Eng2Span: English-to-Spanish Translation with Word Level Confidence Scores

Emma Gifford
Boise State University
emmagifford@u.boisestate.edu

Dylan Gresham
Boise State University
dylangresham@u.boisestate.edu

Abstract

This paper presents Eng2Span, a neural machine translation system designed to translate English text into Spanish while providing word-level confidence scores. Unlike traditional translation tools, Eng2Span aims not only to generate accurate translations, but also to enhance the user's understanding by indicating the model's certainty for each translated word. We have fine-tuned multiple multilingual transformer-based models including MBart, Opus MT, M2M-100, and NLLB-200, on a diverse set of English-Spanish parallel corpora sourced from Kaggle, Hugging Face, and OPUS.

After evaluation using BLEU, METEOR, ROUGE, and TER metrics, the best-performing model is integrated into an interactive Tkinter-based application. The application enables users to input English text, view the Spanish translation, and assess confidence scores derived from output logit distributions. This feedback allows users to identify potentially unreliable translation segments, promoting more informed and effective communication in bilingual contexts. Preliminary evaluations, including human assessments and comparisons with established tools such as Google Translate, suggest that our system offers a blend of accuracy and transparency with plenty of room for improvement.

1 Introduction

Neural network machine translation (NMT) has been used by leading companies such as Google and Microsoft for many years now. Google started as a statistical model based on SYSTRAN (Toma, May 1997) in 2006, and became an artificial neural network in 2016 Wu et al. (2016). The advantage of using a neural network in machine translation is its ability to consume an entire sentence, remember the context, and output a relevant sequence of words in another language.

However, current large-scale machine translation does not provide a way for users who do not know both the source language and the target language to understand the potential accuracy of the translated content. Users who are in the process of learning the target language may be misled by incorrect translations.

In this paper, we work towards addressing this problem by enriching several multi-translational models with English to Spanish translation data. From these fine-tuned models, we select the model that performs the best based on our chosen set of metrics: BLEU (Post, 2018), METEOR (Banerjee and Lavie, 2005; Papineni et al., 2002), ROUGE (Lin, 2004), and TER (Snover et al., 2006). We then extract and transform the logit output of our selected model to generate a confidence score for each translated word. To make the confidence scores available to users, we map the score to a color and paint the translated words to match the confidence scores.

In Section 2, we discuss the background and previous work done in this area. In Section 3, we discuss the datasets used to fine-tune our models and the process we used to clean and combine our data. In Section 4, we discuss our question and the process we used. In Section 5, we discuss how we evaluated the performance of our models and the accuracy of our confidence scores. In Section 6, we discuss what this work could do for the general public. Finally, in Section 7 we provide a summary of our work and provide ideas on future work that could be done.

2 Background

Machine translation is the act of using a computer to map a piece of text from one language to another. Wu et al. (2016) leverages the power of Long Short-

https://github.com/Dylan-Gresham/Eng2Span
https://huggingface.co/dmidge/mbart-large-50-eng2span

Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) and gradient flow (He et al., 2016) to compensate for the shortcomings of statistical machine translation (SMT) (Brown et al., 1988). In 2023, Vaswani et al. (2023) introduced the transformer architecture which outperformed every other architecture at the time while taking less time and resources to construct. Following Vaswani et al. (2023), many papers such as Bao et al. (2021) and Ahmed et al. (2017) have taken the transformer architecture and produced specialized versions of it specifically for machine translation tasks.

Confidence scores are a method of predicting the correctness of a model's output. Lu et al. (2022) proposed learning confidence scores via a second model that gets trained at the same time as the NMT model. Cui and Liang (2024) take a similar approach except instead of training an additional model at the same time, they use BERT models tailored for their task to evaluate translations. Mandelbaum and Weinshall (2017) take a similar approach to ours but, their approach involves computing the Euclidean distance from a point to its neighbors in the training set then estimating based on that density.

Our approach builds off of the M2M (Fan et al., 2020), MBart (Tang et al., 2020), NLLB (Team et al., 2022), and Opus (Tiedemann, 2020) transformer-based machine translation models to provide a unique interpretation of the model's logits as a means of providing confidence. We interpret logits as a learned estimate of distance and use a simple computation to derive confidence scores under this assumption.

3 Data

We utilized English translations from several different sources, combining them into a single comprehensive corpus. These data sets allowed us to get a range of different contexts and specialties while also reinforcing common vocabulary words.

The first data set that we sourced from Kaggle consists of simple English phrases and their Spanish translations (Lonnie, 2021). Some of the English phrases in this data set have several occurrences with different Spanish translations to account for multiple ways of conjugating verbs. This data set gave us $102,904$ sentences pairs.

Two data sets on Hugging Face were available from the Language Technology Research Group at the University of Helsinki. This included the Open Parallel Corpora (OPUS) KDE4 and the OPUS Books data sets from Tiedemann (2012).

The OPUS KDE4 data set contains ninety-two parallel language entries. For the English to Spanish language pairs, the KDE4 data set gives us $218,655$ sentences from the KDE4 localization. This data set focuses on translation for computer applications and computer application documentation. The language phrase pairs are on technical subjects such as computer menu options.

The OPUS Books data set includes many parallel book translations. The English to Spanish pairs were formatted by an id per sentence, and then an associated dictionary with the respective English and Spanish sentence translations. The books for this data set are from the collection of copyright-free books compiled by Andras Farkas. The books in the data set are freely available for personal, educational, and research use. The first book featured in the English to Spanish pairs is Jane Austen's Sense and Sensibility. In total, there are $93,470$ sentences pairs in this data set.

The final data set that we utilized is Google's WMT24++ (Deutsch et al., 2025) data set. This data set contains human translation and post-edit data for 55 English to another language pairs. Each data set row has a language pair identifier (`lp`), a `domain` (of "canary", "news", "social", "speech", or "literacy"), a `document_id`, a `segment_id` to globally identify the segment, `is_bad_source` to indicate lower quality sources, the `source`, the `target`, and the `original_target` reference translation. The data set suggested that we remove rows that are marked as `is_bad_source`, as they were removed from the evaluation in the original paper.

Together, these data sets contained a wide range of topics, covering general conversation, news sources, speeches, books, and technical computer text. This gives our model a varied selection of conversational contexts which it could be useful in.

To preprocess the data, we selected the language pair we wanted to use. Namely, the English and Spanish portions of the data. We then processed each data set individually by removing rows with missing or null values, dropping extra or unnecessary columns, and normalizing the "source" and "target" column names to "English" and "Spanish" before merging them into a single complete data set. Once combined, we removed duplicate entries and allocated 60% of the data for training, 20% of the data for validation, and the remaining 20% of the data for testing. The exact numbers for the data splits can be seen in Table 1.

Split	Count
Train	254,324
Validation	63,582
Test	79,477

Table 1: Number of samples per data split.

Five randomly selected samples from each of the splits can be seen in Table 2 with the proportions of the samples from each split consistent with the true ratios that we utilized. We structured our data set such that each row contains the English text, the corresponding Spanish translation, and then the data split which the sample is in.

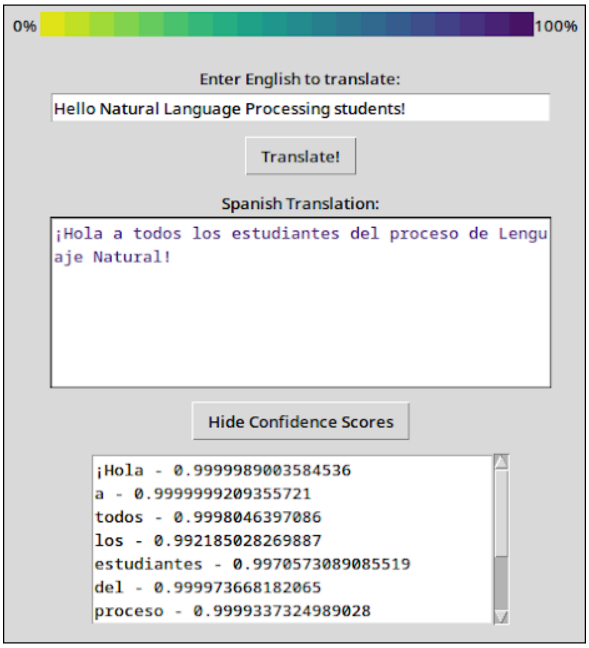
English	Spanish	split
Player settings	Preferencias del reproductor	train
Tom won a free car.	Tom se ganó un auto gratis.	train
I lost the game.	Perdí la partida.	train
You deserve a medal.	Te mereces una medalla.	validation
Why do I have to do it?	¿Por qué tengo que hacerlo?	test

Table 2: Samples from our full dataset.

4 Method

We set out to create a model specializing in English to Spanish translation that can give users a per-word score within the translation. To accomplish this, we fine-tuned several models on English to Spanish data. We then compared these models against several metrics, BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and TER (Snover et al., 2006). We then selected the model with the highest scores as the model to be used in our application.

To make the confidence score outputs more friendly for users, we incorporated the model and its confidence score outputs into a prototype graphical user interface (GUI). This GUI is meant to provide an example of how machine translation models and confidence scores can be utilized in the real world to enhance conversational quality between people of differing primary languages. Having a GUI allows users to have a simple interface for translation where their translations are colored according to confidence. Additionally, we have provided a button that allows users to view the confidence scores directly. Our prototype can be seen in Figure 1.



Hello Natural Language Processing students!

¡Hola a todos los estudiantes del proceso de Lengu
aje Natural!

```
¡Hola - 0.9999989003584536
a - 0.9999999209355721
todos - 0.9998046397086
los - 0.992185028269887
estudiantes - 0.9970573089085519
del - 0.999973668182065
proceso - 0.9999337324989028
```

Figure 1: Our GUI prototype.

Looking further into the logits produced by our model, we noticed that all of the values were extremely low in value. This was interesting as we have typically seen logits that are high in value rather than low. This sparked the question of: What are the logits representing? Our interpretation of logits is that they're a learned estimate of the token's deviation from the truth. In other words, the model is learning a way to predict the distance for each token from the "correct" token. Under this assumption, we defined our formula for per-word confidence scores, c_{score}, as the following.

$$c_{score} = \text{softmax}(1 - logit)$$

Since the output is for a tokenized sequence, we also need to handle the case where a token represents part of a word, a sub-word, and not a full word. To do so, we simply take the average confidence score, not logit value, for each sub-word and use the average confidence score as the score for the overall word.

5 Evaluation

In this section, we will discuss how we evaluated the fine-tuned models, our confidence score derivation method, and how our best model compares to standard tools.

5.1 Task & Procedure

For the task at hand, we selected a pool of models to fine-tune, so that we could select the best one

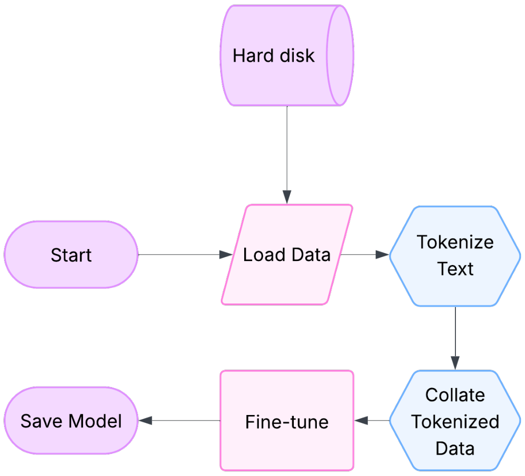


Figure 2: Fine-tuning flow chart.

for our final iteration. Our model pool consisted of the M2M (Fan et al., 2020), mBART (Tang et al., 2020), NLLB (Team et al., 2022), and Opus (Tiedemann, 2020) pretrained transformer-based machine translation models.

We took these pretrained models hosted on HuggingFace and performed fine-tuning on them. We conducted fine-tuning using the `Seq2SeqTrainer` from the Transformers (Wolf et al., 2020) library created by HuggingFace. As we have a larger dataset, we used a relatively low learning rate of 0.00002 along with a weight decay of 0.01 over 3 epochs with each batch containing 32 samples from our dataset. For the training metrics, we utilized all four of the metrics mentioned earlier, BLEU, METEOR, ROUGE, and TER. Each of the four models were fine-tuned individually utilizing Boise State's Borah cluster. Figure 2 shows the overall flow of our fine-tuning process.

At the end of model training, we saved each model in its own file, along with its checkpoints, configuration, and run logs. We also logged the model's final performance on each of the metrics at the end of the fine-tuning evaluation step.

After fine-tuning, we compared each model's performance on the metrics from their baselines and selected the best performing model to run behind our graphical user interface.

5.2 Metrics & Baselines

A translative model that produces a confidence score for each word is great, but how can we be sure that the model's confidence is accurate? For this, we turned to using a combination of language comparison metrics and human evaluation.

The metrics we used when selecting a model from our fine-tuned pool were BLEU, ROUGE, METEOR, and TER. We chose these four metrics because they tend to align with human judgment and are easily reproducible. Furthermore, the metrics complement each other. BLEU assesses at the corpus level, METEOR assesses at the sentence-level, ROUGE assesses content recall and coverage, and TER assesses how much effort would need to be put into the translation after being generated.

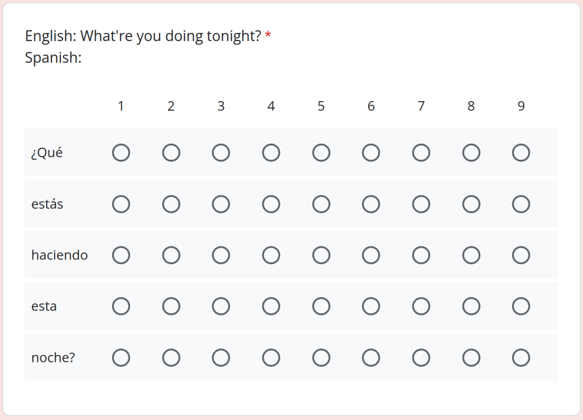
To create baselines, we took each model we selected for our fine-tuning and ran the evaluation step we would use after fine-tuning to generate metrics. This determined each model's baseline.

5.2.1 Human Generated Metrics

In addition, we created a form for human evaluation. We chose a model to utilize within our GUI program, and then we wrote twenty-six English sentences. We then ran each of the sentences through the chosen model to generate full English-Spanish translation pairs. These translation pairs were inputted into a Google Form to allow users to provide confidence scores per word, similar to the model's output. As it would be tedious for participants to enter values as floating point numbers or provide enough check-boxes to cover all possible values, form responders could not provide as fine-grained confidence for each word as the model was able to. As seen in Figure 3 we set up the form as a multiple choice grid, where the rows represented each word in the Spanish translation and the columns represented a value from one to ten. These columns allowed respondents to choose a confidence score for each word from one to ten. One being the lowest confidence, and ten being the highest confidence. Once responses are gathered, the scores for each word will be averaged across all respondents and then compared with the selected model's confidence in that word. In doing so, we aim to evaluate how close our method of deriving confidence scores is to reality.

This form was sent out to nine people with varying Spanish skills. Most of the respondents had several years of Spanish classes at the high school or college level, and three respondents used Spanish regularly or seasonally to communicate with family or colleagues.

In hindsight, there were several issues with our response form. Some of these problems we realized ourselves after the fact, and others came from the



English: What're you doing tonight? *
Spanish:

	1	2	3	4	5	6	7	8	9
¿Qué	○	○	○	○	○	○	○	○	○
estás	○	○	○	○	○	○	○	○	○
haciendo	○	○	○	○	○	○	○	○	○
esta	○	○	○	○	○	○	○	○	○
noche?	○	○	○	○	○	○	○	○	○

Figure 3: Google Form for Human Evaluation

response of form respondents.

During a conversation about the form, we realized that we should have asked the respondents for more than just a confidence score for the response. With the format of the form we sent out, we treated our human respondents as if they were another machine learning model that gave us potential confidence scores at the word level. During form creation, we had considered allowing respondents to provide notes about each translation or their own translation for each language pair. However, we did not want to fatigue respondents with the already lengthy twenty-six form sentence questions. Additionally, once we realized our pool of participants consisted of more non-experienced Spanish speakers than we had thought, we believe we should have included a piece of the form asking for the respondent's self-rating of their Spanish level.

Feedback from respondents said they wanted ways to communicate when the model had provided a generally correct output, but had perhaps missed a word or had an incorrect word in the middle. Some respondents penalized every word after the missing or mistaken word in the model's translation. Other respondents wanted to be able to provide the translation they would give for the English sentence. Mostly, form respondents needed a free form way to communicate ideas about each translated sentence.

5.2.2 Comparison to Standard Tools

Our third method of evaluation is comparing our model's output with Google Translate and Microsoft Translator, which are standard tools used by many people for machine translation. To compare our model with these two tools, we will generate translations using the 26 questions in our survey and then compare the translations to each other using the four benchmark metrics BLEU, METEOR, ROUGE, and TER. In doing so, we aim to show that our model is comparable to both tools.

5.3 Results

In this section, we will first analyze the benchmark metrics, BLEU, METEOR, ROUGE, and TER, to determine which model we used in our GUI before analyzing the results of our confidence score survey with the best performing model. Finally, we examine how our selected model compares to industry standard tools such as Google Translate and Microsoft Translator.

5.3.1 Benchmark Metric Results

The results of this experiment can be found in Table 3. It is important to note that the Opus model (Tiedemann, 2020) was pretrained on the Opus data sets (Tiedemann, 2012) we used during fine-tuning. Due to this, the Opus model's scores on our chosen metrics increased only slightly.

Metric	BLEU	METEOR	ROUGE	TER
Model				
M2M Base	0.260185	0.486152	0.510175	65.919885
M2M FT	0.407935	0.650576	0.674081	**50.590047**
MBart Base	0.029995	0.100678	0.054524	110.127757
MBart FT	**0.427286**	**0.660292**	**0.680227**	50.806106
NLLB Base	0.024615	0.093240	0.079818	106.541898
NLLB FT	0.140845	0.288530	0.285709	95.202080
Opus Base	0.228956	0.588228	0.621103	101.856228
OPUS FT	0.408307	0.655364	0.676027	52.922702

Table 3: Benchmark results. Best scores per-column are bolded.

Based on the scores shown in Table 3, our fine-tuned mBART model is the best performing model, having the best scores on BLEU, METEOR, and ROUGE metrics and falling behind the fine-tuned M2M model on the TER metric by a measly 0.216959. In addition, mBART has shown the greatest improvements across all metrics in comparison to the other models. mBART showed a 53.8863% improvement on TER, 324.53% improvement on BLEU, 555.845% on METEOR, and 1147.58% improvement on ROUGE. From all of this information, we can confidently say that mBART is the model that provides the highest quality translations from the models that we have tested and will be the model that we utilize for our GUI and in our survey.

5.3.2 Survey Results

We ended up with a 77.8% (7/9) participation rate with our survey. After collecting our results into pandas (pandas development team, 2020) DataFrames for each question, a sample of which can be seen in Table 4, we computed the average confidence score error and the thresholded accuracy. We deemed scores within 5% as acceptable and used that as our threshold value.

We found our approach has an average error of 1.408 and a thresholded accuracy of 23.02%. From these numbers, we can determine that our approach to derive confidence scores is **not** accurate.

Spanish Word	Mean Score	Model Score
Él	9.14	10.0
estaba	9.57	10.0
agradecido	9.00	9.87
por	9.86	10.0
el	9.86	9.97
peso	9.86	10.0
y	9.29	10.0
la	9.86	10.0
calentura	8.43	9.44
de	9.86	10.0
la	9.86	10.0
chaqueta.	9.57	10.0

Table 4: Result from the final question in the survey.

5.3.3 Industry Tool Comparisons

We compared the 26 translations from the survey to translations generated using the exact same English sentences from both Google Translate and Microsoft Translator. The metric scores can be seen in Table 5. From the table, we can see that the best scores come from comparing the tools to each other, with the scores from our model to each of the tools being about 10 less than the best score per column. Noticeably, all of the scores are on the higher side of the possible values. This indicates that while our model might not be up to scale, it does not fall too far behind the standard tools from large companies.

6 Implications

The results from Eng2Span suggest significant promise for enhancing machine translation with interpretable outputs, particularly for language learners and bilingual communication contexts. By incorporating word-level confidence scores, our system directly addresses a major limitation of existing NMT tools: the inability for end-users to assess translation reliability without fluency in both languages. This feedback offers users an easy way to critically engage with machine-generated translations.

	BLEU	METEOR	ROUGE	TER
Us/Google	0.517850	0.757513	0.772373	30.708661
Us/Microsoft	0.531258	0.735156	0.757298	30.798479
Microsoft/Google	**0.658967**	**0.862579**	**0.835353**	**21.259843**

Table 5: Metric scores for our model vs Google Translate and Microsoft Translator as well as Microsoft Translator and Google Translate for comparison. Best score per-column is bolded.

Our evaluation uncovered a gap between model-derived confidence scores and human-perceived translation quality. The average error of 1.408 and a low thresholded accuracy of 23.02% in our human evaluation indicate that while the MBart model performs well on standard translation benchmarks, the confidence scores it produces may be misleading or overconfident. This overconfidence is consistent with known issues in neural models and suggests the need for more calibrated or even hybrid approaches to confidence estimation, possibly integrating uncertainty quantification techniques or perhaps ensemble-based measures.

7 Conclusions & Future Work

In this project, we developed and evaluated Eng2Span, a neural machine translation system that not only performs English-to-Spanish translation but also generates per-word confidence scores derived from logit outputs. After fine-tuning several multilingual transformer models, we selected MBart as our best-performing model based on BLEU, METEOR, ROUGE, and TER metrics. We implemented an interactive GUI to present translations alongside a color gradient that represents the model's confidence scores, creating a more informative and user-centric translation experience.

Our experiments indicate that, while MBart delivers strong translation performance, our current method for generating confidence scores does not reliably reflect user-perceived translation accuracy. This highlights an important area for future research, refining confidence estimation. Potential directions include calibrating logits prior to computing a confidence score, incorporating uncertainty modeling techniques, or fusing logit-based scoring with external quality estimation tools.

Additionally, improvements to the human evaluation process, such as a larger pool of participants, broader participant qualifications, richer feedback forms, and participant annotation support, would offer better insight into model behavior and users' expectations.

Ultimately, our project demonstrates that accessible and interpretable translation tools are feasible. Future iterations of this work could extend beyond English-Spanish pairs, incorporate real-time translation contexts, and explore use cases in various industries.

References

Karim Ahmed, Nitish Shirish Keskar, and Richard Socher. 2017. Weighted transformer network for machine translation.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1988. A statistical approach to language translation. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.

Yizhuo Cui and Maocheng Liang. 2024. Automated scoring of translations with BERT models: Chinese and english language case study. *New Developments in Computational Linguistics to Support Decision Making*, 14(5):1925. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages Dialects.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. ISSN: 1063-6919.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

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

Lonnie. 2021. English-spanish translation dataset.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. Learning confidence for transformer-based neural machine translation.

Amit Mandelbaum and Daphna Weinshall. 2017. Distance-based confidence score for neural network classifiers.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

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 (ACL)*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

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.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Dr P P Toma. May 1997. SYSTRAN AS a MULTILINGUAL MACHINE TRANSLATION SYSTEM. *Overcoming the language barrier, Verlag Dokumentation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.