Multi-Model System for Effective Subtitling Compression

Carol-Luca Gasan and **Vasile Păiș**
Research Institute for Artificial Intelligence ”Mihai Drăgănescu”, Romanian Academy
Bucharest, Romania

Abstract

This paper presents RACAI's system used for the shared task of "Subtitling track: Subtitle Compression" (the English to Spanish language direction), organized as part of "the 21st edition of The International Conference on Spoken Language Translation (IWSLT 2024)". The proposed system consists of multiple models whose outputs are then ensembled using an algorithm, which has the purpose of maximizing the similarity of the initial and resulting text. We present the introduced datasets and the models' training strategy, along with the reported results on the proposed test set.

1 Introduction

Subtitles play a vital role in ensuring accessibility and comprehension of audiovisual (AV) content for viewers with diverse needs, including those with hearing impairments or language barriers. However, traditional subtitling methods often generate text exceeding recommended reading speed constraints, hindering comprehension and viewer engagement. This problem becomes particularly pronounced for audiences with slower reading speeds or limited language proficiency.

In the context of the 21st edition of The International Conference on Spoken Language Translation (IWSLT 2024), the Subtitle Compression task, part of the Subtitling track, required participants to propose systems that rephrase subtitles that are non-compliant with the reading speed constraint without limitations on the training data conditions. This paper describes the possibility of using large language models (LLMs) to achieve this while trying to benefit from the initial content in the source language. Sometimes, sentences have formats that make them hard to compress, especially when a translation step has been made. The most fundamental example of such inconvenience is regarding idioms. They might not have perfect equivalents in the target language, and thus, their compression becomes even more challenging to process. Problems of this kind can be partially solved by initially compressing the sentence in the source language and then translating it. Our contribution is twofold: a) we introduce a new method that is able to combine the predictions of multiple models; b) we explore different parameters for the proposed algorithm and present the results on the shared task dataset.

The rest of the paper is structured as follows: Section 2 presents related work, Section 3 describes the method proposed, including dataset description (in Section 3.3), model training (in Section 3.4) and ensemble process (in Section 3.5); results are given in Section 4 and we conclude in Section 5.

2 Related work

In this section, we explore the various methodologies and research efforts that have contributed to the development of compression tasks. Although the compression task is inherently monolingual, we consider not only the works focused on text summarization but also those addressing automatic subtitling, machine translation (MT), and automatic speech recognition (ASR). This is because these domains often employ similar techniques and face comparable challenges in reducing and transforming textual data while maintaining its essential information and coherence.

2.1 Automatic subtitling

Recent advancements in speech translation (ST) have focused on developing systems that can translate spoken language directly into another language, bypassing the need for separate automatic speech recognition and machine translation (MT) steps. This approach, known as end-to-end ST, has shown promising results. Papi et al. (2023a) build on this progress by exploring the use of direct architectures for both simultaneous translation (SimulST) and

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 105–112

automatic subtitling tasks. Their work contributes to the growing body of research on efficient and effective methods for real-time speech translation applications. Bahar et al. (2023) tackle the same task, by proposing En-Ru and En-Pt production models, which support formality control via prefix tokens.

2.2 Text summarization models

Sentence compression has been extensively explored using various transformer-based architectures. The T5 model (Raffel et al., 2023) employs a text-to-text transformer architecture, leveraging its encoder-decoder structure to identify and eliminate redundant information through a process of denoising and reconstruction. Specifically, T5 uses a unified framework that converts all NLP tasks into a text-to-text format, allowing it to adapt to sentence compression tasks through task-specific prompting and fine-tuning.

BART (Lewis et al., 2020) utilizes a novel denoising autoencoder approach, where the input sentence is corrupted through token masking and deletion, and the model is trained to reconstruct the original sentence. During pre-training, BART learns to predict the original tokens from their corrupted versions, developing a robust understanding of sentence structure and semantics. This pre-training objective enables the model to develop a strong ability to recognize and remove redundant information.

The Llama2 model (Touvron et al., 2023) relies on a combination of masked language modelling and denoising objectives to learn a robust representation of language. Specifically, it uses a multi-task learning framework that jointly optimizes masked language modelling, sentiment analysis, and next-sentence prediction tasks. This multi-task learning approach enables Llama2 to develop a comprehensive understanding of language syntax, semantics, and pragmatics.

2.3 Automatic speech recognition

Automatic speech recognition (ASR) has witnessed significant advancements with the emergence of transformer-based architectures. The Whisper model (Radford et al., 2023) employs a conditional waveform-to-text model that leverages a combination of self-supervised learning and supervised finetuning to achieve state-of-the-art performance on various ASR benchmarks. It uses a multi-task learning framework that jointly optimizes masked acoustic modelling, phoneme recognition, and sentence transcription tasks, enabling it to learn robust representations of spoken language that can generalize across different accents, languages, and recording conditions.

2.4 Translation models

Machine translation has seen significant advancements with the development of large-scale transformer-based models. NLLB (No Language Left Behind) (Team et al., 2022) is a family of translation models that aim to bridge the gap between high-resource and low-resource languages. NLLB uses a multilingual masked language modelling objective to pre-train a single model on a massive dataset of 50 languages, enabling it to learn shared representations across languages and achieve state-of-the-art performance on various translation benchmarks. NLLB employs a novel "language-agnostic" approach that treats all languages equally, without relying on language-specific adapters or fine-tuning, making it particularly effective for low-resource languages.

2.5 Summarization Datasets

The development of effective text summarization models relies heavily on the availability of high-quality, linguistically diverse datasets. In this regard, the Google Sentence Compression (Filippova and Altun, 2013) dataset is a prominent resource, comprising approximately 200,000 sentence pairs extracted from news articles. Each pair consists of an original sentence and its corresponding compressed version, with an average compression ratio of 35%. Notably, this dataset is primarily composed of English sentences, with a focus on formal, written language.

TaPaCo (Scherrer, 2020) is a freely available paraphrase corpus that offers a unique resource for natural language processing (NLP) research. Extracted from the Tatoeba database, a crowdsourced platform primarily designed for language learners, TaPaCo provides a vast collection of paraphrases in 73 languages.

The PAWS-X (PAWS eXtended) (Yang et al., 2019) dataset takes a multilingual approach to text summarization, featuring a diverse range of texts from the web in four languages: English, French, German, and Spanish. With over 1 million pairs of original texts and their corresponding summaries, PAWS-X provides a comprehensive benchmark for evaluating cross-lingual summarization perfor-

mance. The dataset's structure is noteworthy, with each instance comprising a source text, a target summary, and corresponding metadata such as language labels and genre information.

3 Method

3.1 Overview

Our proposed method analyzes both the original text in English and the translated text in Spanish in order to have an alternative approach in case the latter is not being compressed within the established limits. Therefore, we had to obtain the initial subtitles in the language of the video through an automatic speech recognition model. With that in mind, we can compress and translate the English text in this exact order such that we obtain a new set of Spanish sentences to be fitted within the time intervals presented in the given SRT file. We define a sentence based on the presence of strong punctuation; a sentence may span over multiple time intervals in the SRT file. Having a series of alternatives for each sentence that has to be processed, we run an algorithm to determine the assignment of the compressed sentences that maximizes the similarity between the reference and the prediction texts. A general representation of the method is presented in Figure 1.

3.2 Performance identifiers and metrics

We focused on multiple metrics to define the performance of our models and to determine a relation of order between sentences with the same meaning.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) is a set of metrics used to evaluate the quality of summarization models. It measures the overlap between the generated summary and the reference summary, focusing on recall (i.e., how much of the reference summary is covered by the generated summary). There are several variants of ROUGE, including:

a) ROUGE-1: measures the overlap of unigrams (single words) between the generated and reference summaries;

b) ROUGE-2: measures the overlap of bigrams (pairs of adjacent words) between the generated and reference summaries;

c) ROUGE-L: measures the longest common subsequence between the generated and reference summaries.

ROUGE scores range from 0 to 1, with higher scores indicating better summarization quality.

BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is a metric used to evaluate the quality of machine translation models, but it can also be applied to summarization tasks. It measures the similarity between the generated summary and the reference summary based on n-gram overlap. BLEU calculates the precision of n-grams (sequences of n items) in the generated summary compared to the reference summary. BLEU scores range from 0 to 1, with higher scores indicating better summarization quality.

MPNet (Quyen and Kim, 2023) is a type of neural network architecture that uses word embeddings to represent words as vectors in a high-dimensional space. In this context, MPNet is used to calculate the distance between words or phrases in the generated summary and the reference summary. The distance calculation can be done using various metrics, such as cosine similarity (in this case), Euclidean distance, or Manhattan distance. The resulting distance score can be used to evaluate the semantic similarity between the generated and reference summaries.

BLEURT (BERT-based Learned Utility for Ranking Translation Outputs) (Sellam et al., 2020) is a metric that evaluates the quality of summarization models using a BERT-based approach. It learns to predict a utility score for each generated summary based on its similarity to the reference summary. BLEURT analyzes different factors, including:

a) Fluency: measures the grammatical correctness and coherence of the generated summary;

b) Relevance: measures the degree to which the generated summary covers the main points and ideas of the original text;

c) Informativeness: measures the amount of new information presented in the generated summary;

d) Coherence: measures the degree to which the generated summary is well-organized and easy to follow.

The BLEURT score is a weighted sum of these individual metrics, providing a comprehensive evaluation of the generated summary's quality.

3.3 Dataset Choice and Creation

As part of the gathered Spanish corpora, PAWS-X and TaPaCo were used as they are, while Google's Sentence Compression dataset was filtered to eliminate pairs of sentences with very low compression rate. In addition to these resources, we created a new one (Sent-Comp-ES) by translating Google's

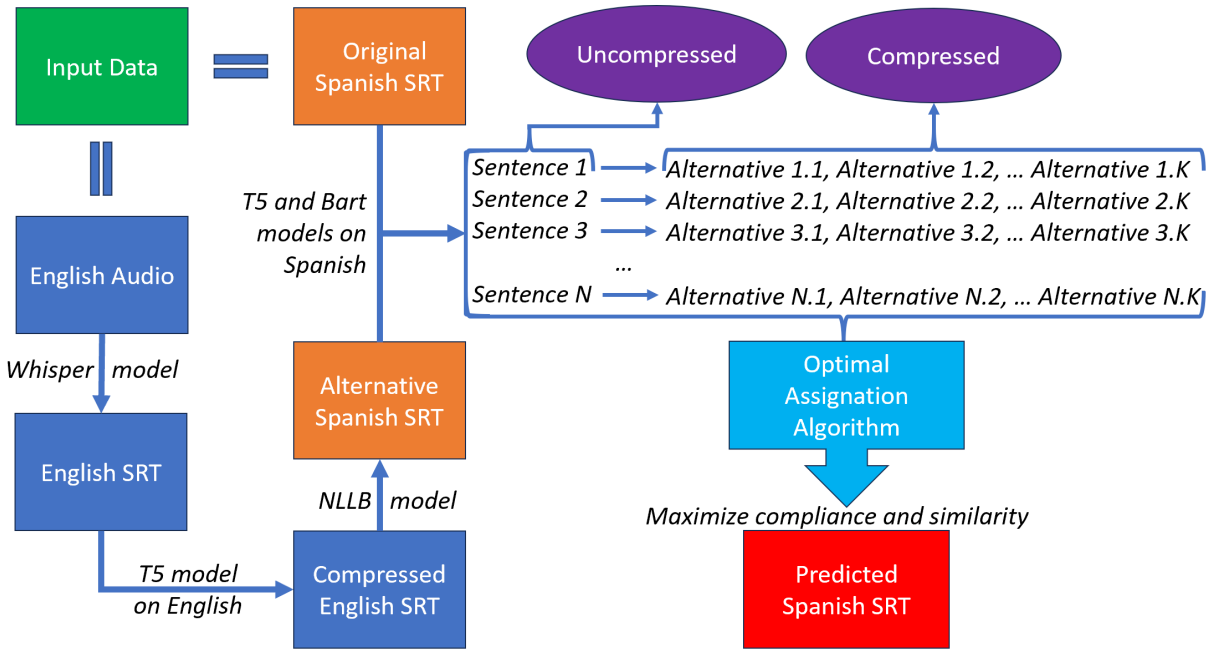

Figure 1: Scheme of the overall transformation.

Dataset	Language	Dimension
Sent. Comp.	English	200k
Sent-Comp-Es	Spanish	53k
TaPaCo	Spanish	85k
PAWS-X (filtered)	Spanish	9k

Table 1: Datasets used for extractive summarization.

Sentence Compression dataset (referred as Sent. Comp. later in the paper) for extractive summarization (i.e., the task of selecting a subset of words from a sentence to form a summary). In the transformation process, multiple rules have been established such that the quality of the data is preserved, with the downside of obtaining less data than the initial resource. The conducted steps are in exact order:

a) Eliminate the English pairs with an associated compression rate smaller than 10% for sentences with at least 10 characters;

b) Eliminate the English pairs with an associated ROUGE score smaller than 0.8;

c) Translate the remaining sentences to Spanish using Facebook's NLLB model;

d) Eliminate the Spanish pairs not respecting the extractive summarization pattern (i.e., eliminate those pairs for which the compressed sentence is not a subsequence of words from the initial sentence);

e) Check again for the associated compression rate and ROUGE score while keeping the same constraints as aforementioned;

In the end, from 200k pairs of English sentences, we formed 53k pairs of Spanish sentences that can be used for extractive summarization training. Furthermore, all processed data can be as well used for abstractive summarization.

3.4 Model Choice and Training

Since this paper focuses on an ensemble selection system, we had to define the models we want to use and train. Regarding the Spanish text models, we finetuned the base checkpoints of T5 and Bart, while for Llama2, we chose the 13B parameters checkpoint. Through the previous models, we propose to tackle both extractive and abstractive summarization. On the other hand, for the audio processing, since it can be assumed that for generating the given Spanish text, a variant of the original English text is already composed, we decided to go with a pre-trained large checkpoint of the Whisper v2 model. Wee feed the model pre-segmented audio by taking timestamps of the original Spanish SRT, without activating the internal VAD. For the English text summarization, a pre-trained large checkpoint of T5 was used.

T5 and Bart were trained on a joint dataset containing TaPaCo, PAWS-X and Sent-Comp-ES, totaling at 147k pairs of sentences, with a simple prompt, namely *"comprimir: "* (en: *"compress: "*).

Model	Learning Rate	Epoch	Avg. Compression	ROUGE	BLEU	MPNET
T5-base	1e-4	4	48%	0.60	0.23	0.81
T5-base	2e-5	4	47%	0.61	0.20	0.74
T5-base	1e-4	15	46%	0.61	0.24	0.81
Bart-base	2e-5	4	46%	0.60	0.24	0.82
Bart-base	2e-5	15	47%	0.62	0.25	0.84
Llama2-13B	1e-4	1	18%	0.57	0.23	0.80
Llama2-13B	1e-4	4	33%	0.60	0.24	0.85

Table 2: Metrics obtained on the gathered corpora while training for Spanish sentence compression.

We also finetuned Llama2 on all the data available (200k pairs) using QLoRA (Dettmers et al., 2023), with a more complex prompt trying to settle the context and the general task:

```
### TAREA: Parafrasee la frase de entrada para hacerla lo más corta posible en términos de número de caracteres, conservando el significado inicial y teniendo una gramática y puntuación correctas. Si no es posible o no está seguro, mantenga la frase sin cambios.
### SENTENCIA SIN COMPRIMIR: <UNCOMP>
### SENTENCIA COMPRIMIDA: <COMP>

(Note: the <UNCOMP> and <COMP> tokens are replacing the uncompressed and compressed sentences respectively.)
```

Table 2 contains the results acquired during training. According to the reported performance and considering Llama2's inference time, we decided to exclude it from the prediction system. Another important reason is that Llama2 was trained for abstractive summarization, which makes the reconstruction of the SRT file from sentences really difficult.

3.5 Algorithm Development

In order to present the proposed algorithm, let us standardize the problem to be solved. We have N sentences distributed among M time intervals, where a sentence might be covering multiple intervals. Each sentence can be written as a set of word sequences, representing its splits among the time intervals it overlaps. Using the summarization models, we obtain for each given sentence a set of at most K other sentences split in the same manner (possible because the extractive summarization preserves the order of the words), along with some metrics defining the resemblance to the uncompressed text. Considering known the time intervals' lengths, we can determine if a split is compliant by taking into account the dimension of the newly formed word sequence. We define the following notion as well: the score of an assignment is the weighted sum of similarity scores where the weights are length-based. The score is between 0 and 1, a score of one being obtained for the initial sentences. The length of a sentence is defined as the number of characters.

A baseline approach is to go through all the possible combinations of assigned sentences and choose the one with a maximal score that is also compliant. The complexity of this algorithm is in terms of $O((M+N)*K^N)$. Our proposed algorithm achieves a complexity of $O((M+N)*K*\alpha)$, where α is the maximum length of a split. The main idea of the algorithm is to denote critical points as the time intervals that contain words from more than one sentence. Then, we just have to analyze the best obtainable score until a certain checkpoint, while consuming a certain number of characters from the maximum allowed within that time interval. This is achievable using dynamic programming and it reduces the complexity to the one previously mentioned. The pseudo code for obtaining the maximum score can be seen in Figure 2. The optimal solution can be easily reconstructed by maintaining a backward array during the update of the *dp* array, which allows backtracking from the final state to the initial state to retrieve the sequence of selected sentences.

4 Results

The dev set proposed within the shared task consists of 7 SRT files, part of the EuroParl Interviews (EPI) en-es test set, whereas the test set concerns AV docs from the ITV entertainment series, all generated by the non-participating (Papi et al., 2023b). The reported results of our submission can be seen in Table 3, where ChrF is a metric introduced by

Algorithm 1 Compute Maximum Score

Input:
`capacity` - array of size $M + 1$ (stores time interval capacities)
`interval` - array of pairs (stores start and end times for each sentence)
`quality` - matrix of size $(N + 1) \times (K + 1)$ (stores similarity scores)
`quantity` - 3D matrix of size $(N + 1) \times (K + 1) \times (\texttt{interval}[i].\text{right} - \texttt{interval}[i].\text{left})$ (stores word sequence lengths)
`dp` - matrix of size $(N + 1) \times$ (maximum capacity across intervals) (stores maximum score achievable)
Initialize `dp`
`dp`$[0][0] = 0$ ▷ base case, no sentences processed
for $i = 1$ to N **do**
    `dp`[i] $\leftarrow$ array filled with $-\infty$ ▷ initialize scores for current sentence
end for
Loop through sentences
for $i = 1$ to N **do**
    if `interval`[i].left $\neq$ `interval`[i - 1].right **then**
        Get maximum score from previous sentence ▷ no addition possible
        for $j = 1$ to K **do** ▷ loop through candidate sentences for current
            Update `dp` with score from previous + current sentence similarity
        end for
    else if `interval`[i].left $<$ `interval`[i].right **then**
        for $j = 1$ to K **do** ▷ loop through candidate sentences for current
            for last $= 0$ to `capacity`$[i$.left$] -$ `quantity`$[i][j][0]$ **do** ▷ check if candidate fits
                Update `dp` with score from previous + current sentence similarity
            end for
        end for
    else
        for curr $=$ `capacity`$[i$.left$]$ down to 0 **do** ▷ loop through capacities
            for $j = 1$ to K **do** ▷ loop through candidate sentences for current
                if candidate fits current capacity **then**
                    Update `dp` with score from previous + current sentence similarity
                end if
            end for
        end for
    end if
end for
Find maximum score across all capacities for the last sentence
max_score $\leftarrow -\infty$
for $i = 0$ to `capacity`$[M]$ **do**
    max_score $\leftarrow$ max(max_score, `dp`$[N][i]$)
end for
Output: max_score

Figure 2: Pseudo code for obtaining the maximum score.

Method	BLEU	ChrF	TER	BLEURT	CPS
ref	-	-	-	-	89.98
ori test-set	8.71	29.18	81.08	0.213571	69.97
baseline	7.70	27.52	81.27	0.18917	100.00
RACAI	7.51	26.60	80.33	0.194613	94.29

Table 3: Reported results on the proposed test set.

(Popović, 2015) , and TER (Translation Edit Rate) represents the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references. In addition, the methods' acronyms in Table 3 respect the following notations:

a) ref: reference subtitles used to compute BLEU/ChrF/TER/BLEURT scores;

b) ori test-set: original subtitles to be compressed;

c) ori test-set-1line: original subtitles where those segmented in more lines are unsegmented in 1-line;

d) baseline: hard cut at max number of charsq compatible with subtitle duration;

e) RACAI: subtitles generated with the system described in this paper.

5 Conclusion

This paper presents RACAI's system for the "Subtitling track: Subtitle Compression" shared task, focusing on compressing subtitles from English to Spanish while maintaining readability within reading speed constraints. Our system leverages multiple large language models (LLMs) to generate alternative compressed sentences for the original text. An ensemble selection algorithm then chooses the most suitable compressed options based on similarity metrics. This approach allows us to benefit from the strengths of various models and address potential shortcomings of individual models.

Future work could explore the incorporation of additional metrics or quality estimation techniques within the ensemble selection algorithm. Additionally, investigating the effectiveness of the system on different language pairs or domains could be valuable, such as including the Romanian language. We previously had an interest for processing Romanian language speech using Whisper (Gasan and Păiș, 2023). Overall, this work contributes to the development of automatic subtitling systems that ensure accessibility and comprehension for diverse audiences.

References

Parnia Bahar, Patrick Wilken, Javier Iranzo-Sánchez, Mattia Di Gangi, Evgeny Matusov, and Zoltán Tüske. 2023. Speech translation with style: AppTek's submissions to the IWSLT subtitling and formality tracks in 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 251–260, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.

Carol-Luca Gasan and Vasile Păiș. 2023. Investigation of Romanian speech recognition improvement by incorporating Italian speech data. In *The 18th International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR-2023)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023a. Direct Speech Translation for Automatic Subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Sara Papi, Marco Gaido, and Matteo Negri. 2023b. Direct models for simultaneous translation and automatic subtitling: FBK@IWSLT2023. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 159–168, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Van Toan Quyen and Min Young Kim. 2023. Mpnet: Multiscale predictions based on feature pyramid network for semantic segmentation. In *2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 114–119.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.