Pivot Machine Translation in INTERACT Project

Chao-Hong Liu, Andy Way
ADAPT Centre
Dublin City University, Ireland
chaohong.liu@adaptcentre.ie
andy.way@adaptcentre.ie

Catarina Silva and André Martins
Unbabel
Lisbon, Portugal
catarina@unbabel.com
andre.martins@unbabel.com

Abstract

The INTERnAtional network on Crisis Translation (INTERACT) project under EU Marie Skłodowska-Curie Actions (MSCA) Research and Innovation Staff Exchange (RISE) Programme aimed at researching translation in crisis scenarios. In this extended abstract, we present the work on Pivot Machine Translation under the INTERACT project.

1 Introduction

The EU INTERACT project is a staff exchange project which brings together researchers from different disciplines to collaborate on the issues arise in delivering translation in crisis scenarios. The project is led by Dr. Sharon O'Brien at Dublin City University and the consortium also includes University College London, Unbabel, Microsoft, Translators without Borders, Cochrane, University of Auckland, Arizona State University. The project starts from April 2017 for 36 months.

There are many issues addressed in different work packages in the project, e.g. involving civil translator and ethics. In this extended abstract, we describe the work on automatic translation, which includes building and adapting machine translation (MT) systems for health-related contents in scenario language translation pairs.

As a target scenario to bring translation for crisis situations, we focused on building MT systems for translation of health-related contents where parallel corpora for a language pair are very small or do not even exist.

There are several approaches to build MT systems under this circumstance. For example, in Wu and Wang (2007), a small parallel corpus is used to interpolate translation probabilities of a target translation pair in statistical machine translation (SMT). Utiyama and Isahara (2007) also compared two different SMT strategies, phrase-translation and sentence-translation. pivot machine translation (Pivot MT) is the sentence-translation strategy which builds two cascading MT systems (A-to-B and B-to-C) to realize translation from A to C, with the assumption that the parallel corpora (A—B and B—C) are much larger than direct A—C parallel corpus. The zero-shot neural MT (NMT) approach is another method to build MT systems without any direct parallel corpus (Johnson, Schuster, Le, et al.) However, experiments on the UN Parallel Corpus showed that pivoting with sentence-translation strategy is still the best practice under this circumstance.

We took three major steps to improve MT quality for health-related contents. First, manually collected and edited parallel corpora for Arabic—English and Greek—English are curated. These are small corpora at the scale of two thousand sentence pairs, and each of them are separated into development and test sets for the training of the MT systems.

Second, we used data selection methods (term frequency–inverse document frequency, cross-entropy difference and feature decay algorithm) to select data from a large parallel corpus and used the selected subset to train the resulting MT

This work has partially received funding from the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie Actions (Grant No. 734211; the EU INTERACT project). The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund.

systems. The results show that it also improved performance of NMT systems, in terms of BLEU, although the gain is smaller compared to that of SMT (Silva, Liu, Poncelas, and Way, 2018).

Third, we adapted the MT models using the curated development sets to improve the performance of in-domain (health-related contents) translation. We also compared pivot MT systems using UN Parallel Corpus 1.0 and participated in the China Workshop on Machine Translation (CWMT) shared task on pivot MT (Liu, Silva, Wang, and Way, 2018). Our pivot MT system took the first place in terms of METEOR and translation edit rate (TER) in the shared task.

In this project, we have reviewed, identified and built MT systems where only small parallel corpora are available in the scenario language translation pairs. We also manually curated Arabic—English and Greek—English parallel corpora of health-related contents in crisis situations, as development and test sets for MT system training. The performance is improved in terms of BLEU using the strategies described above and is now being evaluated by professional linguists for detailed assessment.

References

Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas and Andy Way, "Extracting In-domain Training Corpora for Neural Machine Translation Using Data Selection Methods," pp. 224–231, The Third Conference on Machine Translation (WMT18).

Chao-Hong Liu, Catarina Cruz Silva, Longyue Wang, and Andy Way, "Pivot Machine Translation Using Chinese as Pivot Language," The 14th China Workshop on Machine Translation (CWMT 2018). In: Chen J., Zhang J. (eds) Machine Translation. CWMT 2018. Communications in Computer and Information Science, vol 954. Springer, Singapore.

Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics 5, 339–351 (2017)

Utiyama, M., Isahara, H.: A comparison of pivot methods for phrase-based statistical machinetranslation. In: Proceedings of Human Language Technologies, The Conference of the NorthAmerican Chapter of the Association for Computational Linguistics (NAACL 2007). pp.484–491. Rochester, USA (2007)

Wu, H., Wang, H.: Pivot language approach for phrase-based statistical machine translation.Machine Translation21(3), 165–181 (2007)