How good is NLLB-200 for low-resource languages? A study on Genoese

Davide Buscaldi[1], Paolo Rosso[2]

[1]LIPN, CNRS UMR 7030, Université Sorbonne Paris Nord, 99 av. Jean-Baptiste Clément, 93430, Villetaneuse, France
[2]PRHLT, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022, Valencia, Spain

Abstract

English. In this paper we analyze the performance of the NLLB-200 models from Meta AI on a manually built parallel corpus of Ligurian (specifically, the Genoese variant), consisting in 283 sentences and their respective Italian translation. Our experiments highlight some issues with NLLB-200, especially regarding local knowledge, deriving from some choices done for the training process.

Italiano. In quest'articolo analizziamo la performance del modello NLLB-200 di Meta AI su un corpus parallelo, costruito manualmente, di 283 frasi in genovese e la loro rispettiva traduzione in italiano. Mostriamo i punti deboli di NLLB-200, in particolare il trattamento dei toponimi ed altri termini in relazione con un contesto locale ligure, evidenziando alcuni problemi derivati dalle scelte fatte nel training di questo modello.

Keywords

NLLB, Machine Translation, Genoese, Endangered Languages

1. Introduction

NLLB (No Languages Left Behind) [1] is a collection of language models created by Meta AI to fill the void left in Machine Translation (MT) for some low- and very low-resource languages. NLLB-200 is the latest model and is able to provide MT for 200 languages, including some that had never been considered before. One of these languages is Ligurian, an endangered language that is spoken mainly in the Liguria region in Italy, Monaco (where it is called Monegasque), and some small islands in the Mediterranean Sea (Carloforte and Calasetta in Sardinia).

The content in Ligurian on the web is very scarce. The main source is Wikipedia, which has only $11,172$[1] articles in Ligurian, many of them being a bit more than drafts. In comparison, the number of articles in Welsh, which has an estimated equivalent number of native speakers ($500,000$) is twenty times as much as Ligurian.[2] This difference is easily explained by the fact that Welsh is an official language, supported by the local government while Ligurian is mostly orally spoken.

The rarity of content in Ligurian is not the only problem that may affect MT tools and methods. In particular, the syntax of Ligurian has not been completely standardised: many variants of the same word may exist, even when they are pronounced in the same way, due to various reasons. First of all, the local variants of Ligurian, but also because the language has been passed down from a generation to another one mostly in an oral way. For instance, in Monegasque, the word "white" is written as *giancu* while in Genoese (the predominant variant) it is written as *gianco*[3]. This problem has been well exposed in the work of [2], which also cite the lack of regulatory bodies as one of the sources of variations. In their study, they propose a corpus of normalized and unnormalized texts in Ligurian to train a neural model for the normalization of Ligurian texts. The example in Figure 1 allows to appreciate the high variability of Ligurian spelling.

Unna	rondaniña	affammâ	a s' é pösâ in	sciô	teito de coppi
Ûnn-a	rōndaninn-a	affammâ	a s' é pösâ in	sciö	teito de cōppi
Ûnn-a	rondaninn-a	affammâ	a s' è pösâ in	sciö	teito de coppi
Ûnn-a	rondaninn-a	affamâ	a s' è pösâ in	sce-o	teito de coppi
Ûña	rundaniña	affammä'	a s' é pösä	in sce o	téyto de cuppi

Figure 1: Examples of 4 variants of Ligurian from [2], with the reference standardised spelling on top. In our work, we did not standardise the texts but used them "as they are".

Given this premise, it is important to evaluate whether the NLLB-200 model is able to deal with these problems. For this reason, we conducted an evaluation by composing a test dataset in Genoese that was not used for NLLB training. The copyright-free subset of this

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

buscaldi@lipn.fr (D. Buscaldi); prosso@dsic.upv.es (P. Rosso)
https://lipn.fr/~buscaldi (D. Buscaldi);
https://www.prhlt.upv.es/paolo-rosso/ (P. Rosso)
0000-0003-1112-3789 (D. Buscaldi); 0000-0002-8922-1242 (P. Rosso)

[1]as of 20/07/2023

[2]https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_speakers_per_article

[3]https://fr.wikipedia.org/wiki/Monegasque

dataset is available at the following address: https://github.com/dbuscaldi/zeneize.

In the following section we describe the NLLB-200 model and how the training has been carried out. In Section 3 we show the experiments carried out on our data set and discuss the results. Finally, in Section 4 we draw some conclusions of this analysis and propose some ideas to improve the model on the basis of this experimentation.

2. NLLB-200

To be able to interpret the results we must first take a better look at NLLB-200: the dataset on which it was trained, and the characteristics of NLLB-200 and the models used in our experiments.

2.1. Training Data

The training data for Ligurian were created by composing a set of 6,193 professionally-translated sentences in the Wikipedia domain, named *NLLB-Seed*[4] [1]. Data for NLLB-Seed was sampled from Wikimedia's "List of articles every Wikipedia should have", a collection of 10,000 Wikidata IDs corresponding to notable topics in different fields of knowledge and human activity. These are split into 11 categories such as People, History, Philosophy and Religion, Geography. NLLB developers note that half the data for Ligurian were first translated from English to Italian, then translated from Italian to Ligurian while the other half was translated directly from English. It can be noted that this process is covering English language domain knowledge rather than knowledge related to the local language. we will come back to this aspect later during our evaluation.

2.2. Models

The NLLB-200 model is an encoder-decoder model that makes the most out of the LASER-3 embeddings [3]. LASER-3 are multilingual embeddings that focus on training multiple language *family-specific* representations. This means that embeddings trained for Italian will still have a degree of similarity to other embeddings in the same family of languages (i.e., Romance languages). The final translation models come in various configurations, distilled and non-distilled. The non-distilled models have 3.3B and 1.3B parameters, while the distilled ones have 1.3B and 600M parameters. Distillation for these models is based on online word-level distillation [4], which means that the student model is trained on the training data but with an additional objective: to minimize the cross-entropy with respect to the word-level distribution of the teacher model.

3. Experiments and Results

We collected a dataset of texts in Genoese from three different sources. First of all, 95 lines from two of the most famous songs of Fabrizio De André, "Crêuza de mä" (small path to the sea) and "'A çimma" (the cima is a typical Genoese dish) with their respective translation in Italian found on the official Fondazione De André page[5] (retrieved on 2022-08-13), and the popular song "Trilli trilli". Then, 188 sentences from "Zêna e contórni", a translation in Genoese from Charles Dickens section on Genoa, Italy, from his "Pictures of Italy" work. We used the wikisource text[6] and an Italian translation obtained from the English one with Deepl[7].

We applied the NLLB-200 models on this dataset, obtaining the results in Tables 1 and 2. We calculated the results using SacreBLEU [5], in particular the measures spBLEU [6] with flores-200 tokenization (as in the NLLB paper), the character n-gram based measure chrF [7], and TER (Translation Error Rate) [8].

The time required to run the translation varied considerably from the dist-600M (about 20 minutes) to the 3.3B model (about 3 hours and 45 minutes). The intermediate size models, 1.3B parameters took about 45 minutes on average to process the whole dataset. All these values were obtained on a CPU 2,6 GHz Intel Core i7 (no GPU acceleration used) and 16GB RAM.

As it can be seen, the results are quite appalling even with the largest model, casting some doubts on the usability of the NLLB-200 model for Genoese. It can be observed that all models are having more problems with the translation from Italian to Genoese than in the opposite direction. As expected, in most cases, the larger the model, the better the results, although the improvements in the Italian-Genoese translation are lower than in Genoese-Italian. TER values higher than 100 indicate that the models are overgenerating, producing sequences that are longer than the reference ones. This is particularly evident in the songs subset, in the Italian to Genoese direction.

An inspection of the results in Genoese shows some interesting outputs of the models. Toponyms are often translated incorrectly. For instance, let's consider the Ligurian capital, Genova, which is mentioned 9 times in our dataset. In the Italian-Genoese direction, only NLLB 3.3B translates it correctly in 3 out of 9 cases. NLLB 1.3B once translates "Genova" into "Genoa" instead of "Zena" and it always keeps "Genova" elsewhere. NLLB

[4] https://github.com/facebookresearch/fairseq/tree/nllb

[5] http://www.fabriziodeandre.it/

[6] https://lij.wikisource.org/wiki/Zêna_e_contórni

[7] https://www.deepl.com/translator

Table 1
Genoese to Italian results

	De André + Trilli			Dickens			Full Dataset		
Model	**spBLEU**	**chrF**	**TER**	**spBLEU**	**chrF**	**TER**	**spBLEU**	**chrF**	**TER**
NLLB dist-600M	11.2	35.1	74.4	14.2	40.5	73.6	15.2	39.9	73.7
NLLB dist-1.3B	16.3	**38.3**	67.7	18.6	**44.9**	68.0	**20.2**	**44.2**	**67.9**
NLLB 1.3B	14.0	37.9	**66.2**	18.9	44.0	68.9	18.4	43.3	68.5
NLLB 3.3B	**14.2**	35.9	71.4	**20.9**	44.8	**67.4**	**20.2**	43.8	**67.9**

Table 2
Italian to Genoese results

	De André + Trilli			Dickens			Full Dataset		
Model	**spBLEU**	**chrF**	**TER**	**spBLEU**	**chrF**	**TER**	**spBLEU**	**chrF**	**TER**
NLLB dist-600M	4.2	24.5	144.4	2.4	25.3	96.3	4.4	25.2	102.4
NLLB dist-1.3B	4.1	26.4	108.6	3.8	**26.9**	**93.4**	5.5	**26.8**	95.3
NLLB 1.3B	7.8	24.9	**98.8**	4.8	25.9	93.5	5.3	25.8	**94.2**
NLLB 3.3B	**9.0**	**27.7**	102.2	**5.3**	26.6	95.8	**5.9**	26.7	96.6

dist-600M is never able to translate correctly “Genova" into “Zena", it always translates it the same as in Italian (Genova). Finally, NLLB dist-1.3B correctly translates it 6 times. The problem seems also to affect other proper nouns, such as Saint Peter (San Pietro in Italian) which is correctly translagted in Genoese as “San Pê". In fact, is translated as “San Peixe", which is Portuguese for fish, in both distilled models. The 3.3B model translates it as “San Pêo" and the 1.3B one translates it as “San Peçio". Both these translations make no sense in Genoese.

In the Genoese-Italian direction, only NLLB dist-1.3B translates “Zêna" correctly in 3 out of 9 cases. Both 1.3B models translate in one case it as “Ginevra" (Geneva, in Switzerland). Other spelling errors show “Giena" by the dist-600M model and “Gênes" (in French instead of Italian) by the 3.3B model. Looking into the tokenizer we observed that “Genova" is not in the dictionary and is tokenized as *Gen-ova*, and “Zêna" is tokenized as *Z-êna*. On the other hand, “San Pê" is correctly translated by all models. Due to the output occurring sometimes in different languages than the target ones, we suspect that the previous errors may result from the LASER-3 embeddings which are language-family based.

Both distilled models fall into repetitions. For instance, in the Dickens text, NLLB dist-600M translates “Ma, per il momento, gironzolo qui intorno, in tutti i buchi e gli angoli del quartiere, in un perpetuo stato di forzata sorpresa" (“But, as yet, I stroll about here, in all the holes and corners of the neighbourhood, in a perpetual state of forlorn surprise") into “*Ma, pe-o momento, o l'é in sciâ çitæ, in tutti i buchi e in tutti i cantoni do quartiere, in un stato de sorpresâ forçâ pe pe pe pe pe pe pe pe pe...*" (“But, for now, he is on top of the city, in all the holes and corners of the neighbourhood, in a state of forced surprise for for for for..."). The larger model (dist-1.3B) is not immune to this behaviour although it happens only 2 times instead of 9. The non-distilled models don't present this problem. The fact that the models fall into this kind of repetition could be due to the lack of sufficient training data for the word-based online distillation process. Therefore the probability distribution for the tokens is skewed towards some frequent words (“pe" - *for*, “ti" - *you*, “ben" - *well*). We observed that the minimum frequency in the NLLB-seed dataset of words that are repeated is 39 (for the word “sciâ": probably as part of “in sciâ" - on top of).

4. Conclusions

From our preliminary analysis, carried out on a dataset specific to the Genoese culture, we can affirm that currently NLLB-200 is not good enough to deal with Genoese texts or to translate text into Genoese. In particular, we found out that local toponyms are difficult to translate: how good is an MT tool that is not able to correctly translate the name of the largest city where the language is spoken or the name of the language itself? Given the information provided regarding NLLB-200 models, we can identify two main elements explaining this behaviour. The first one is the training data: they do not cover local information, but general English Wikipedia articles, so they lack to provide the context in which Genoese is usually spoken. The second one is the tokenization process and the LASER-3 embeddings: given the high spelling variability of the Ligurian language, we suspect that the tokenization process may not be precise and that it may map some tokens into a position in the embedding space that does not correspond to their actual “meaning", maybe also because of a sort of interference from

other Western Romance languages that are very close to Ligurian.

However, NLLB-200 is a big step forward making endangered languages such as Ligurian and its variants available to everyone. From our point of view, we think that NLLB-200 could be improved in various ways, for instance fine-tuning the model on more “local" datasets; and possibly including knowledge regarding Out-Of-Vocabolary words that are often named entities, for instance with the methods proposed by [9], or integrating dictionaries to deal with named entities.

Acknowledgments

This work is supported/ partially supported by a public grant overseen by the IdEx Université Paris Cité (ANR-18-IDEX-0001) as part of the Labex Empirical Foundations of Linguistics - EFL.

References

[1] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022).

[2] S. Lusito, E. Ferrante, J. Maillard, Text normalization for low-resource languages: the case of ligurian, in: Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages, 2023, pp. 98–103.

[3] K. Heffernan, O. Çelebi, H. Schwenk, Bitext mining using distilled sentence representations for low-resource languages, arXiv preprint arXiv:2205.12654 (2022).

[4] Y. Kim, A. M. Rush, Sequence-level knowledge distillation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1317–1327. URL: https://aclanthology.org/D16-1139. doi:10.18653/v1/D16-1139.

[5] M. Post, A call for clarity in reporting bleu scores, arXiv preprint arXiv:1804.08771 (2018).

[6] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2018, pp. 66–71.

[7] M. Popović, chrf: character n-gram f-score for automatic mt evaluation, in: Proceedings of the tenth workshop on statistical machine translation, 2015, pp. 392–395.

[8] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, 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, 2006, pp. 223–231.

[9] J. Waldendorf, A. Birch, B. Hadow, A. V. Micele Barone, Improving translation of out of vocabulary words using bilingual lexicon induction in low-resource machine translation, in: Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), Association for Machine Translation in the Americas, Orlando, USA, 2022, pp. 144–156. URL: https://aclanthology.org/2022.amta-research.11.