eSTÓR: Curating Irish Datasets for Machine Translation

Abigail Walsh[1], Órla Ní Loinsigh[1], Jane Adkins[1], Ornait O'Connell[1], Mark Andrade[1], Teresa Clifford[1], Federico Gaspari[1], Jane Dunne[1], Brian Davis[1]
[1]ADAPT Centre, Dublin City University
firstname.lastname@adaptcentre.ie

Abstract

Minority languages such as Irish are massively under-resourced, particularly in terms of high-quality domain-relevant data, limiting the capabilities of machine translation (MT) engines, even those integrating large language models (LLMs). The eSTÓR project, described in this paper, focuses on the collection and curation of high-quality Irish text data for diverse domains.

1 Introduction

Despite the growing ubiquity of digital technologies, the Irish language lacks robust language technology that serves Irish speakers adequately in the digital sphere, with Irish language classified as in the "weak or no support" category of European languages (Lynn, 2022). This digital disconnect poses a significant threat to the vitality and sustainability of the Irish language resulting in the very real threat of *digital extinction* in the medium to long term.

The Digital Plan for Irish 2023-2027 (Ní Chasaide et al., 2022) is a detailed guide regarding areas in Irish language technology that require development. The eSTÓR (*Sonraí Teanga Óstáilte i gcomhair Ríomhphróiseála* "Hosted Language Data for Digital Processing") project is funded by the Irish government (Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media) from 2021 to 2025, to address the lack of high-quality data described in the Digital Plan, by providing a digital platform (`https://estor.ie/`) for **sharing bilingual and monolingual Irish text data**. In addition to data collection, the project aims to **further research and technological innovation**, **promote language accessibility**, and **educate members of the public and government bodies** on the value of Irish language data. Additionally, the project collaborates with the European Commission to share language data with the online machine translation system, eTranslation (Commission) in order to enhance the performance and accuracy of their EN<>GA engine.

2 Data Curation and Back-translation Experiments

Text data shared to the eSTÓR platform originates mainly from individuals or organisations in the public or government bodies of Ireland, often through direct contact. To date, 188 parallel language resources, totalling 185,343 Translation Units have been uploaded and processed on the eSTÓR platform, and 201,719 words of monolingual data. While sourcing the data from trusted language producers encourages reuse of existing high-quality data sources, and spreads awareness of the importance of data sharing, this resource-intensive approach is difficult to scale in order to meet the increasing demands for large data collections. Additionally, the text types collected from these sources offer limited variety of style, tone, and topic, resulting in unbalanced coverage in NLP models.

To address these issues, the eSTÓR project has begun experimenting with alternative methods of data collection and production. Web crawling is a popular method of sourcing large quantities of language data that can be largely automated, but the quality is difficult to ensure. Employing a blend of manual inspection and automatic filtering, the eSTÓR project has experimented with selecting high-quality articles from Irish Wikipedia Vicipéid (`https://ga.wikipedia.org/`), and using the eTranslation Irish-to-English General model to perform back-translation to generate synthetic parallel datasets covering diverse topics. This dataset can then be employed as test data to investigate coverage of existing Irish MT models.

Proceedings of Machine Translation Summit XX Volume 2, pages 115–116, June 23-27, 2025

3 Data Cleaning

Much of the importance of the eSTÓR project lies in the data cleaning work, which requires meticulous attention to detail and a robust knowledge of Irish and English language to ensure correctly-aligned, relevant, and high-quality clean data. The impact of this essential work in the development of powerful NLP tools and applications is still underplayed in the larger NLP community (Sambasivan et al., 2021), and precise details of the task can often be glossed over or omitted in reporting. There are many components to process the data in its raw form, although at a minimum the following steps are undertaken:

Initial Input Assessment: Typically files uploaded to eSTÓR are aligned (e.g. translation memories, spreadsheets, aligned plain-text), unaligned editable (e.g. word processing types, unaligned plain-text), or unaligned uneditable (e.g. PDF generated by software). Raster formats are typically not accepted due to the additional challenge of running Optical Character Recognition, although scans of hard-copy data are currently being processed as part of collaborative digitisation work.

Text Extraction and Normalisation: While text file types are processed using hand-written text extraction tools, library support is used for extracting text from binary file types (e.g. PDFs). Normalisation of text encoding maintains consistency throughout the data, and helps prevent incorrectly encoded characters or Unicode-equivalence errors.

Language Identification: Text is sampled at regular intervals throughout the dataset to perform language identification and ensure that the file contains the correct language. A standardised Irish language model was trained for this task, using the langdetect[1] Python port of the original tool (Nakatani, 2010).

Sentence Splitting: It is often necessary to reconstruct sentence boundary information in order to produce the sentence-aligned output. This task can be as trivial as splitting on sentence-final punctuation (e.g. '.', '?'), but becomes more challenging when processing text containing e.g. abbreviations (e.g. 'etc.', 'Dr.'). As abbreviations in Irish differ from English (e.g. *uimhir* 'number' is abbreviated as 'uimh.'), it was necessary to define bespoke rules to process most of these cases automatically.

Document and Sentence Alignment: While unmatched Irish text files can be published as monolingual data, any files uploaded in English must be aligned with an Irish file to be considered of use. Sentence alignment ensures that the text on each line of each aligned file pair corresponds with the text on that same line in the other language, employing the Hunalign (Varga et al., 2005) tool.

Verification: The final step is to assess aligned document pairs to ensure that the data has been correctly processed according to specified criteria (e.g. numerals appearing on one side should appear on the other). The text is checked through a series of automatic checks, and potential bad alignments are flagged for manual review.

4 Conclusion

We present the eSTÓR project, an effort in curating and cleaning high-quality Irish text data for the development of language technology, including improved MT engines. The project has many components, but this paper focuses on the data cleaning and selection tasks, which constitutes a vital step in the development of any NLP applications.

[1] https://pypi.org/project/langdetect/

References

European Commission. eTranslation - The European Commission's Machine Translation System. https://commission.europa.eu/resources-partners/etranslation_en.

Teresa Lynn. 2022. Report on the Irish language. https://european-language-equality.eu/deliverables/. Technical Report D1.20, European Language Equality Project.

Shuyo Nakatani. 2010. Language detection library for java.

Ailbhe Ní Chasaide, Neasa Ní Chiarán, Elaine Uí Dhonnchadha, Teresa Lynn, and John Judge. 2022. Digital Plan for the Irish Language Speech and Language Technologies 2023-2027. Available at https://assets.gov.ie/241755/e82c256a-6f47-4ddb-8ce6-ff81df208bb1.pdf.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *"Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems"*, pages 1–15.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP-2005)*, pages 590–596.