ZPJa: Bi-text retrieval for machine translation

Student Name
xtikho00@fit.vut.cz

Supervisor Name
kesiraju@fit.vut.cz

Abstract

To train state-of-the-art machine translation models, high-quality parallel corpora are needed. In most cases, parallel data is being mined from multilingual websites. The data is automatically aligned for several language pairs. However, the quality of the resulting aligned corpus can be disappointing: paired sentences may not be aligned, there could be noisy elements, such as markup tags, HTML entities, escape sequences. The objective of this research was to audit the web-crawled parallel corpus and to implement filters in order to clean it and compare the performance of MT models trained on the same corpus before and after cleaning.

1 Introduction

The large-scale bilingual corpora are abundant nowadays, partially because of availability of web-crawled datasets, however, quality of those may be disappointing: there could be a lot of unrelated symbols in text, markup tags and misaligned sentences. The purpose of this research is to audit the quality of the given web-crawled dataset (CCMatrix (Schwenk et al., 2019)) and implement the filters to clean and analyze the performance of models trained on both cleaned and uncleaned datasets.

2 Task Definition

Motivated by the low quality of parallel corpora obtained by web-crawling (Kreutzer et al., 2022), the main goal of this specific variant of assignment was to implement the filters to clean these kinds of datasets and, first, compare the average sentence-similarity between of cleaned and uncleaned data, and second, evaluate the performance of models trained both on cleaned and uncleaned data.

3 Method

This chapter describes findings about the chosen dataset and details about filters implementation.

3.1 Data quality examination

As a language pair I chose English-Basque, this choice is motivated by the fact that Basque is not a widespread and can be even considered low-resource. For the sake of simplicity of conducting experiments I wasn't using the entire CCMatrix dataset, instead I extracted first 130 thousand lines and used it throughout this research.

After analysing the first 3000 lines of the CCMatrix subset I discovered the following problems:

Misaligned sentences The most common issue was the sentence misalignment, the situation when pair of two sentences are not mutual translations.

Duplicate sentences Although the full duplicates (when the entire pair is duplicates) may only cause longer model training, partial duplicates (when, for example, only source side is duplicated) may harm the resulting performance of the model.

Noisy sentences Noise in sentences is represented by many entities, let's list few of them:

- enumeration or dashes/hyphens at the start of sentences.
- sentences (or parts of them) inside parentheses.
- obscene content.
- unpaired symbols (quotation marks, parentheses).
- junk characters (emojis, bad encodings).
- timestamps.

Partially misaligned sentences Some sentences in pairs, having similar semantic meaning to their aligned sentences, may have some parts of them that do not correspond to those in sentences in other language.

Misaligned sentence	**Source**	Sinesmen abrahamikoak[aldatu | aldatu iturburu kodea]
	Target	those who have the faith of Abraham,
Duplicate sentences	**Source**	Uste al duzu Jainkoaren lege moralak zure onerako direla?
	Target	Do you believe that God's judgment of you is for your benefit?
	Source	Uste al duzu Jainkoaren lege moralak zure onerako direla?
	Target	Do you believe that God's moral will is good?
Noisy sentences	**Source**	Zorionekoak ikusi gabe sinesten dutenak|[[Joanen Ebanjelioa]], 20:24,29}}
	Target	Blessed are they who believe without seeing.
	Source	- Jainkoaren izena : YAHVE.
	Target	God's s name is (YHWH) Yahweh.
Part. m.-aligned sentences	**Source**	Zer egin zien Faraonek [Jakoben semeai]?
	Target	And what does the Pharaoh have?

Table 1: Examples of encountered problems.

3.2 Filtering methods

This subsection describes the filtering methods through which I tried to address the aforementioned problems.

3.2.1 Main methods

Deduplication According to observations, duplicate sentences appear near each other, so, to address the issue of duplicate documents, cleaner, in advance, before applying filters on the dataset, iterates through it by chunks (each consists of approximately 1000 lines), and, after the removal of any delimiters (spaces, punctuation), checks the similarity ratio of extracted sentences. If the similarity ratio is higher than 0.9, then it removes the sentence pair from dataset:

$$F = \frac{(2 * \textit{Number of matched characters})}{\textit{Total length of both sequences}} > 0.9$$

Every sentence that was filtered is added to the set of encountered duplicates, so all other entries with these sentences will also be deleted.

Deduplication reduced the size of the dataset by 7% (from 130000 to 120752 entries).

Regex filters Regex filters are useful when we know in advance that some repeating patterns or unnormalized entities may be encountered in dataset. Such entities as markup (HTML, latex, XML, ...) tags are fixed either by replacements with tags' contents (e.g. \textbf tag in latex) or by its complete removal (XML/HTML tags). There are also regex patterns that pursue the text normalization by replacing the symbols that are rare, list of few replacements that were used:

- different types of quotation marks (guillemets, quotes and even doubled "<" or ">" signs) will be unified to standard quotes (").
- special ellipsis signs are replaced by three periods.
- different types of signs for hyphens/dashes are replaced by standard ones.
- double apostrophes are either converted to single ones (when are used in English contractions) or to quotes.

Getting rid of misaligned sentences In order to address the issue of misaligned sentences I utilised the LASER (Feng et al., 2022) and LaBSE (Feng et al., 2020) models (Python interfaces provided by *laserembeddings*[1] and *huggingface-transformers*[2] respectively) in order to calculate cosine-similarity

[1] https://github.com/yannvgn/laserembeddings
[2] https://github.com/huggingface/transformers

of sentences in pairs and sort out the misaligned sentences. Cosine-similarity between two sentences is given by:

$$sim(x_1, x_2) = \frac{x_1 \cdot x_2}{||x_1|| \cdot ||x_2||}$$

Also, in order to address the issue sentences which are highly polluted with unrelated sets of expressions or symbols (e.g. random sequence of punctuation symbols, serial biblical in-text citations) I used the number of tokens as a evaluation parameter for each sentence.

The final score that is assigned to the sentence pair is calculated using the following:

$$S = (sim_{laser}(x_1, x_2) + 2 \cdot sim_{labse}(x_1, x_2)) \cdot penalty(x_1, x_2) \quad (1)$$

where

$$penalty(x_1, x_2) = \begin{cases} 0.8, & \text{if } \frac{N(x_2)}{N(x_1)} > 2 \text{ or } \frac{N(x_1)}{N(x_2)} > 2 \\ 1, & \text{otherwise} \end{cases}$$

where N is the number of tokens in given sentence.

Doubling the similarity provided by LaBSE model is motivated by the fact that CCMatrix corpus was already filtered using LASER model and it outputs comparatively high scores for almost all sentence pairs.

The removal of misaligned sentences cut the size of training dataset even further, down to 102000 pair sentences.

Profanity filter To solve the issue of profane content encountered in dataset I used the profanity filter for English language from *profanity-check* library[3]. Whenever it encounters sentence that contains obscene content, it removes the entire pair from the dataset.

3.2.2 Other methods

This subsection devoted to methods that have been tested and turned out to be either not very useful or even harmful for given dataset.

Language detection Language detection is an essential part for dataset processing, especially for low-resource languages like Basque. It was expected that due to great influence of more widespread bordering languages (Spanish, French) Basque side of sentence pairs would be polluted with expressions in these languages.

However, these expectations were not met. The languages of all sentence pairs were recognized by *fasttext* classifier (Joulin et al., 2016) as Basque on one side and English on the other.

Although the language classifier was useless for this particular dataset, it is still necessary to check the language of sentences, so I decided to leave language classifier in the pipeline.

Spelling correction The spelling correction is based on *TextBlob* library[4] and was used for both languages. The English spelling correction is provided out of the box by the TextBlob. In order to assemble the vocabulary for Basque language I utilised the method described by (Norvig, 2007), in which to create the vocabulary composed of words and count of their occurrences we have to parse the provided text. For that purpose I used these corpora: bible-uedin (Christodoulopoulos and Steedman, 2014), XLEnt (El-Kishky et al., 2021) and flores-200 (NLLB Team, 2022).

The spelling correction itself is provided by TextBlob and is performed when the word that is not in language dictionary is within edit distance 2 from the word that is in dictionary (it will be replaced with word with the highest count assigned to it).

This method was not effective, depending on dictionary used, it was fixing from 0 to 1500 words in the entire dataset, which is negligible. This can simply be explained by the fact that both English and Basque sentences don't have any orthographic errors.

Fixing parallel sentences using aligner To investigate the issue of partially misaligned sentences even further I tried to use *awesome-align* toolkit (Dou and Neubig, 2021) to extract the source-to-target word alignments.

Awesome-align uses word representations from the 8th layer of mBERT (Devlin et al., 2018) in order to extract alignments.

In order to remove the tokens that do not have their reflection the opposite sentence, I utilised the

[3] https://github.com/vzhou842/profanity-check

[4] https://github.com/sloria/TextBlob

probability matrix that awesome-align aligner provides. If the probability of some word alignment is very low and there is no other word that has alignment with it, it will be removed.

The testing aligner's capabilities showed that, although in most cases it provides right alignment pairs, in others it may misalign the words, which will only worsen the quality of particular sentences. Because of that it was decided to not include the aligner in final cleaning pipeline.

4 Experimental Setup

4.1 Dataset

As were mentioned before, I used the first 130000 sentences pairs from CCMatrix corpus as uncleaned training data and 102000 processed sentence pairs as cleaned training data. For the validation set I have chosen flores-200 dataset (2009 sentence pairs) because of high data quality. As the test set I used Tatoeba corpus (Artetxe and Schwenk, 2019) (2066 sentence pairs) because it is known for its relative syntactic and semantic simplicity. All corpora were downloaded from OPUS (Tiedemann, 2012).

4.2 Training

Training was conducted with supervised English-to-Basque MT objective without pre-training. I utilised *XLM* toolkit[5] to train the model with following parameters:

- embeddings dimension: 128.
- number of layers: 4.
- number of attention heads: 8.
- epoch size: 100000.
- batch size: 32.
- learning rate: 0.00001.
- validation metric: BLEU.
- stopping criteria: 50 failed attempts to improve the score.
- number of BPE codes: 50000.

The way the data were obtained and the models were trained is described in `experiments/experiments.ipynb`.

	Sentence-similarity		Model performance: valid/test	
	Laser	Labse	BLEU	chrF
Original data	0.8322	0.8164	3.06/14.45	32.26/42.92
Cleaned data	0.8279	0.8256	3.10/14.56	33.15/43.76

Table 2: Results of dataset cleaning and training on cleaned/uncleaned data

5 Results and Analysis

As you can see in the Table 2, after data cleaning sentence-similarities "shifted" more to the LaBSE side, it can be explained by the fact that scoring function 1, that was used to sort sentence pairs by score and to remove pairs with low scores, doubles the similarity obtained from LaBSE model, thus increasing its importance.

Regarding the validation scores, the difference between the scores of two models is negligibly small, it may even be caused by parameter initialisation with a different seed.

The inability to improve the model performance by dataset cleaning may be caused either by my misunderstanding the main problems CCMatrix's subset was stuffed with or by the fact that the dataset is too polluted by too many patterns for filters to solve this problem.

Although the resulting performance of the model has not been improved, the reduction of dataset significantly shortened the training time (from 275 epochs to 230).

6 Conclusion

In this research I analysed the polluted CCMatrix subset and discovered its main issues and implemented filters to address these problems and to create the cleaned version of this dataset. Then I trained models on both versions and compared their performance on validation and test data.

Although I find the products of this research rather unsatisfactory: the main issues – dataset's pollution with noisy elements and misalignment of sentences were solved only partially (by removal of sentences with low heuristic quality score), it still provided some improvement in model's training time.

[5] https://github.com/facebookresearch/XLM

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Christos Christodoulopoulos and Mark Steedman. 2014. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49:1–21.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora.

Ahmed El-Kishky, Adi Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining cross-lingual entities with lexical-semantic-phonetic word alignment. In *Preprint*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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 Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Peter Norvig. 2007. How to write a spelling corrector. *De: http://norvig.com/spell-correct.html*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web.

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).