Community-driven machine translation for the Catalan language at Softcatalà

Xavi Ivars-Ribes, Jordi Mas,
Marc Riera Jaume Ortolà, David Cànovas
Softcatalà
{xavivars,jmas,marcriera,
jaumeortola,davidcanovas}
@softcatala.org

Mikel L. Forcada
Softcatalà and
Prompsit Language Engineering
mlf@prompsit.com

Abstract

Among the services provided by Softcatalà, a non-profit 25-year-old grassroots organization that localizes software into Catalan and develops software to ease the generation of Catalan content, one of the most used is its machine translation (MT) service, which provides both rule-based MT and neural MT between Catalan and twelve other languages. Development occurs in a community-supported, transparent way by using free/open-source software and open language resources. This paper briefly describes the MT services at Softcatalà: the offered functionalities, the data, and the software used to provide them.

1 Introduction

Softcatalà[1] is a non-profit organization dedicated to promoting the use of Catalan in the realm of computing, internet, and new technologies. This organization leverages a volunteer workforce composed of computer specialists, philologists, translators, students, and others. These volunteers contribute to the translation of software interfaces and documentation into Catalan, while also developing tools that facilitate the creation and use of Catalan-language content.

In addition to a long history of providing Catalan-localized versions of popular software, Softcatalà offers a number of web-based documentation and language-related services, such as a grammar and spelling checker for Catalan, a video and audio transcription service, or machine translation (MT) systems. The service is mainly ad-supported, free to use, and very popular in the Catalan language community.[2] This paper describes Softcatalà's MT service: the functionalities offered, the data, and the software used to provide them.

[1] https://www.softcatala.org

2 The service

Softcatalà's MT service[3] provides MT between Catalan and Aragonese, Occitan (both Aranese and Languedocian), French, English, Italian and Spanish using the free/open-source rule-based MT platform Apertium, and between Catalan and Dutch, English, French, Galician, German, Italian, Japanese, Portuguese, and Romanian using neural MT. Most of our users are students, teachers, and public workers.

3 The inner workings

3.1 Rule-based machine translation

The rule-based MT systems powering Softcatalà's machine translation services are all based on the free/open-source machine translation platform Apertium (Forcada et al., 2011).[4] The first such service, between Catalan and Spanish, was launched in 2010[5] (Ivars-Ribes and Sánchez-Cartagena, 2011). Softcatalà has contributed massively to this platform, particularly by improving the language data (dictionaries, rules) and the configuration of the MT pipelines used for language pairs involving Catalan. Due to the commitment of key Softcatalà developers as part of the Apertium community, their contributions soon propagate to other services based on Apertium, improving their performance; for instance, SALT.usu, the official Spanish↔Valencian[6] MT system of the Valencian

[2] More than 220,000 translations/day in 2023.

[3] https://www.softcatala.org/traductor/

[4] https://apertium.org

[5] Previously, the (now extinct) machine translation service interNOSTRUM was offered since 2000.

[6] Valencian is the name given in the Valencia region to the local variety of Catalan; in writing, the standards used in Valen-

regional government. Anyone can install these improved systems locally off the Apertium webpage.

3.2 Neural machine translation

Softcatalà's neural MT systems[7] use the `OpenNMT-tf`[8] sequence learning toolkit, which in turn is based on TensorFlow[9] version 2; they are trained on publicly-available parallel corpora,[10] using a publicly documented training procedure;[11] all text-processing and training software is free/open-source. Trained models[12] and Docker containers are available for anyone to download and deploy in their own servers. Memory- and speed-optimized models (using `CTranslate2`[13]) are also provided, which allow for CPU-only inference and therefore produce a smaller CO_2 footprint during inference.

4 Evaluation

In Table 1 we report the latest automatic evaluation results (BLEU using SacreBLEU's default 13a tokenization (Post, 2018)) for Softcatalà's MT systems, and a comparison with Google Translate, Meta's NLLB model `nllb-200-3.3B`, and Opus-MT,[14] using the Flores200 test set.[15]

Note that the results correspond to models used in production with modest hardware (CPU), which strike a balance between accuracy and speed; BLEU could be improved with slower models. Also note that Flores200 was produced translating from English to many of the other languages, and that means that in language pairs not containing English, say, `pt`–`ca`, sentence pairs are quite different from what would be obtained if translating directly from `pt` to `ca`; results have therefore to be taken with additional caution.

As can be seen in Table 1, the BLEU scores obtained by Softcatalà's systems, when evaluated against Flores 200: (a) are consistently better than Opus-MT's freely-available models; (b) lag well behind Google Translate, a much larger commercial model for most pairs, but get quite close to it for `es`–`ca`, `ca`–`es`, `ca`–`gl`, and `en`–`ca`; (c) are competitive compared to those by Meta's much larger NLLB model. Note that Google regularly updates their MT models; the results shown are about one year old (for details, see `https://github.com/Softcatala/nmt-models`). We plan to publish additional metrics in the next training round.

Pair	SC	F200	Goo	NLLB	Opus	Sent. pairs
de–ca	34.8	28.9	35.5	30.7	18.5	3142257
ca–de	28.5	25.4	32.9	29.1	15.8	3142257
en–ca	46.9	43.8	46.0	41.7	29.8	7856208
ca–en	47.4	43.5	47.0	48.0	29.6	7856208
fr–ca	41.3	31.6	37.3	33.3	27.2	2566302
ca–fr	41.4	35.4	41.7	39.6	27.9	2566302
gl–ca	74.1	31.4	36.5	33.2	N/A	2710149
ca–gl	80.7	31.9	33.1	31.7	N/A	2710149
it–ca	39.7	26.5	30.6	27.8	22.0	2584598
ca–it	36.2	24.5	27.5	26.0	19.2	2584598
ja–ca	24.9	17.8	23.4	N/A	N/A	1997740
ca–ja	21.3	19.8	32.5	N/A	N/A	1997740
nl–ca	30.4	20.3	27.1	24.8	15.8	2208538
ca–nl	27.6	18.2	23.4	21.8	13.4	2208538
oc–ca	74.9	32.5	N/A	36.2	N/A	2711350
ca–oc	78.8	28.9	N/A	27.8	N/A	2711350
pt–ca	41.6	33.9	38.7	34.5	28.1	2043019
ca–pt	39.0	32.3	40.0	36.5	27.5	2043019
es–ca	88.8	22.6	23.6	25.8	22.5	7596985
ca–es	87.5	24.2	24.2	25.5	23.2	7596985

Table 1: BLEU scores for the latest versions of Softcatalà's (SC's) MT systems. **SC**: SC using SC's own test sets; `F200`: SC using the Flores 200 test set; `Goo`, `NLLB` and `Opus-MT`: results of these three systems using the Flores200 test set.

5 Concluding remarks

The community-driven effort of Softcatalà, a grassroots organization devoted to digitally enable the Catalan language, has managed to provide the community with competitive, freely-available, open machine translation systems that anyone can use or even improve using free/open-source software.

References

Forcada, Mikel L, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Ivars-Ribes, Xavier and Victor M. Sánchez-Cartagena. 2011. A widely used machine translation service and its migration to a free/open-source solution: the case of softcatalà. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 61–68, Barcelona.

Post, Matt. 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.

cia and Catalonia or the Balearic Islands are not too different.

[7] `https://github.com/Softcatala/nmt-softcatala`

[8] `https://github.com/OpenNMT/OpenNMT-tf`

[9] `https://www.tensorflow.org/`

[10] `https://github.com/Softcatala/parallel-corpus`

[11] `https://github.com/Softcatala/nmt-models/`

[12] `https://github.com/Softcatala/nmt-models/`

[13] `https://github.com/OpenNMT/CTranslate2`

[14] `https://github.com/Helsinki-NLP/Opus-MT`

[15] `https://github.com/facebookresearch/flores`