PRINCIPLE: Providing Resources in Irish, Norwegian, Croatian and Icelandic for the Purposes of Language Engineering

Andy Way & Federico Gaspari
ADAPT Centre
School of Computing
Dublin City University
Dublin 9, Ireland
{Firstname.Lastname}@adaptcentre.ie

Abstract

PRINCIPLE is a new 2-year project starting in September 2019 funded by the European Commission under the Connecting Europe Facility (CEF) programme. Parallel data for Croatian, Icelandic, Irish and Norwegian are in relatively short supply, so that the quality of the eTranslation machine translation (MT) engines is less than would be the case if larger parallel corpora were available. PRINCIPLE will gather parallel data for these languages and English, evaluate the quality of the gathered resources via MT, and deliver corpora deemed to be of high quality to eTranslation for improved MT engine training.

1 Languages, Activities and Partners

The PRINCIPLE project focuses on the identification, collection and processing of language resources (LRs) for four under-resourced European languages: Croatian, Icelandic, Irish, and Norwegian (covering both varieties: Bokmål and Nynorsk).

It focuses on providing data to improve translation quality in two Digital Service Infrastructures (DSIs)[1] – eJustice and eProcurement – via domain-specific MT engines, over a 2-year period (September 2019 to August 2021).

The main activities in PRINCIPLE are:

(i) use-case analysis, data requirements and data preparation,

(ii) identification and collection of LRs,

(iii) development, evaluation and deployment of MT systems,

(iv) exploitation and sustainability, and

(v) dissemination.

The project is coordinated by the ADAPT Centre at Dublin City University (Ireland), and the partners are Iconic Translation Machines (Dublin, Ireland), the University of Zagreb (Croatia), the National Library of Norway in Oslo, and the University of Iceland in Rejkyavik.

2 Data Collection and Verification

PRINCIPLE will provide high-quality curated data via ELRC-SHARE,[2] a repository for documenting, storing, browsing and accessing LRs collected through the European Language Resource Coordination[3] network to feed the CEF eTranslation engines. MT engines will be offered to the 'early adopter' public administration partners in the four countries to validate the LRs collected based on the specific use-cases determined by public bodies within each country.

While public administrations are already able to upload their data sets directly to ELRC-SHARE, for low-resource languages such as those of focus in the project, this has been relatively unsuccessful to date. In PRINCIPLE, partners will

[1] https://ec.europa.eu/digital-single-market/en/news/connecting-europe-facility-cef-digital-service-infrastructures

[2] https://elrc-share.eu/

[3] http://lr-coordination.eu/

avail of their local contacts in each (relatively small) country to try to persuade key stakeholders of the benefit of releasing corpora in their possession, negotiating in each case the most permissive terms possible for distribution and further reuse.

However, rather than just acting as data collectors, and passing data blindly to ELRC-SHARE, the ADAPT MT team at DCU, Iconic and the University of Zagreb all have ample experience of building MT engines, including for the low-resource language pairs of the project. Dowling et al. (2018) compare statistical MT and neural MT performance for English-Irish; Klubička et al. (2017) built Croatian-English neural MT systems with superior quality to Google Translate;[4] and Gupta et al. (2019) addresses the issue of robustness in real commercial neural MT systems.

Accordingly, PRINCIPLE will build in-house baseline MT engines for each language pair and domain, add incremental amounts of data gathered, retrain the MT engines, and only submit data to eTranslation if improvements in MT quality are clearly visible via both automatic metrics and human evaluation.

Once the utility of the datasets has been verified in this way by the project partners, parallel data in 50K sentence-pair batches will be uploaded to ELRC-SHARE for use by the eTranslation engines which will be important to break down language barriers via MT capability to provide multilingual access to all DSIs by European and national public administrations for the languages covered under this project.

As well as these clear benefits to eTranslation, public administrations which have agreed to partner with the project will be able to use the in-house MT engines developed by the PRINCIPLE technical partners for the duration of the project.

3 Relationship with other CEF Projects

The experience of evaluating commercial MT systems for deployment in public administrations in the iADAATPA project[5] (cf. Castilho et al., 2019) will greatly benefit PRINCIPLE.

PRINCIPLE intends to promote awareness and use of National Relay Stations (NRSs), which are designed to effectively collect, process and share language resources that can be used for MT training under the European Language Resource Infrastructure (ELRI) project (Etchegoyen et al., 2018).[6] NRSs have already been made available and promoted by ELRI in Ireland, France, Portugal and Spain. PRINCIPLE will encourage the extension of new NRSs to Croatia, Iceland and Norway for their respective languages.

Acknowledgements

PRINCIPLE is generously co-financed by the European Union Connecting Europe Facility under Action 2018-EU-IA-0050 with the specific grant agreement INEA/CEF/ICT/A2018/1761837.

References

Castilho, Sheila, Natalia Resende, Federico Gaspari, Andy Way, Tony O'Dowd, Marek Mazur, Manuel Herranz, Alex Helle, Gema Ramirez-Sanchez, Victor Sanchez-Cartagena, Marcis Pinnis, and Valters Sics. 2019. Large-scale Machine Translation Evaluation of the iADAATPA Project. In *Proceedings of MT Summit XVII*, Dublin, Ireland.

Dowling, Meghan, Teresa Lynn, Alberto Poncelas and Andy Way. 2018. SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018),* Boston, MA., pp.12—20.

Etchegoyhen, Thierry, Borja Anza Porras, Andoni Azpeitia, Eva Martínez Garcia, Paulo Vale, José Luis Fonseca, Teresa Lynn, Jane Dunne, Federico Gaspari, Andy Way, Victoria Arranz, Khalid Choukri, Vladimir Popescu, Pedro Neiva, Rui Neto, Maite Melero, David Perez Fernandez, Antonio Branco, Ruben Branco and Luis Gomes. 2018. ELRI - European Language Resources Infrastructure. In *Proceedings of The 21st Annual Conference of the European Association for Machine Translation (EAMT 2018),* Alicante, Spain, p.351.

Gupta, Rohit , Patrik Lambert, Raj Patel and John Tinsley. 2019. Improving Robustness in Real-World Neural Machine Translation Engines. In *Proceedings of MT Summit XVII*, Dublin, Ireland.

Klubička, Filip, Antonio Toral and Victor Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics* **108** (1): 121—132.

[4] http://translate.google.com
[5] http://iadaatpa.com/
[6] http://www.elri-project.eu/