English-Catalan Neural Machine Translation in the Biomedical Domain through the cascade approach

Marta R. Costa-jussà, Noe Casas and Maite Melero*
TALP Research Center - Universitat Politècnica de Catalunya, Barcelona
* OTG Plan de Tecnologías del Lenguaje
marta.ruiz@upc.edu; contact@noecasas.com; maite.melero@upf.edu

Abstract

This paper describes the methodology followed to build a neural machine translation system in the biomedical domain for the English-Catalan language pair. This task can be considered a low-resourced task from the point of view of the domain and the language pair. To face this task, this paper reports experiments on a cascade pivot strategy through Spanish for the neural machine translation using the English-Spanish SCIELO and Spanish-Catalan El Periódico database. To test the final performance of the system, we have created a new test data set for English-Catalan in the biomedical domain which is freely available on request.

Keywords: Neural Machine Translation, Biomedical, English-Catalan

1. Introduction

Neural machine translation (Sutskever et al., 2014; Cho et al.,) has recently emerged as a stronger alternative to standard statistical paradigm (Koehn et al., 2003). Among other advantages, neural MT offers an end-to-end paradigm which seems to be able to generalize better from data (Bentivogli et al., 2017). However, deep learning techniques face serious difficulties for learning when having limited or low resources and machine translation is not an exception (Koehn and Knowles, 2017).

English has become the *de facto* universal language of communication around the world. In Catalonia, out of 7.5 million population only around 30% of people have knowledge of English in all competences[1]. Therefore, there are many situations where professional or automatic translations are still necessary. One of them is in medical communication patient-physician at the level of primary health care. Also in the biomedical domain it is worth mentioning that Catalonia has become a hub of global biomedical research as proven by the nearly 1% of global scientific production, 9,000 innovative companies or the fact that the sector raised a record of 153 million of euros in 2016 [2]. Therefore, English-Catalan translation in the biomedical domain is of interest not only in health communication but to properly disseminate the work in such a relevant area for Catalan economy.

English-Catalan in general —and even more in a closed domain as the biomedical one—can be considered to be a limited resourced language pair. However, there are quite large amount of resources for English-Spanish and Spanish-Catalan language pairs. Therefore, English-Catalan could take advantage of them by using the popular pivot strategies which consist in using one intermediate language to perform the translation between two other languages.

Pivot strategies have been shown to provide a good performance within the phrase-based framework (Costa-jussà et al., 2012) and also for the particular case of English-Catalan (de Gispert and no, 2006). While in the phrase-based context, pivot strategies have been widely exploited, this is not the case for the neural approaches. Pivot studies are limited to (Cheng et al., 2017) which considers a single direct approach (the cascade) contrasted with a joint trained model from source-to-pivot and pivot-to-source. Other alternatives when having no parallel data for a language pair are the multilingual approximations where several language pairs are trained together and the system is able to learn non-resourced pairs (Wu et al., 2016).

Another related research area for this study is precisely training translation systems domain-specific tasks, where there are scarce in-domain translation resources. A common approach in these cases consists in training a system with a generic corpus and then, use a small in-domain corpus to adapt the system to that particular domain. In this direction, there is a huge amount of research in the statistical approach (Costa-jussà, 2015) and also starting in the neural approach (Chu et al., 2017). Finally, there is am emerging line of research in the topic of unsupervised neural MT (Lample et al., 2018; Artetxe et al., 2018).

This study designs and details an experiment for testing the standard cascade pivot architecture which has been employed in standard statistical machine translation (Costa-jussà et al., 2012).

The system that we propose builds on top of one of the latest neural MT architectures called the Transformer (Vaswani et al., 2017). This architecture is an encoder-decoder structure which uses attention-based mechanisms as an alternative to recurrent neural networks proposed in initial architectures (Sutskever et al., 2014; Cho et al.,). This new architecture has been proven more efficient and better than all previous proposed so far(Vaswani et al., 2017).

2. Neural MT Approach

This section provides a brief high-level explanation of the neural MT approach that we are using as a baseline system, which is one of the strongest systems presented recently (Vaswani et al., 2017), as well as a glance of its differences with other popular neural machine translation architectures.

[1] Data taken from https://www.idescat.cat/

[2] http://cataloniabio.org/ca/publicacions

Table 1: Size of the parallel training corpora

Language Pair	**Corpus**	Language	Segments	Words	Vocab
En-Es	Biomedical	En	$932 \cdot 10^3$	$20,5 \cdot 10^3$	$296 \cdot 10^3$
		Es		$21,9 \cdot 10^3$	$309 \cdot 10^3$
Es-Ca	El Periódico	Es	$6,5 \cdot 10^3$	$16,5 \cdot 10^3$	$736 \cdot 10^3$
		Ca		$17,9 \cdot 10^3$	$713 \cdot 10^3$

Table 2: Size of the test set

Language Pair	**Corpus**	Language	Segments	Words	Vocab
En-Es	Biomedical	En	1000	$26,1 \cdot 10^3$	$6,1 \cdot 10^3$
		Es		$27,4 \cdot 10^3$	$6,6 \cdot 10^3$
Es-Ca	El Periódico	Es	2244	$56 \cdot 10^3$	$12,2 \cdot 10^3$
		Ca		$60,7 \cdot 10^3$	$11,7 \cdot 10^3$

Sequence-to-sequence recurrent models (Sutskever et al., 2014; Cho et al.,) have been the standard approach for neural machine translation, especially since the incorporation of attention mechanisms (Bahdanau et al., 2014; Luong et al., 2015), which enables the system to learn to identify the information which is relevant for producing each word in the translation. Convolutional networks (Gehring et al., 2017) were the second *paradigm* to effectively approach sequence transduction tasks like machine translation.

In this paper we make use of the third *paradigm* for neural machine translation, proposed in (Vaswani et al., 2017), namely the Transformer architecture, which is based on a feed-forward encoder-decoder scheme with attention mechanisms. The type of attention mechanism used in the system, referred to as *multi-head attention*, allows to train several attention modules in parallel, combining also self-attention with standard attention. Self-attention differs from standard attention in the use of the same sentence as input and trains over it allowing to solve issues as coreference resolution. Equations and details about the transformer system can be found in the original paper (Vaswani et al., 2017) and are out of the scope of this paper.

For the definition of the vocabulary to be used as input for the neural network, we used the sub-word mechanism from `tensor2tensor` package, which is similar to Byte-Pair Encoding (BPE) from (Sennrich et al., 2016). For the English-Spanish language pair, two separate 32K sub-word vocabularies where extracted, while for Spanish-Catalan we extracted a single shared 32K sub-word vocabulary for both languages.

3. Pivot Cascade Approach

Standard approaches for making use of pivot language translation in phrase-based systems include the translation cascade. The cascade approach consists in building two translation systems: source-to-pivot and pivot-to-target. In test time, the cascade approach requires two translations. Figure 1 depicts the training of the pivot systems while figure 2 shows how they are combined to devise the final one.

4. Experimental Framework: data resources

This section reports details on data to be employed in the experiments.

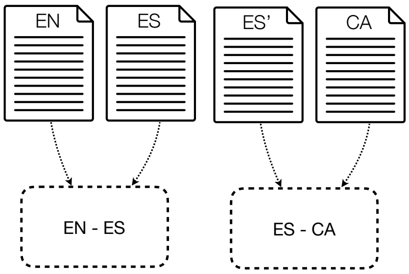

Figure 1: Pivot translation systems training.

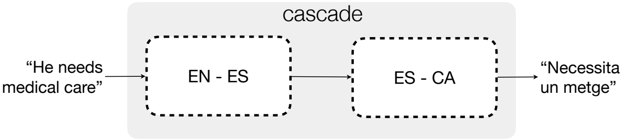

Figure 2: Pivot cascading approach at inference time.

Experiments will be performed for the English-Catalan language pair on the biomedical domain. Resources are the SCIELO database from the WMT 2016 International Evaluation Campaign (Costa-jussà et al., 2016). The English-Spanish corpus is the compilation of the corpora assigned for the WMT 2016 biomedical shared task, gathered from the Scielo database. The Spanish-Catalan corpus is extracted from ten years of the paper edition of a bilingual Catalan newspaper, El Periódico (Costa-jussà et al., 2014). The Spanish-Catalan corpus is partially available via ELDA (Evaluations and Language Resources Distribution Agency) in catalog number ELRA-W0053.

The size of the corpora is summarised in Table 1. The corpora has been pre-processed with a standard pipeline for Catalan, Spanish and English: tokenizing and keeping parallel sentences between 1 and 50 words. Additionally, for English and Spanish we used Freeling (Padró and Stanilovsky, 2012) to tokenize pronouns from verbs (i.e. *preguntándose* to *preguntando* + *se*), we also split prepositions and articles, i.e. *del* to *de* + *el* and *al* to *a* + *el*. This

Table 3: Sample end-to-end English-Catalan translations.

English	Catalan
crustacean diversity and population peaks were within the range of examples found in worldwide literature .	la diversitat de crustacis i els pics poblacionals van estar dins del rang d ’ exemples trobat en la literatura mundial .
multivariate analysis and Post-Hoc Bonferroni tests were used and relative risk and attributable fraction were calculated .	es va utilitzar anàlisis multivariat i post-Hoc de Bonferroni i es va calcular el risc relatiu i la fracció atribuïble .
this qualitative study used semi-structured interviews , with eight coordinators of the Tuberculosis Control Program in six cities of the state of Paraíba .	es tracta d ’ un estudi qualitatiu amb entrevistes semiestructurades , amb vuit coordinadors del Programa de Control de la Tuberculosi en sis municipis de l ’ estat de Paraíba .
there is no statistically significant difference in global and event free survival between the two groups .	no hi ha diferència estadísticament significativa en la supervivència global i lliure d ’ esdeveniments entre els dos grups .

was done for similarity to English. For Spanish and Catalan, we used Freeling to tokenize the text but no split with pronouns, prepositions or articles was done. The test sets come from WMT 2016 biomedical shared task in the case of English and Spanish. Since we required a gold standard in English-Catalan, we translated the Spanish test set from WMT 2016 biomedical shared task into Catalan. The translation was performed in two steps: we did a first automatic translation from Spanish to Catalan and then a professional translator postedited the output. This English-Catalan test set on the biomedical domain is freely available on request to authors. Details on the test sets are reported in Table 4.

5. Results

The results of each of the pivotal translation systems as well as the combined cascaded translation are summarized in table 4, which shows the high quality of the translations of the attentional architecture from (Vaswani et al., 2017).

The English-to-Spanish translation obtains a BLEU score of 46.55 in the test set of the WMT Biomedical test set while the Spanish-to-Catalan translation obtains a BLEU score of 86.89 in the El Periódico test set. The cascaded translation achives a BLEU score of 41.38 in the translated WMT Biometical test set.

Table 4: BLEU results.

Language	System	BLEU
EN2ES	Direct	46.55
ES2CA	Direct	86.89
EN2CA	Cascade	41.38

All BLEU scores are case-sensitive and where obtained with script `t2t-bleu` from the tensor2tensor framework, whose results are equivalent to those from `mteval-v14.pl` from the Moses package.

In order to illustrate the quality of the cascaded translations quality, some sample translations are shown in table 3.

6. Conclusion

This paper describes the data resources and architectures to build an English-Catalan neural MT system in the medical domain without English-Catalan parallel resources. This descriptive paper provides details on latest architectures in neural MT based on attention mechanisms and one standard pivot architecture that has been used with the statistical approach. The paper reports results on the baseline system of the cascade approach with the latest neural MT architecture of the Transformer.

Further experiments are required to fully characterize the potential of pivoting approaches. One of the future lines of research is to apply the pseudo-corpus approach, which consists in training a third translation system on a synthetic corpus created by means of the pivotal ones. A second future line of research is the use of recently proposed unsupervised machine translation approaches (Lample et al., 2018; Artetxe et al., 2018), which do not require large amount of parallel data.

7. Acknowledgements

This work is supported by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, through the postdoctoral senior grant *Ramón y Cajal*, the contract TEC2015-69266-P (MINECO/FEDER, UE) and the contract PCIN-2017-079 (AEI/MINECO). This work is also supported in part by an Industrial PhD Grant from the Catalan Agency for Management of University and Research Grants (AGAUR) and by the United Language Group (ULG).

8. Bibliographical References

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2017). Neural versus phrase-based mt quality: An in-depth analysis on english–german and english–french. *Computer Speech and Language*.

Cheng, Y., Yang, Q., Liu, Y., Sun, M., and Xu, W. (2017). Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3974–3980.

Cho, K., van Merrienboer, B., and Caglar Gulcehre and Dzmitry Bahdanau and Fethi Bougares and Holger Schwenk and Yoshua Bengio, title = Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, b. . P. p. . . y. . .).

Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of simple domain adaptation methods for neural machine translation. *CoRR*, abs/1701.03214.

Costa-jussà, M. R., Henríquez Q, C. A., and Banchs, R. E. (2012). Evaluating indirect strategies for chinese-spanish statistical machine translation. *J. Artif. Int. Res.*, 45(1):761–780, September.

Costa-jussà, M. R., Poch, M., Fonollosa, J. A., Farrús, M., and Mariño, J. B. (2014). A large Spanish-Catalan parallel corpus release for Machine Translation. *Computing and Informatics Journal*, 33.

Costa-jussà, M. R., España Bonet, C., Madhyastha, P., Escolano, C., and Fonollosa, J. A. R. (2016). The talp–upc spanish–english wmt biomedical task: Bilingual embeddings and char-based neural language model rescoring in a phrase-based system. In *Proceedings of the First Conference on Machine Translation*, pages 463–468, Berlin, Germany, August. Association for Computational Linguistics.

Costa-jussà, M. R. (2015). Domain adaptation strategies in statistical machine translation: a brief overview. *Knowledge and Engineering Review*, 3(05):514–520.

de Gispert, A. and no, J. B. M. (2006). Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proceedings of LREC*.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54.

Lample, G., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *International Conference on language Resources and Evaluation*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.