A high-quality Seed dataset for Italian machine translation

Edoardo Ferrante
Council for Ligurian Linguistic Heritage
info@conseggio-ligure.org

Abstract

This paper describes the submission of a high-quality translation of the OLDI Seed dataset into Italian for the WMT 2024 Open Language Data Initiative shared task.

The base of this submission is a previous version of an Italian OLDI Seed dataset released by Haberland et al. (2024) via machine translation and partial post-editing. This data was subsequently reviewed in its entirety by two native speakers of Italian, who carried out extensive post-editing with particular attention to the idiomatic translation of named entities.

1 Language overview

This paper presents an Italian version of the OLDI Seed dataset (Maillard et al., 2023; NLLB Team et al., 2024).

Italian is a Romance language, recognised as an official language of the Italian Republic, the Republic of San Marino and the Canton of Ticino in Switzerland (Maiden, 2014). Modern Italian fundamentally represents cultured Florentine, as first attested by 14th century authors (Dante, Petrarch and Boccaccio) and later scholars (Coletti, 2022). Although it is a variety of Tuscan, standard Italian is purged of the more typical features of Tuscan, at a phonetic level represented above all by the so-called *gorgia* (i.e. the fricative pronunciation of certain occlusive consonants in intervocalic position) (Marotta, 2008).

The presence of a curated Italian version in the Seed dataset is of great importance for the regional languages of the Italian peninsula, six of which are already represented in the same dataset.[1] The creation of an Italian version enables the training of machine translation models for these languages to and from Italian, a direction which is more culturally relevant than English-centric MT, as the vast majority of speakers of such languages (or prospect learners) are also native Italian speakers (Haberland et al., 2024; Ramponi, 2024).

2 Data creation

The original source of the data was an initial Italian version of the Seed dataset released by Haberland et al. (2024). The authors created it by machine translating the original English version with an OpusMT bilingual English-Italian model (Tiedemann and Thottingal, 2020), combined with partial post-editing. Through personal correspondence with the authors we learned that their post-editing, which only affected a small percentage of the overall data, involved two steps:

1. A check of the length ratios of Italian and English sentences, followed by manual checking and post-editing of sentence pairs with outlier length ratios.

2. A spellchecker run using Hunspell (Ooms et al., 2017), followed by manual checking and post-editing of sentence pairs where spelling mistakes were found.

The submission described in this paper constitutes a further refinement of the dataset of Haberland et al. (2024), in order to bring it to a level that could be seen as comparable to that of translations produced by highly proficient bilingual individuals.

This project involved the participation of two annotators, henceforth A1 and A2, both native speakers of Italian with a university level of education. The refinement process followed these steps:

1. A manual, sequential review of the entire dataset by A1, followed by post-editing where necessary.

2. Following Haberland et al. (2024), a targeted review of sentence pairs with outlier length ratios, followed by post-editing where necessary.

[1] These are Friulian, Ligurian (Genoese), Lombard, Sicilian, Sardinian, Venetian (Maillard et al., 2023; NLLB Team et al., 2024).

Proceedings of the Ninth Conference on Machine Translation, pages 567–569
November 15-16, 2024 ©2024 Association for Computational Linguistics

3. A targeted review of sentences involving specific subsets of the corpus which were found to have a high incidence of mistranslated strings: date and time expressions, sentences about mathematics and sentences about the history of cinema.

4. A final targeted review of sentence pairs which were found to be of low-quality using a series of Quality Estimations approaches using LLMs, as described in Zhao et al. (2024).

Apart from the first item above, which was carried out by annotator A1 alone, the workload for all subsequent tasks was split equally between both annotators.

3 Experimental validation

In order to experimentally validate the quality of this Seed dataset, we replicate the baseline experiments of Haberland et al. (2024), by training an Italian-Ligurian machine translation model on a combination of the 6,193 paired Italian-Ligurian sentences from the Seed data and the same 1,520 paired Italian-Ligurian sentences from the Tatoeba project[2] used by the authors. The translation model is trained using Fairseq (Ott et al., 2019), with the exact same architecture and overall setup of Haberland et al. (2024).

Model	FLORES
NLLB-3.3B	13.9
Haberland et al. (2024)	14.5
Ours	**15.0**

Table 1: Italian-Ligurian translation performance measured with BLEU on FLORES `devtest`.

In Table 1 we compare the BLEU scores[3] obtained by three models on the FLORES (NLLB Team et al., 2024) `devtest` data. The first model, provided only for context, is the massively multilingual 3.3B version of NLLB (NLLB Team et al., 2024), which was trained on much larger amounts of data but without any direct Italian-Ligurian supervision. The second is the baseline model of Haberland et al. (2024). The final row reports the performance of the best of three training runs of our model, which is a re-training of Haberland et al.'s, the only difference being the use of the improved Seed data.

As can be observed in the results, our model achieves a performance of 15 BLEU points on the FLORES devtest set, 1.1 points higher compared to NLLB-3.3B (NLLB Team et al., 2024) and half a point higher compared to the baseline model of Haberland et al. (2024). The relatively small degree of improvement compared to the latter baseline can be attributed to the fact that, in general, machine translation for a high-resource language pair such as English-Italian is of high quality, so that manual post-editing (especially in a formal domain such as Wikipedia text) leads to only minor changes.

This result, although numerically marginal, confirms that our post-editing of the seed data for improved idiomaticity does not hurt the downstream performance of models trained on it but does, in fact, slightly improve it.

4 Data samples

We provide a selection of samples whose translations proved to be particularly hard for the OpusMT bilingual English-Italian model.

Acknowledgments

We thank Haberland et al. (2024) for the helpful discussions about post-editing and for helping us replicate their experiments.

References

Vittorio Coletti. 2022. *Storia dell'italiano letterario*. Einaudi.

Christopher R. Haberland, Jean Maillard, and Stefano Lusito. 2024. Italian-Ligurian machine translation in its cultural context. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 168–176, Torino, Italia. ELRA and ICCL.

M. Maiden. 2014. *A Linguistic History of Italian*. Longman Linguistics Library. Taylor & Francis.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Giovanna Marotta. 2008. Lenition in tuscan italian (gorgia toscana).

[2] `https://tatoeba.org/`

[3] SacreBLEU (Post, 2018) signature `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0`.

English (original)	Italian
He made a series of two-reel comedies, including One Week (1920), The Playhouse (1921), Cops (1922), and The Electric House (1922).	Realizzò una serie di commedie a due bobine, tra cui Una settimana (1920), Il teatro (1921), Poliziotti (1922) e La casa elettrica (1922).
The development of a regulatory framework concerning genetic engineering began in 1975, at Asilomar, California.	Lo sviluppo di un quadro normativo sull'ingegneria genetica è iniziato nel 1975, ad Asilomar, in California.
But the next major advance in the theory was made by Georg Cantor; in 1895 he published a book about his new set theory, introducing, among other things, transfinite numbers and formulating the continuum hypothesis.	Ma il successivo importante passo avanti nella teoria fu compiuto da Georg Cantor, che nel 1895 pubblicò un libro sulla sua nuova teoria degli insiemi, introducendo, tra l'altro, i numeri transfiniti e formulando l'ipotesi del continuo.
Aside from Steamboat Bill, Jr. (1928), Keaton's most enduring feature-length films include Our Hospitality (1923), The Navigator (1924), Sherlock Jr. (1924), Seven Chances (1925), The Cameraman (1928), and The General (1926).	Oltre a Io… e il ciclone (1928), tra i lungometraggi più duraturi di Keaton vi sono La legge dell'ospitalità (1923), Il navigatore (1924), Sherlock Jr. (1924), Le sette probabilità (1925), Il cameraman (1928) e Come vinsi la guerra (1926).
These chains of extensions make the natural numbers canonically embedded (identified) in the other number systems.	Queste catene di estensioni rendono i numeri naturali canonicamente immersi (identificati) negli altri sistemi numerici.

Table 2: Dataset samples.

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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Jeroen Ooms et al. 2017. Hunspell: High-performance stemmer, tokenizer, and spell checker.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Alan Ramponi. 2024. Language Varieties of Italy: Technology Challenges and Opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Haofei Zhao, Yilun Liu, Shimin Tao, Weibin Meng, Yimeng Chen, Xiang Geng, Chang Su, Min Zhang, and Hao Yang. 2024. From handcrafted features to llms: A brief survey for machine translation quality estimation.