Creating and Evaluating a Multilingual Corpus of UN General Assembly Debates

Hannah Béchara
Hertie School
Berlin, Germany
bechara
@hertie-school.org

Krishnamoorthy Manohara
Hertie School
Berlin, Germany
manohara
@hertie-school.org

Slava Jankin
University of Birmingham
Birmingham, UK
v.jankin
@bham.ac.uk

Abstract

This paper presents a multilingual aligned corpus of political debates from the United Nations (UN) General Assembly sessions between 1978 and 2021, which covers five of the six official UN languages: Arabic, Chinese, English, French, Russian, and Spanish. We explain the preprocessing steps we applied to the corpus. We align the sentences by using word vectors to numerically represent the meaning of each sentence and then calculating the Euclidean distance between them. To validate our alignment methods, we conducted an evaluation study with crowd-sourced human annotators using Scale AI, an online platform for data labelling. The final dataset consists of around 300,000 aligned sentences for En-Es, En-Fr, En-Zh and En-Ru. It is publicly available for download.

1 Introduction

Multilingual corpora are valuable resources for natural language processing (NLP) research and applications, as they enable the development and evaluation of cross-lingual and low-resource models and systems. However, creating and maintaining large-scale and high-quality multilingual corpora is a challenging task, as it involves collecting, processing, and aligning texts from multiple languages and domains, while ensuring their accuracy, consistency, and relevance. In this paper, we align and evaluate a multilingual corpus that is based on the plenary sessions of the United Nations (UN) General Assembly, which is the main organ of the UN where all member states have equal representation and voice. The plenary sessions are held every year and are translated and transcribed in the six official languages of the UN: Arabic, Chinese, English, French, Russian, and Spanish. These sessions cover a wide range of global issues, such as peace and security, human rights, development, climate change, and health, and reflect the views and positions of different countries and regions on these issues. Therefore, our corpus provides a rich source of multilingual texts within the political domain that can be used for various NLP tasks, such as machine translation, in-domain text classification, question-answering, and multilingual argument mining.

We describe the methods we used to collect, clean, segment, and align the plenary sessions across languages. We then cover our approach to bilingual sentence alignment using embeddings and Euclidean distance, and the special considerations and difficulties we encountered for the different languages. For example, we faced some challenges in aligning Arabic with the other languages, due to technical issues in converting the Arabic documents into a suitable format for alignment. We also noticed some differences in the order and structure of sentences across languages, which made the alignment more difficult. We discuss how we addressed these challenges and how we validated the quality of our alignment.

2 Related Work

Previous multilingual parallel corpora have been based on United Nations data, including the MultiUN (Eisele and Chen, 2010) and the United Nations Parallel Corpus (Ziemski et al., 2016).

In the MultiUN Corpus, data was retrieved from the United Nations Official Document Sys-

tem (ODS), a web-based repository of official documents of the United Nations. The data was filtered by publication symbols, which are unique identifiers that indicate the issuing body, the type of document and the year of publication. The paper selected documents with publication symbols that correspond to official records and other parliamentary documents of the UN. The multilingual sentence alignment starts with pairwise alignments based primarily on sentence lengths and then on a dictionary. Pairwise alignments are available, and later merged into multilingual alignments across all six languages. The updated version of MultiUN, v2, contains documents up to and including 2011 (Chen and Eisele, 2012).

The United Nations Parallel Corpus (UNPC) is composed of official records and other parliamentary documents of the United Nations that are in the public domain. It contains sentence-level alignments for content between 1990 and 2014. The corpus contains 799,276 documents in six languages and contains 86,307 documents that have translations across all six languages. The sentence-level alignments were generated using GIZA++ (Och and Ney, 2003) to align sentences based on word co-occurrences. The authors validate their dataset through a number of machine translation baselines, with BLEU scores results varying between 29 and 61, depending on the language pair.

While robust, neither of these corpora offer a full evaluation of the accuracy and precision of their alignment, nor are they recent enough to include the later documents. Furthermore, sentence alignment methods have come a long way since their collection. More modern methods of multilingual sentence alignments are based on multilingual pretrained language models, such as mBERT, that can learn cross-lingual representations of sentences. These methods have been shown to outperform GIZA++ (Schwenk, 2018; Guo et al., 2018). Artetxe and Schwenk (2019) use a sequence-to-sequence architecture to train a multilingual sentence encoder on an initial parallel corpus. The encoder maps sentences from different languages into a shared embedding space, where similar sentences are close to each other. The authors then use a margin-based scoring method to measure the similarity between sentence embeddings. The authors evaluate their their method on three tasks, the BUCC mining task, the UN reconstruction task, and the ParaCrawl filtering task, and show that the proposed method outperforms existing methods on all three tasks by a large margin.

3 Corpus Collection

The United Nations (UN) plenary meetings are meticulously recorded in each of the six official UN languages, making them an ideal source for a multilingual corpus. The records are then made available on the official website[1], in the form of PDF files, separated by language and session. All the documents are public domain. We downloaded the documents in all 6 languages and converted them using an OCR-based tool in a pdf editor, discarding pictures, tables and style markers. In total, we processed 2113 documents between 1978 and 2021.

However, due to the age of the documents and the limitations of the OCR-based tool, we were unable to convert enough Arabic-language documents for use in alignment. Furthermore, the documents we did manage to convert were of poor quality. As a result, we were unable to align the Arabic sentences and eventually discarded the Arabic language documents until such a time that we can properly convert them.

We then processed these documents using the Natural Language Toolkit (NLTK) package (Bird et al., 2009). We used the toolkit to separate the documents into individual sentences, as the documents only provide paragraph boundaries. We then tokenised the sentences and removed stop words to made them ready for alignment.

4 Sentence Alignment

The documents described in Section 3 are not translated at the sentence level, but rather at the level of individual speeches taken as a whole. This means that each speech in one language has a corresponding speech in another language, but not necessarily each sentence. Therefore, in order to create a parallel corpus at the sentence level, we need to match each sentence in one language with the equivalent sentence in another language. This is a challenging task, as the sentences may not have the same order, structure, or length across languages. Furthermore, translations are not always a one to one mapping. Sometimes a sentence can be represented by multiple sentences in the other language, or multiple sentences can be condensed

[1] https://gadebate.un.org/en

into a single sentence in another. Therefore, a simple probabilistic model based on sentence length would fail across languages with different scripts and language families. To solve this problem, we use a semantic similarity approach that aligns the sentences based on their meaning and content, rather than their form or position. We do this by using word vectors to represent the meaning of each sentence as a numerical vector. Then, we calculate the euclidean distance between the vectors of each language pair. The sentence pair with the smallest distance is the correct match.

For word vectorisation, we used Language-Agnostic Sentence Representations (LASER)[2], an open-source NLP toolkit developed by Facebook and trained on the Tatoeba corpus [3]. LASER performs sequence-to-sequence processing with an encoder-decoder approach. The encoder network, which is used to generate the embeddings we need, is a five layered bi-directional Long-Short-Term Memory (BiLSTM) network whose input is a string and output is a fixed-size vector in a 1024 dimensional space. Crucially, this space is shared by all languages, meaning that sentences with similar meaning in two different languages would be mapped to very near points in the space, regardless of how different the languages are.

The vectors are normalised and stored in a matrix, where each row represents a sentence and each column represents a language. We calculate the Euclidean distance between each sentence in one language and a window of 25 sentences in another language for each language pair. We select the pair of sentences with the smallest distance as the match. The window size is implemented to decrease time complexity as well as improve accuracy by not considering sentences too far away to have been the intended translation. This is done to prevent long, vague sentences that may be close to several other sentences from being matched numerous times, while also allowing for genuine cases where a sentence in one language has legitimately been represented by multiple sentences in the other. We also perform anchoring, where we identify special entities such as dates and numbers, and include only sentences in the target language that that contain the same terms to be considered for matching.

[2] https://github.com/facebookresearch/LASER

[3] https://tatoeba.org/en/

Table 1: Statistics of Pairwise Aligned Sentences

	Sentences	Source Tokens	Target Tokens
En–ES	322,379	8,051,597	8,782,297
En–FR	325,968	8,145,802	8,885,067
En–ZH	300,281	6,849,901	6,503,222
En–RU	316,031	7,938,417	6,849,994

5 Validation

After matching, we performed a simple validation by training a linear regression model to predict sentence length for a translation in a target language, based on the length of the original sentence in English. We use this model to estimate the likelihood that a target language sentence is the correct translation for an English sentence. If the other language sentence length is either less than 50% or more than 150% of how much it is predicted to be by the model, it is discarded. We keep the remaining matches and add them to the corpus.

The statistics for all validated language pairs are presented in Table 1, which shows the number of sentences for each language pair, along with the number of tokens for each of the language pairs.

6 Evaluation

We evaluated the quality of our final validated dataset using crowd-sourced human annotations. To obtain reliable and consistent evaluations, we used Scale AI[4], an online platform whose purpose is to generate labelled datasets for training AI models. Scale AI allows for the labelling of data such as images, videos, texts and 3D models.

We uploaded our parallel documents to Scale AI and requested the annotators to mark the sentences that are translations of each other in each language pair. We also provided them with clear instructions and examples of how to perform the task. We received the annotations from Scale AI in a JSON format, which we converted into a tab-separated format for further analysis. Scale also selects a "training set" of 20 sentence pairs, which it chooses from the corpus, for its crowd-sourced users, and discards results from users that perform below a threshold of 70% on the training set.

We designed the task to present annotators with two sentences, the "original" sentence in English and the "target" sentence in one of the target languages: French, Russian, Spanish or Chinese. The

[4] https://scale.com/

annotators were asked to read the sentences carefully and decide whether or not the two sentences are a match, a partial match, no match, or if they were unsure. In the instructions, the annotators were given detailed guidelines about what constitutes a match. If two sentences are direct translations of each other, or if all the information in the target sentence is present in the original sentence, they are considered to be a match. Furthermore, if the target sentence conveys the full meaning of the original sentence, annotators are to consider them a match. Partial matches occur when some information in the target sentence is not present in the original sentence, or vice versa. If the sentences are neither a full match nor a partial match, then annotators were to choose "no match". We also included an "unsure" option, and discarded any responses that included it. Figures 1 and shows an example of the instructions for the English–French evaluation, the way they appear next to each sentence pair in the task. In addition to these instructions, further workflow directions were made available.

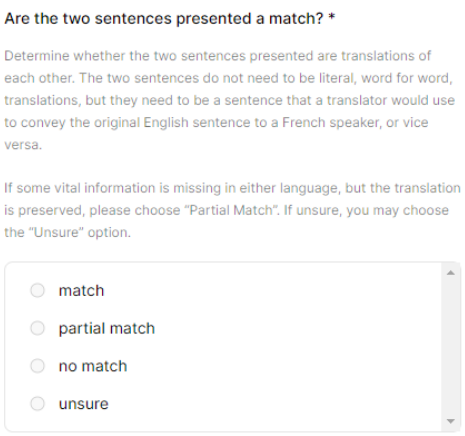

Figure 1: Instructions as they are presented to annotators alongside each sentence pair.

The task required 3 reviews per sentence, meaning three different annotators had to agree on a label for it to be accepted. Annotators were required to have a basic proficiency in the source language and native proficiency in the target language. However, due to crowd-sourcing, there was no way to verify their actual proficiency. The agreement between reviewers was pretty high, with Cohen's Kappa at 0.87 across all four language pairs. Furthermore, the evaluators found that over 80% of the presented sentences were a match, and less 5% were completely unaligned.

The number of evaluated sentences varied across languages, as it depended on the number of available annotators that Scale was able to train for each task. As a result, while we only had 1000 sentences evaluated for English–French, we managed to evaluate upwards of 6000 sentences for English–Chinese. Table 2 shows the number of sentences we aligned, and the percentage of this total that we managed to evaluate.

Table 2: Percentage of Sentence Pairs Evaluated Across Languages

	Evaluation Set	Percentage of Total
En–Fr	1000	0.3%
En–Es	7000	2.3%
En–Zh	6000	2%
En–Ru	9000	3%

The evaluators found that on average, over 85% of aligned sentences were a complete match, with around 6% of sentences being completely misaligned. The English–Spanish language pair had the highest percentage of correctly aligned sentences, at 91.4% of sentences being a total match. Conversely, the English–Russian language pair showed the highest number of misalignment, with only 78.5% of sentences matching. Table 3 shows the percentage of correctly aligned sentences by language pair.

Table 3: Scale AI Evaluation of Alignment

	Complete Match	Partial Match	No Match
En–Fr	86.4%	8.5%	5.1%
En–Es	91.4%	2.6%	6%
En–Zh	86.3%	9%	4.7%
En–Ru	78.5%	13%	8.5%

7 Limitations

We acknowledge some limitations of our dataset that we aim to overcome in future iterations. One of the main limitations is that we could not include Arabic as one of the languages in our corpus, due to technical difficulties in converting the Arabic documents into a suitable format for alignment. This means that our dataset does not cover all six official languages of the United Nations, and thus misses an important and widely spoken language in the world. We hope to solve this problem by finding a more reliable way to process the Arabic documents and align them with the other languages. Another limitation of our dataset is

that we relied on crowd-sourcing for evaluating the quality of our alignment. While crowd-sourcing is a convenient and cost-effective way to obtain human judgments, it also comes with some drawbacks, such as inconsistency and bias among the annotators. We tried to mitigate this issue by auditing the results and filtering out the outliers, but we could only review a small fraction of the evaluations. Therefore, our evaluation may not reflect the true quality of our dataset, and may be influenced by the subjective opinions of the annotators. We plan to address this issue by conducting a more rigorous and systematic evaluation of our dataset, using multiple sources of human feedback and objective metrics.

8 Conclusion

In this paper, we introduced a novel parallel corpus that consists of texts from the plenary sessions of the United Nations General Assembly. Our corpus covers five languages: English, French, Spanish, Russian, and Chinese. We described the process of extracting and preprocessing the sentences from the original documents, and aligning them based on semantic similarity using a state-of-the-art cross-lingual sentence encoder. We evaluated the quality of our dataset using two methods: a simple validation that uses a regression model to predict sentence length based on the source language and the target language, and a crowd-source human evaluation that measures the accuracy and precision of our alignment.

The resulting aligned dataset has a high degree of accuracy across languages, and can be used for various natural language processing tasks, such as machine translation, cross-lingual information retrieval, and multilingual text summarisation.

Our work contributes to the field of multilingual natural language processing by providing a large-scale and high-quality parallel corpus that covers multiple languages in the field of political discourse and debate. We believe that our corpus can facilitate the development and evaluation of cross-lingual models and applications. In the future, we plan to solve the problem of Arabic-language documents that prevented us from completing our dataset for all six official languages of the United Nations. We also intend to extend our corpus to include more languages and more sources of multilingual texts. The current version of our dataset is available for download[5].

9 Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101057131, Climate Action To Advance HeaLthY Societies in Europe (CATALYSE).

References

Artetxe, Mikel and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July. Association for Computational Linguistics.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Chen, Yu and Andreas Eisele. 2012. MultiUN v2: UN documents with multilingual alignments. In *International Conference on Language Resources and Evaluation*.

Eisele, Andreas and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Guo, Mandy, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium, October. Association for Computational Linguistics.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29, 03.

Schwenk, Holger. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia, July. Association for Computational Linguistics.

Ziemski, Michal, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *International Conference on Language Resources and Evaluation*.

[5] https://github.com/KrishnaM313/UN_multilingual_corpora