Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish to English Cardiology Corpora

Gökhan Doğru, Adrià Martín-Mor, Anna Aguilar-Amat
Universitat Autònoma de Barcelona, Universitat Autònoma de Barcelona, Universitat Autònoma de Barcelona
gokhan.dogru@uab.cat, adria.martin@uab.cat, anna.aguilar-amat@uab.cat

Abstract

High quality clean parallel corpora is a must for creating statistical machine translation or neural machine translation systems. Although high quality parallel corpora is largely available for official languages of the European Union, the United Nations and other organization, it is hard to encounter enough amount of open parallel corpora for languages such as Turkish, which, in turn, leads to lower quality Machine Translation for these languages. In this study, we use automatic and semi-automatic procedures to collect and prepare parallel corpora in cardiology domain. We crawl a journal website and obtain 6500 Turkish abstracts and their English translations by using HTTrack. By aligning these abstracts and converting them into a translation memory in a computer-aided translation tool environment, we make it possible to use the corpora for machine translation training as well as term extraction. We argue that new tools integrating and streamlining the web crawling, alignment and cleaning steps are needed in order to support the preparation of parallel corpora for low-resource languages.

Keywords: Parallel Corpora Preparation, Medical terminology, Low-resource languages

1. Introduction

Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) approaches have yielded important advances for creating automated translations with higher qualities compared to previous approaches. SMT is a mature study field and there are open source SMT systems such as Moses as well as free platforms based on Moses such MTradumàtica[1], which has a user friendly interface for training parallel corpora and creating an SMT engine. Both of these systems are corpus-based, namely, they need a large parallel corpus to learn from. As Forcada observes "[i]n both NMT and SMT, a target sentence is a translation of a source sentence with a certain probability of likelihood" (Forcada, 2017). Although research on NMT is relatively new, there are already some open source systems such as OpenNMT, AmuNMT and Nematus. Up until now, the automatic and human quality evaluation studies on the engines trained by these systems have reported that NMT engines yield slightly better quality results, though there are some studies claiming that SMT still gives higher quality. Although this discussion will likely carry on in the near future, there is one thing that everyone agrees: Since both of these approaches are data-driven, the quality of the parallel corpora used for training the systems plays a key role in the success of the respective system. Both of the approaches rely on high quality parallel corpora for the training of the SMT or NMT system. For this reason, the selection and preparation of the parallel corpora conditions the quality of resulting MT engine. This can be illustrated by the implementation of these systems to data-rich language pairs which have large amounts of high-quality open and free parallel corpora available on the internet.

With the availability of these huge amounts of data, domain-specific parallel corpora can easily be selected and the overall training data can be prepared in a short time (it is widely accepted that MT works better with domain-specific parallel corpora). However, not all language pairs have these readily available large amounts of data especially minoritized languages and less translated languages. And with the lack of parallel corpora, it becomes very difficult to create a machine translation engine for the low-resource language pairs. In the scope of this study, low-resource languages are defined as the languages with limited amount of high quality open parallel corpora on the internet. We think that with ever-increasing amount of open tools for MT, the high-quality parallel corpora selection and preparation will be very important to close the gap between low-resource languages including minoritized languages. Besides, we believe that one of the reasons why low-resource languages score lower in machine translation evaluation is that generally machine translation systems including these languages are trained with either small amount of data or low quality data.

Turkish language is spoken by nearly 90 millions of people. However, the parallel corpora having Turkish language (e.g. Turkish to English translations) in OPUS CORPUS, the biggest open parallel corpora on the internet, is very limited compared to other data-rich languages. And the available Turkish corpora is mostly obtained from volunteer or crowdsourced translation

[1] m.tradumatica.net. See Martin-Mor (2017) for the details of the project.

projects such Open Subtitle and Wikipedia. For achieving higher scores in machine translation quality evaluations, a Turkish to English machine translation engine will need domain-specific high quality parallel corpora. For example, it will be more convenient to create a medical domain English into French machine translation engine not only because these languages are grammatically similar but also because of the large amount of open parallel corpora available on the internet. By using open tools and state-of-the-art CAT tool features, we achieved to create parallel corpora of Turkish to English cardiology ready for machine translation training. We have selected cardiology because it is a very specific area with its own terminology and textual conventions.

2. Objective

The objective of this study is to show the methods that we have used in order to gather and prepare medical parallel corpora for the purpose of machine translation training. In this study, we report the automatic and semi-automatic methods we use for creating domain-specific (medical) custom translation memories as well as bilingual terminology lists which include web-crawling, document alignment in CAT tools and term extraction. We have crawled the web and obtained 6500 Turkish cardiology abstracts and their human translations into English using HTTrack, then we have aligned these abstracts using LiveDocs feature of Memoq and have created a translation memory of 1.200.000 words, of which we have extracted terms to be used both in the MT training and evaluation steps. This parallel corpora will be used in NMT and SMT training in the future, and evaluation of the quality of the engines through terminological quality. Using this method, it is possible to prepare large parallel corpora for language pairs lacking free and open training data.

3. Tools&Methods

Created in computer-aided translation (CAT) tools, translation memories are the most common parallel corpora types and constitute the most valuable data for machine translation training. They are exchanged between different CAT tools in TMX format which makes that highly reusable both in other translation projects and in machine translation or text extraction projects (and in other corpus analysis projects). However, most of the translation memories created in the industry are proprietary and are not open to be used by third parties. For this reason, finding open domain-specific parallel documents and process them to create translation memories will be beneficial for many shareholders including freelance translators, machine translation practioners as well as academic researchers.

Web crawling is an important step in collecting parallel corpora. There are many multilingual websites with open textual data. Nevertheless, although these websites may have a well-structured form, it is generally hard to first crawl them and then align their multilingual/bilingual content (say, English and Turkish) automatically. Bitextor[2] is a free and open automatic bitext generator which yields a TMX file after crawling and aligning a bilingual website. However, it only works with a few languages and does not include Turkish-English language pair. Besides, the quality of the automatic alignments is not enough to be used practically. For these reasons, we did not include this tool in our study. Yet, we think that the development of similar tools and the addition of an interface to be able to correct/validate alignment pairs will accelerate a lot the process of parallel corpora preparation.

The first step of our study is the selection of the data. Since corpus based machine translation systems function better with more specific data, we have focused on cardiology abstracts. We have used a free and open-source website copier called HTTrack to crawl the website of Archives of the Turkish Society of Cardiology[3] which both in Turkish and English. Since these abstracts have been translated from Turkish to English directly (without the intermediacy of another language), revised for an academic journal and published in this journal, we have assumed high quality in translation. In this decision, we take into consideration the "publishable quality" concept suggested by TAUS[4] for postediting machine translation output. We think that this criteria can also be applied during the selection of translations for MT training. We obtained 6500 abstracts (in total, 13000) in the form of HTML as a result of the web crawling. We used regular expression filters and cascading filters Memoq CAT tool to extract only the relevant abstracts, and finally aligned them using the LiveDocs feature of the same tool. Although segment by segment alignment has been very successful most of the time, a human intervention has been needed to fix the misalignments. However, this intervention has been needed at a minimum level because the automatic alignment has been very successful and the alignment editing interface of LiveDocs is very user-friendly. Upon completing the alignment process, it is possible to export the aligned segments directly into translation memory. We have obtained a TMX file in the end, which needed one more step to be cleaned since some HTML tags remained in our translation memory. We have used Olifant which is a translation memory editor to clean up the HTML tags. But not all statistical machine translation customisation platforms accept TMX files. Hence, depending on the platform to be used, a further step may be needed. While KantanMT[5] and Microsoft Translator

[2] http://bitextor.sourceforge.net
[3] http://www.archivestsc.com
[4] www.taus.net. Translation Automation User Society
[5] kantanmt.com

Hub[6] support TMX files, MTradumàtica currently requires aligned source language file and target language file separately. Okapi Framerwork[7], especially Tikal, can convert TMX files into two separate language files ready to be used for training. With these steps, we have created a translation memory of 1.200.000 words. This amount of corpus is of course not sufficient to train a state-of-the-art machine translation engine but by finding more data and speeding up these process, we are planning to gather more data in the future.

There are also two ways of extracting terminology out of the aligned segments. Firstly, it is possible to run an automatic monolingual term extraction feature to search for the most frequent words and then add their target term counterpart and validate them. The second strategy is to select the source term and target term manually in the LiveDocs interface and save them as term. Although the second option seems to be time-consuming, in fact, it can be done very fast and a glossary/terminology list can be created in a short time. And then, this terminology/glossary list can be used in the statistical machine translation training and increase the terminological quality of the output.

4. Turkish-English Parallel Corpora in Cardiology Domain

In corpus-based approaches of machine translation, the more specific the training corpus domain, the better the translation output will be. As Wolk and Marasek highlights, "SMT systems should work best in specific, narrow text domains and will not perform well for a general usage" (Wolk and Marasek, 2015). Similar observation is made by Lumeras and Way (2017): "It is well-known that MT systems work best when tested on data that is very similar to the corpora on which they are trained. For example, it would be foolish to translate weather forecasts using an SMT system that had been trained on parliamentary texts. So, for a particular company, it might not make sense to talk about their single English-to-French engine, but rather a whole suite of engines for this language pair depending on the domain (or in the field of Language for Specialised Purposes, what is known as "text genre"), e.g. patents, trials, white papers, personnel texts, product documentation, legal texts, contracts etc". However, in this point, we come across a paradox: although we will be more likely to create a better engine when we have a specific domain, the more our domain is specific, the less amount of text we will likely have. This is especially true for low-resource languages like Turkish. If we would like to create high-quality custom engines in the future, we will need to create both sufficient amount of domain-specific open parallel corpora and tools and methods to create this corpora. Since professional translators and translation companies continue to translate through their CAT tools, which, in turn, leads to the accumulation of new translation memories (which constitute the most widely used parallel corpora). Yet, these translation memories are proprietary and it is very unlikely that they will be made public due to the privacy agreements between translators/translation companies and clients. Hence, academic studies should be made to create (from openly available data) the relevant parallel corpora for different text genres for low-resource languages so that new corpus-based MT systems (SMT and NMT) can be studied in academic fields or implemented in industrial environments.

There are different text types (domains, as MT practitioners call them) with different textual conventions and terminologies. Since 1950s, many different text typologies have been suggested, one of the most well-known being Reiss and Vermeer's (2004) three-fold text type distinction where they classify texts as informative, operative and expressive texts. We believe that an understanding and use of text typology in the process of preparing parallel corpora are essential and are the added-values to be brought by the translators to the development of machine translation systems.

With this in mind, we have kept our focus very narrow in our study. Instead of concentrating on a more general text domain like medical domain, we focus on cardiology domain which is an area of expertise with its own terminology. We have crawled only Turkish and English abstracts in a scientific cardiology journal. Scientific abstracts tend to have their own text conventions including word choice, character or word limitations, and (and when applicable just like the case of medical articles) conventional headings such as "objective", "methodology", "results" and "conclusion".

We can also resort to another classification to understand better our data. Montalt (2010) groups medical genres under four classes in his social function oriented distinction: research, professional, educational and commercial. This classification is helpful in the parallel corpora preparation phase since it constrains the text type selection, and as mentioned above, the more the data is of a specific domain, the better the statistical machine translation engine is. Cardiology articles and abstracts are research-based medical texts and their translations. These are "those used by researchers to communicate their findings and arguments: original articles, case reports, doctoral theses, etc." (Montalt, 2010). Note that in such a MT engine preparation process, the work begins by selecting the specific text domain. The next step will be to search, crawl and collect open resources considering the amount and type of the domain.

5. Conclusion

Most of the studies concentrate on the evaluation of machine translation systems as well as procedures for postediting. We think that concentrating on the parallel corpora selection, collection and preparation processes is equally important and may have a positive impact on the later processes. And this is where translators can prove

[6] hub.microsofttranslator.com

[7] http://okapiframework.org

their added-value to the partial automation of translation process. The further creation of free and open parallel corpora will also make it possible for freelance translators and small language service providers to be able to use open source machine translation platforms such as MTradumatica, and to customize their own engines without paying large amounts of money. Hence, they will be able to compete in respective areas.

Another important aspect of optimizing the linguistic data (parallel corpora in our case) preparation is to help minority languages which have scarce data or no data to be able to make use of the data-driven approaches such neural machine translation and statistical machine translation, translation technologies and corpus technologies which, all together, have a potential to empower these languages.

Acknowledgements

This work has been supported by the FI-DGR-2017 grant cofinanced by European Social Fund and Generalitat de Catalunya.

6. Bibliography

1. The Opus Corpus. http://opus.nlpl.eu [last access : 05.02.2018]
2. Olifant. Okapi Translation Memory Editor. http://okapi.sourceforge.net/Release/Olifant/Help/ [last access : 05.02.2018]
3. Memoq. http://memoq.com [last access : 05.02.2018]
4. MTradumatica. http://m.tradumatica.net [last access : 05.02.2018]
5. Martín-Mor, Adrià (2017) MTradumàtica: Statistical Machine Translation Customisation for Translators. *Skase journal of translation and interpretation*, 11, p. 25-40
6. Forcada, Mikel L. (2017) Making sense of neural machine translation. *Translation Spaces*, 6, p. 291-309
7. Wołk, K. and Marasek, K. (2015) Polish - English Statistical Machine Translation of Medical Texts. In Aleksander Zgrzywa, A., Choroś, K., Siemiński, A. [eds.] *New Research in Multimedia and Internet Systems*. pp. 169-179.
8. Reiss, K. and Vermeer, H. (2004) *Towards a General Theory of Translational Action. Skopos Theory Explained*. Routledge.
9. Montalt, V. (2010) "Medical translation and interpreting". In Gambier, Y. and Doorslaer, L. [eds.] Handbook of Translation Studies. Vol. 2. pp. 79-83.
10. TAUS Postediting Guidelines. https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines (last access: 06.03.2018)