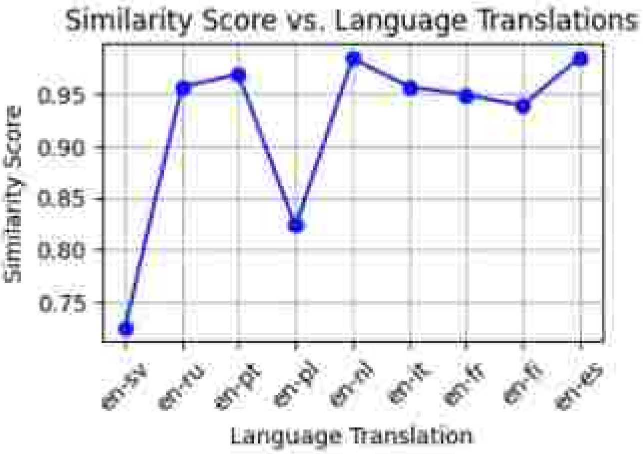

Figure 2. Line Graph Plot Between Similarity Score Versus Language Translations. (English To Swedish, Russian, Portuguese, Polish, Dutch; Flemish, Italian, French, Finnish, Spanish; Castilian)

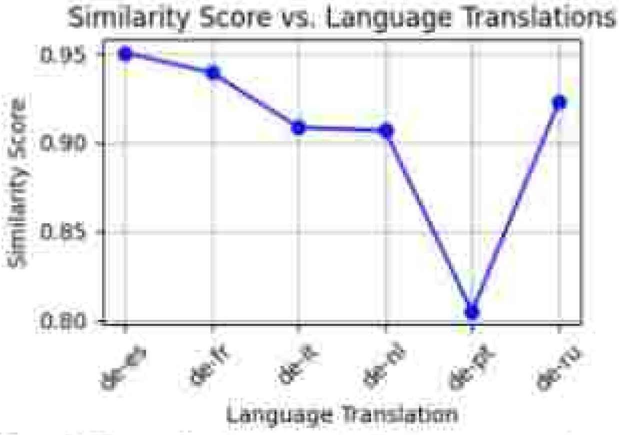

Figure 3. Line graph plot between similarity score versus language translations. (German to Spanish, Castilian, Flemish, Portuguese, Russian, French, Hungarian, Italian, Dutch)

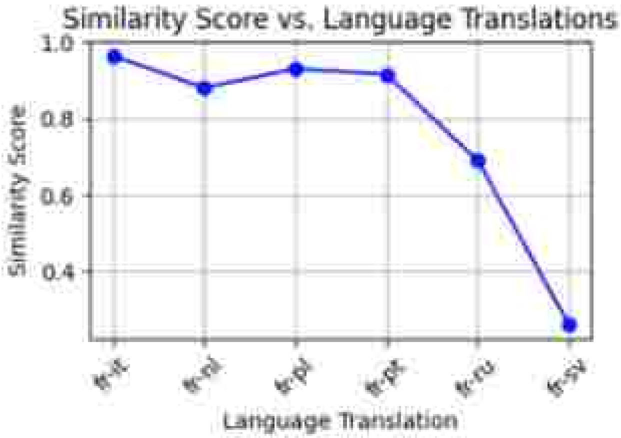

Figure 4. Line graph plot between similarity score versus language translations. (German to Spanish, Castilian, Flemish, Portuguese, Russian, French, Hungarian, Italian, Dutch)

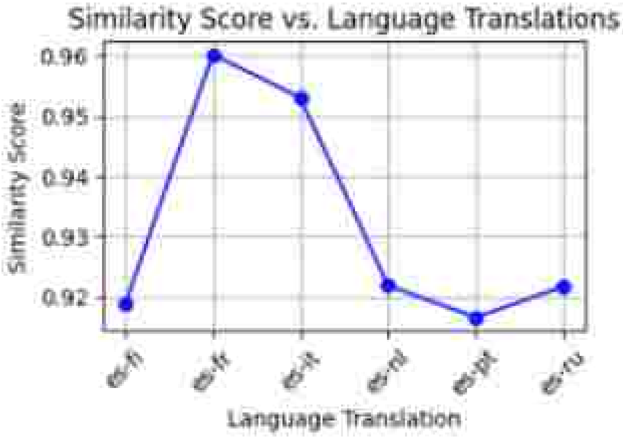

Figure 5. Line graph plot between similarity score versus language translations (Spanish to Finnish, French, Italian, Dutch; Flemish, Portuguese, Russian)

IV. CONCLUSION

Certain specific translations like English to Flemish; Dutch and Spanish; Castilian; German to French; Spanish, Castilian; and French always have very high similarity scores which means the translation is correct. But this time translation into Portuguese from German and Russian from French does not perform well as their scores are low. The performance of the model across different language pairs seems to vary largely due to source and target languages. While in some pairs results are consistently good, in other cases they can vary more. In general, although this model has potential with many languages, it still needs further development especially where there is a lesser extent of likeness between human-created versions and machine-generated implementations. Additional work could strengthen quality within all test languages through revision and examination procedures being done on them numerous times.

V. REFERENCES

[1] L. Barrault et al., ": Multilingual Expressive and Streaming Speech Translation," arXiv preprint arXiv:2312.05187, 2023.

[2] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 2837-2846, 2017.

[3] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language state-of-the-art and the challenge," Computer Speech & Language, vol. 27, no. 1, pp. 4-39, 2013.

[4] X. Ma, J. Pino, J. Cross, L. Puzon, and J. Gu, "Monotonic Multihead Attention," arXiv preprint arXiv:1909.12406, 2019.

[5] X. Ma, A. Sun, S. Ouyang, H. Inaguma, and P. Tomasello, "Efficient Monotonic Multihead Attention," in Proceedings of the IEEE, vol. 1, pp. 1-1, 2023, doi: 10.1109/TPAMI.2023.

[6] D. Cokely, "The effects of lag time on interpreter errors," Sign Language Studies, pp. 341-375, 1986.

[7] J. Hutchins, "Multiple uses of machine translation and computerised translation tools," Machine Translation, pp. 13-20, 2009.

[8] A. Thakur, "Opus Books Dataset," [Online]. Available: https://github.com/abhishekkrthakur [Aug. 2020].

[9] N. Arivazhagan, C. Cherry, W. Macherey, C.-C. Chiu, S. Yavuz, R. Pang, et al., "Monotonic Infinite Lookback Attention for Simultaneous Machine Translation," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1313-1323, Association for Computational Linguistics, 2019, doi: 10.18653/v1/P19-1126.

[10] A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.

[11] M. Honnibal and I. Montani, "Models and Languages," Feb 15, 2023, https://spacy.io/usage/models

[12] X. Ma et al., "Monotonic Multihead Attention," arXiv preprint arXiv:1909.12406, 2019.

[13] Z. Zhang et al., "Speak Foreign Languages with Your Own Voice: Cross-lingual Neural Codec Language Modeling," CoRR, vol. abs/2303.03926, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.03926

[14] C. Wang, J. Pino, A. Wu, and J. Gu, "CoVoST: A Diverse Multilingual Speech-to-Text Translation Corpus," in Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 4197-4203, European Language Resources Association, 2020, ISBN 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.517

[15] R. Ishida, "Language Codes," W3C, Jun. 13, 2014, [Online]. Available: https://www.science.co.il/language/Codes.php