Similar Language Translation for Catalan, Portuguese and Spanish Using Marian NMT

Reinhard Rapp
Athena Research Center
Magdeburg-Stendal University of Applied Sciences
University of Mainz
reinhardrapp@gmx.de

Abstract

This paper describes the SEBAMAT contribution to the 2021 WMT Similar Language Translation shared task. Using the Marian neural machine translation toolkit, translation systems based on Google's transformer architecture were built in both directions of Catalan–Spanish and Portuguese–Spanish. The systems were trained in two contrastive parameter settings (different vocabulary sizes for byte pair encoding) using only the parallel but not the comparable corpora provided by the shared task organizers. According to their official evaluation results, the SEBAMAT system turned out to be competitive with rankings among the top teams and BLEU scores between 38 and 47 for the language pairs involving Portuguese and between 76 and 80 for the language pairs involving Catalan.

1 Introduction

In recent years, neural machine translation (NMT) has become the state of the art in machine translation (MT). Using toolkits such as Marian NMT (Junczys-Dowmunt et al., 2018), it is relatively straightforward to construct end-to-end NMT systems which need only little pre-processing of the training corpora and post-processing of the system output. As NMT is a supervised approach to MT based on machine learning technology, training is usually conducted using sentence aligned human translations. Given a large number of source/target-language sentence pairs, the neural system fully automatically learns how to translate.

The SEBAMAT submission to the Similar Language Translation (SLT) task[1] of the 6th Conference on MT is based on work conducted as part of the SEBAMAT project[2] (semantics-based machine translation; Rapp & Tambouratzis, 2020). This project has a focus on experiments introducing semantics into MT but, for comparative purposes, also deals with standard NMT systems. The latter were used as the basis for the current shared task. During the SEBAMAT project a number of MT systems had been developed for language pairs involving English, French, German, Greek and Spanish, but there had been no prior work on Catalan and Portuguese. The aims of the participation in the SLT shared task were the following:

- See in how far the SEBAMAT-based MT systems are competitive.
- Extend the number of SEBAMAT languages by Catalan and Portuguese.
- Find out whether systems for new language pairs can be developed in a very short time.
- See whether reasonably well working systems can be developed without much proficiency of the respective languages on the developer side.

2 Resources

2.1 Corpora

For the training of the NMT systems sentence-aligned parallel corpora are required. We used all parallel corpora suggested by the SLT shared task organizers who in their task description explicitly stated that no additional parallel corpora were allowed for training.

For Catalan–Spanish the following parallel corpora were used:

- Wiki Titles v3 (476,475 sentence pairs) (Barrault et al., 2020)[3]
- ParaCrawl (6,870,183 sentence pairs) (Bañón et al., 2020)
- DOGC v2 (10,933,622 sentence pairs) (Tiedemann, 2012).

[1] http://www.statmt. org/wmt21/similar.html

[2] https://cordis.europa.eu/project/id/844951

[3] http://data.statmt.org/wikititles/README

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 292–298
November 10–11, 2021. ©2021 Association for Computational Linguistics

For Portuguese–Spanish, these parallel corpora were used:

- Europarl v10 (1,801,845 sentence pairs) (Koehn, 2005)
- News Commentary v16 (48,259 sentence pairs) (Tiedemann, 2012)
- Wiki Titles v3 (649,833 sentence pairs) (Barrault et al., 2020)
- Tilde MODEL (13,464 sentence pairs) (Rozis & Skadiņš, 2017).
- JRC-Acquis (1,650,126 sentence pairs) (Steinberger et al., 2006; Tiedemann, 2012)[4]

The above length specifications were taken from the SLT 2021 website's corpus download page. We did not use any of the comparable corpora provided by the SLT task organizers which, among others, included about 65 million sentences of Spanish news crawl.[5] The reason is that in the SEBAMAT project we achieved fairly good translation results when training with the Europarl corpus only. For example, we obtained BLEU scores (Papineni et al., 2002) well above 40 for Spanish–English and Greek–English when evaluated with randomly held out data. The Europarl corpus comprises in the order of 2 million sentences per language pair for many languages. As the parallel corpora for the shared task were much larger than this (with the Portuguese–Spanish language pair even including the respective language parts of the Europarl corpus), we saw no need to extract additional parallel sentences from comparable corpora. Such sentences are usually much noisier than parallel sentences based on human translations and could therefore possibly even reduce the quality of the NMT training in this high resource scenario.

Given the good quality of the training data provided by the shared task organizers, we only had to convert some of the files from a two-column translation memory format to the standard Moses format, and then concatenate all files of the source language as well as all files of the target language to form a large parallel training set. The concatenation was done in the order as listed above. However, as Marian NMT by default randomly shuffles the sentence pairs for training, the order of concatenation should not be of importance in our scenario.

2.2 Hardware

Marian NMT supports training using CPUs or GPUs. According to our experiments in the SEBAMAT project, training times in NMT can typically be reduced by about two orders of magnitude by conducting the training on a current GPU rather than on a (single) CPU. We therefore used a PC with an nVidia RTX 3090 GPU, supported by an i9 CPU. With 24 GB of memory, 28.3 billion transistors and 35.58 TFLOPS FP32 (float) performance, this GPU is state of the art in 2021, so – depending on parameter settings – with a single GPU we typically had training times of only a few hours per language pair. As our operating system we used Ubuntu 20.04 LTS.

As a side note, let us mention that performance in CPU-based training can be increased by using several CPU cores in parallel, which is supported by Marian NMT. With the 16 cores of the i9 processor, this looks promising if an appropriate GPU is not at hand. However, according to our experiments, each of the processors requires the full amount of memory. Therefore, if we assume 8 GB of memory per CPU core (which is typically the minimum for serious NMT work), we would require a total of 128 GB of RAM if we wished to use all 16 cores.

2.3 Software

As the translation engine we used the Marian NMT toolkit as it is well established and, for the reason that it is implemented in the C++ programming language, runs very fast (Kim et al., 2019), thereby substantially reducing training times. This is a particularly important consideration in a shared task where time tends to be very limited.

Marian NMT was installed on the above PC together with the nVidia driver and CUDA software. Our pre-processing pipeline involves the following steps: tokenization, cleaning, i.e. removal of very long sentences and sentence pairs with very different lengths, true-casing, and byte-pair encoding. For the first three steps we used the Moses tools *tokenizer*, *clean-corpus-n*, and *truecase* (Koehn et al., 2007).[6] For byte-pair encoding we used Rico Sennrich's Python program *bpe* (Sennrich et al., 2016). For post-processing of the translations, the tokenization and true-casing was reversed using the Moses tools *detruecase* and *detokenizer*.

[4] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

[5] http://data.statmt.org/news-crawl/README

[6] http://statmt.org/moses/

```
# train model
if [ ! -e "model/model.npz.best-translation.npz" ]
then
  $MARIAN_TRAIN \
  --devices $GPUS --sync-sgd --seed 1111 \
  --model model/model.npz --type transformer \
  --train-sets data/corpus.bpe.pt data/corpus.bpe.es \
  --max-length 100 \
  --vocabs model/vocab.ptes.yml model/vocab.ptes.yml \
  --mini-batch-fit -w 10000 --maxi-batch 1000 \
  --early-stopping 10 --cost-type=ce-mean-words \
  --valid-freq 5000 --save-freq 5000 --disp-freq 500 \
  --valid-metrics translation ce-mean-words perplexity cross-entropy \
  --valid-sets data/corpus-dev.bpe.pt data/corpus-dev.bpe.es \
  --valid-script-path "bash ./scripts/validate.sh" \
  --valid-translation-output data/valid.bpe.es.output --quiet-translation \
  --valid-mini-batch 64 \
  --beam-size 6 --normalize 0.6 \
  --log model/train.log --valid-log model/valid.log \
  --enc-depth 6 --dec-depth 6 \
  --transformer-heads 8 \
  --transformer-postprocess-emb d \
  --transformer-postprocess dan \
  --transformer-dropout 0.1 --label-smoothing 0.1 \
  --learn-rate 0.0003 --lr-warmup 16000 \
  --lr-decay-inv-sqrt 16000 --lr-report \
  --optimizer-params 0.9 0.98 1e-09 --clip-norm 5 \
  --tied-embeddings-all \
  --exponential-smoothing \
  --overwrite --keep-best
fi
```

Table 1: Parameters for Marian NMT training (Portuguese → Spanish).

The Moses tokenizer works well for Spanish and Portuguese, but has some problems with Catalan as there are some peculiarities in this language, most notably the interpunct (as used e.g. in the word *cel·la*). An insightful discussion on this can be found on GitHub.[7] As we did not have time to adapt the tokenizer to Catalan, to account for a few obvious errors, we did some minimalistic automatic post-processing (replacing a few short character sequences) as described in Section 4.

3 Experiments

To obtain any information on the training data and on the development and test sets, it was required to register for the shared task. We did so on July 13, 2021, so had seven days until July 19 when the submission of the results was due. This did not leave us much time for parameter optimization which is why we mostly took the standard parameters as suggested in the Marian NMT documentation for the Transformer architecture (Vaswani et. al, 2017).

We only did a few test runs with various settings concerning the number of merge operations in byte pair encoding (later to be referred to as *vocabulary size*), but did not have time to systematically optimize it. In our primary submissions, this parameter is set to 40,000, whereas in the comparative submissions, as in some previous SEBAMAT work, it is set to 85,000. For all language pairs and data sets (development and test), the smaller size performed better in terms of BLEU scores, although the exact size appears to be not very critical within a wide range of values.

To provide details on the core part of our experiments, in Table 1 we show the script for the Marian NMT training. As the parameters are well described in the Marian NMT documentation, we do not discuss them here. Let us only mention that during training BLEU scores are computed periodically on the development set, and that the training stops if the best score cannot be improved within ten iterations.

[7] https://github.com/alvations/sacremoses/issues/43

4 Results

We pre-processed the corpora as described in section 2 and trained the system for the language pairs Catalan→Spanish, Spanish→Catalan, Portuguese →Spanish and Spanish→Portuguese. We then inspected the translation results. For all language pairs the translations looked ok except for Spanish→Catalan. For this language pair we noticed, apparently because the tokenizer was not well suited for Catalan, the following three types of systematic errors in the output:

- There were extra spaces within Catalan words such as *paral·lel* because the *interpunct* (*punt volat*) was incorrectly interpreted as a separator between words rather than between syllables, which is why during tokenization blanks were inserted around it. We corrected this by replacing all «␣·␣» sequences («␣» stands for blank) in the translation output by «·».
- We found extra spaces before the apostrophe in phrases such as «l'efectiu» or after the apostrophe in phrases such as «d'informes». We therefore removed all spaces before and after apostrophes in the translation output.
- We noticed that whereas in the translated output «'» was used as the apostrophe, the sample translations in the development set used «´» instead. The reason is probably a discrepancy between training corpora and development sets. Assuming that the test set would have the same characteristics as the development set, we replaced in the translation output all occurrences of the former by the latter.

Vocabulary size	Language pair	BLEU score
40,000 (primary submission)	ca–es	80.72
	es–ca	83.32
	pt–es	50.37
	es–pt	44.96
85,000 (contrastive submission)	ca–es	79.40
	es–ca	81.21
	pt–es	47.29
	es–pt	42.77

Table 2: Results for the development sets. ca = Catalan, es = Spanish, pt = Portuguese. Without the interpunct and apostrophe substitutions, the BLEU score of es–ca (primary) is 69.40 and the BLEU score of es–ca (contrastive) is 68.01.

Vocabulary size	Language pair	BLEU	RIBES	TER	Rank
40,000 (primary)	ca–es	78.65	94.76	15.805	2
	es–ca	79.69	95.76	14.632	1
	pt–es	46.51	86.31	41.235	2
	es–pt	40.35	84.99	45.258	2
85,000 (contrastive)	ca–es	76.78	94.46	17.067	5
	es–ca	77.32	95.35	16.744	3
	pt–es	43.12	84.99	45.068	4
	es–pt	38.90	83.89	47.044	3

Table 3: Shared task results for the test sets as computed by the shared task organizers.

Especially the substitution of the apostrophes resulted in an improvement of several BLEU points, whereas the effects of blank removal before and after apostrophes differed depending on the software used for automatic evaluation. When using *Tilde's interactive BLEU score evaluator*[8] this change had no effect, whereas with the Moses *multi-bleu-detok.perl* tool, which we used in our scripts, a small improvement was obtained. The discrepancy can be explained by assuming that tools for computing BLEU scores often introduce some forms of tokenization or de-tokenization by themselves, and that these operations can slightly differ between tools.

Table 2 shows the BLEU scores obtained with the *multi-bleu-detok.perl* tool on the development sets for the four language pairs and the two parameter settings (byte pair encoding vocabulary sizes of 40,000 vs. 85,000). Table 3 shows the official BLEU scores for the test sets as computed by the shared task organizers who also provided scores for the RIBES (Isozaki et al., 2010) and TER (Snover et al., 2006) measures on the evaluation section of the SLT webpage. These scores we cite in Table 3. The last column shows our submissions' ranks among the other teams participating in the competition. As can be seen, our primary submissions (byte pair encoding vocabulary size of 40,000) won the competition for Spanish → Catalan, and ranked second for the other three language pairs.

As can be expected from the evaluation scores, the translation quality is to the most part very good. This is particularly true for the language pairs involving Catalan. Table 4 shows a translation example for Portuguese → Spanish which is harder than Catalan → Spanish. For all language pairs, we

[8] https://www.letsmt.eu/Bleu.aspx

found a few occurrences of hallucinations (Raunak et al., 2021), mainly of the type where the beginning of a long sentence is translated well but towards the end a phrase translation is repeated over and over again. To give an idea how often this happened: manual inspection of the results for Catalan → Spanish (vocabulary size 85,000) showed that such hallucinations occurred in 7 of the 970 sentence translations of the development set.

First three segments of test set (Portuguese)	SEBAMAT translation (Spanish)
O plano pretende contribuir para a realização dos objetivos da política comum das pescas e, em especial, para garantir que a exploração dos recursos biológicos marinhos vivos restabeleça e mantenha as populações das espécies exploradas acima de níveis que possam produzir o rendimento máximo sustentável (MSY).	El plan pretende contribuir a la consecución de los objetivos de la política pesquera común y, en particular, garantizar que la explotación de los recursos biológicos marinos vivos restablezca y mantenga a las poblaciones de especies explotadas por encima de niveles que puedan producir el rendimiento máximo sostenible (RMS).
Coordenar, em consulta com a Comissão, os métodos de diagnóstico da doença de Newcastle nos Estados-Membros, nomeadamente, mediante:	Coordinar, en consulta con la Comisión, los métodos de diagnóstico de la enfermedad de Newcastle en los Estados miembros, en particular mediante:
Por conseguinte, atualmente é preciso cofinanciamento da União para assegurar que a Finance Watch e a Better Finance recebam os recursos necessários para atingirem os objetivos pretendidos nos próximos anos, e para proporcionar estabilidade financeira a ambas as organizações, aos seus peritos e ao seu pessoal administrativo, que até agora conseguiram lançar as suas atividades relevantes em pouco tempo.	Por lo tanto, hoy es necesario cofinanciar la Unión para garantizar que Finance Watch y Better Finance reciban los recursos necesarios para alcanzar los objetivos deseados en los próximos años y para proporcionar estabilidad financiera a ambas organizaciones, a sus expertos y a su personal administrativo, que hasta ahora han logrado lanzar sus actividades relevantes en poco tiempo.

Table 4: First three Portuguese segments of the SLT 2021 test set and their translations to Spanish as produced by the primary SEBAMAT NMT system.

The hallucinations could be detected by looking at the ratio of sentence lengths between a source language sentence and its translation, and/or by detecting repetitive phrases towards the end of a sentence translation. However, according to Raunak et al. (2021), hallucinations are a problem of training data quality, which to improve would have been too time-consuming. We thought of greedy solutions such as cutting off repetitive sentence ends, but did not implement them for lack of time and as they would be hard to justify.

5 Discussion and conclusions

Given the observation that the language pairs involving Catalan achieved considerably higher evaluation scores than those involving Portuguese, the question arises how this can be explained. Our somewhat speculative answer is as follows: As Catalan's grammar and sentence structure is very similar to Spanish, with differences mainly on the vocabulary side, extremely high scores can be achieved because in many cases, often in the form of a word-by-word translation, there is just one obvious way how to translate a sentence for both man and machine. This is only to a lesser extend true for Spanish–Portuguese, so the lower scores are likely caused by more variability in acceptable translation options, rather than by lower translation quality.

When comparing the BLEU scores in Tables 2 and 3, it can be seen that our system performed significantly worse on the test sets than it did on the development sets. From this we conclude that probably the test data is less representative of the training data than the development data. Problems with overfitting seem unlikely as in the previous SEBAMAT work we had usually used randomly held out sentences of the training data for both development and testing. In such a scenario, the results were very similar in both cases, with only minor unsystematic discrepancies in BLEU scores.

Finally, let us try to answer the questions raised in the introduction. As it was ranked first for Spanish → Catalan and second for the other three language pairs, it appears that especially our primary systems (with vocabulary size 40,000) are competitive. Let us mention, however, that all participating systems showed similarly convincing evaluation scores, so that minor differences in parameter choice (such as vocabulary size) or in the organizers' selection of the test set may have had a noticeable impact on the rankings.

Like many other studies, this work provides once again evidence how powerful NMT is, and how well the Marian toolkit works: Within a week it was possible for a single developer to add two new language pairs (in two directions each) to the SEBAMAT portfolio, despite mediocre proficiency of Spanish and hardly any proficiency of Catalan and Portuguese. Of course, achieving good translation quality was considerably facilitated by the similarity of the languages and by the good quality and size of the training data provided by the shared task organizers.

Acknowledgments

This work was funded by the Marie Curie Individual Fellowship SEBAMAT (grant agreement number 844951) within the European Commission's Horizon 2020 Framework Programme. I would like to thank my colleagues at Athena R.C., the SLT and the WMT organizers, and all researchers who generously made available the tools and resources used in this work.

References

Bañón, Marta; Chen, Pinzhen; Haddow, Barry; Heafield, Kenneth; Hoang, Hieu; Esplà-Gomis, Miquel; Forcada, Mikel L.; Kamran, Amir; Kirefu, Faheem; Koehn, Philipp; Ortiz Rojas, Sergio; Pla Sempere, Leopoldo; Ramírez-Sánchez, Gema; Sarrías, Elsa; Strelec, Marek; Thompson, Brian; Waites, William; Wiggins, Dion; Zaragoza, Jaume (2020). ParaCrawl: Web-scale acquisition of parallel corpora. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4555–4567.

Barrault, Loïc; Biesialska, Magdalena; Bojar, Ondřej; Costa-jussà, Marta R.; Federmann, Christian; Graham, Yvette; Grundkiewicz, Roman; Haddow, Barry; Huck, Matthias; Joanis, Eric; Kocmi, Tom; Koehn, Philipp; Lo, Chi-kiu; Ljubešić, Nikola; Monz, Christof; Morishita, Makoto; Nagata, Masaaki; Nakazawa, Toshiaki; Pal, Santanu; Post, Matt; Zampieri, Marcos (2020). Findings of the 2020 Conference on Machine Translation (WMT20). *Proceedings of the Fifth Conference on Machine Translation*, 1–55.

Isozaki, Hideki; Hirao, Tsutomu; Duh, Kevin, Sudoh, Katsuhito; Tsukada, Hajime (2010). Automatic evaluation of translation quality for distant language pairs. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 944–952.

Junczys-Dowmunt, Marcin; Grundkiewicz, Roman; Dwojak, Tomasz; Hoang, Hieu; Heafield, Kenneth; Neckermann, Tom; Seide, Frank; Germann, Ulrich; Aji, Alham; Bogoychev, Nikolay; Martins, André; Birch, Alexandra (2018). Marian: Fast Neural Machine Translation in C++. *Proceedings of ACL 2018, System Demonstrations*, 116–121.

Kim, Young Jin; Junczys-Dowmunt, Marcin; Hassan, Hany; Fikri Aji, Alham; Heafield, Kenneth; Grundkiewicz, Roman; Bogoychev, Nikolay (2019). From research to production and back: ludicrously fast neural machine translation. *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 280–288.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Mo-ran, Richard Zens, Chris Dyer, Ondrej Bojar, Alex-andra Constantin, Evan Herbst. (2007). Moses: Open Source Toolkit for Statistical Machine Trans-lation. Annual Meeting of the Association for Compu¬tational Linguistics (ACL), demonstration session, Prague, 177–180.

Koehn, Philipp (2005). Europarl: a parallel corpus for statistical machine translation. *Proceedings of the Tenth Machine Translation Summit,* Phuket, Thailand, 79–86.

Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, 311–318.

Rapp, Reinhard; Tambouratzis, George (2020). An overview of the SEBAMAT project. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 491–492.

Raunak, Vikas; Arul Menezes, Marcin Junczys-Dowmunt (2021). The curious case of hallucinations in neural machine translation. *NAACL-HLT*, 1172–1183.

Rozis, Roberts; Skadiņš, Raivis (2017). Tilde MODEL – multilingual open data for EU languages. *Proceedings of the 21st Nordic Conference of Computational Linguistics (NODALIDA)*, Gothenburg, Sweden, 263–265.

Sennrich, Rico; Barry Haddow, Alexandra Birch (2016). Neural machine translation of rare words with subword units. *Proc. of the 54th Annual Meeting of the ACL (Vol. 1: Long Papers)*, 1715–1725.

Snover, Matthew; Dorr, Bonnie; Schwartz, Rich; Micciulla, Linnea; Makhoul, John (2006). A study of translation edit rate with targeted human annotation. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, 223–231.

Steinberger, Ralf; Pouliquen, Bruno; Widiger, Anna; Ignat, Camelia; Erjavec, Tomaž, Tufis, Dan; Varga, Dániel (2006). The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006).* Genoa, Italy, 24–26.

Tiedemann, Jörg (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, 2214–2218.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jacob; Jones, Llion; Gomze, Aidan N. G.; Kaiser, Lukasz; Polosukhin, Illia (2017). Attention is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 6000–6010.