M1.4 – MT Approaches towards Cross-lingual IR in Pubpsych

Cristina España-Bonet
Universität des Saarlandes

– v1.0 –
August 2018

Abstract

This document describes the architecture options and final choices for implementing the machine translation (MT) system aimed to translate articles' titles and abstracts. The alternatives are presented and a comparison among the two most promising architectures, Statistical MT and Neural MT, is given.

Contents

1 Introduction

The main purpose of the project is to provide PubPsych with a Cross Language Information Retrieval (CLIR) functionality to allow users to specify their information needs in their preferred language while retrieving relevant documents matching their needs in languages different from the query language.

One of the possible approaches to CLIR is Machine Translation (MT). One can either translate the query or translate the text:

(*i*) Translating the query: When a search expression is entered, the portal extends the query by injecting translations of the terms or phrases found in it, and runs the modified version against the search index

(*ii*) Translating the metadata: Before indexing, some or all of the content (e.g. keywords, titles & abstracts) of the portal is translated, so that all query languages are represented in the index

PubPsych offers multilingual content in English, French, German and Spanish. All its metadata is in one or more of the four languages. The original works, such as articles, book chapters, data sets, might be in many different languages (more than 50 in PubPsych), but the metadata are always available in at least either English, French, German or Spanish. So, both in Scenario (*i*) and (*ii*) the translation needs to be done among the four languages. Again, there are two possibilities:

(*i*) Training in the 12 translation directions to be able to perform any combination,

(*ii*) Using English as a pivot language

Choosing the best option depends on the nature of the translation system. To give an example, an MT system for translating the abstracts of our documents can be built either with statistical or neural architectures. Statistical systems would offer a higher quality when using English as pivot, as we do not have parallel corpora for the six language pairs and the translation from German-to-Spanish would be done via English anyway. Besides, the number of systems to maintain diminishes from 12 to 3. On the other hand, neural systems are able to learn to translate among all the languages even when there is no direct parallel data for a language pair (with a lower quality though). In this case, a single multilingual neural system is enough to translate among the 4 languages. More details are given in Section 2.1.

2 The CLuBS Translation Proposal

When the proposal for the project was written, the state of the art for machine translation was Statistical Machine Translation systems (SMT). At the time, SMT was envisaged as the main architecture for translating parts of the metadata (titles and abstracts). For translating queries and keywords, multilingual thesauri and controlled vocabularies were chosen as the source input is not made by complete sentences.

During the course of the project, we have entered the boom of deep learning, especially for MT. Other fields benefited from deep learning before, but for MT, Neural Machine Translation (NMT) became state-of-the-art in 2016. To take this development into account, we have decided to follow both approaches and chose which architecture to integrate after its evaluation on the retrieval performance. Notice that the translation quality does not need to be related to the final retrieval quality, and a translation system with a high

	RBMT	SMT	NMT
Data Amount	small	large	large
Training Time	–	days	weeks
CPU/GPU	CPU	CPU	GPU
Cost	expensive (in people)	cheap	expensive (in hardware)
Maintainability	weak	strong	superstrong
Grammaticality	strong	medium	strong
Reordering	strong	weak	strong
Consistency	strong	medium	weak
Coverage	weak	strong	weak
Multilinguality	medium	none	strong

Table 1: Comparison of the characteristics of the main kinds of translation engines: rule-based (RMBT), statistical (SMT) and neural (NMT). The top rows show the characteristics to be taken into account for deployment and the bottom rows the quality achieved for different linguistic issues.

adequacy could be better for retrieval than a system with a very high fluency, which is usually preferred by humans.

Section 2.1 summarises the main differences between translation architectures for titles and abstracts. For queries and keywords, we sketch the translation proposal via mapping approaches in deliverable M1.5.

2.1 Neural Machine Translation vs. other Systems

The quality of neural systems is currently superior to other translation systems for language pairs with large amounts of parallel data. NMT is specially better than SMT in fluency which makes its output more appealing to humans. The decoder side of an NMT system is basically a language model, so, by construction, good fluency is expected. However, NMT shows inconvenient characteristics regarding adequacy. Since embeddings, that is, vectorial representation of words, take care of the alignments and similarities, synonyms are more likely to appear, but not only synonyms, any kind of related word. So, it is not strange that the sentence "I have 72 books in my library." is translated as "Ich habe 79 Bücher in meiner Bibliothek." as the embedding for 72 and 79 will be very similar. NMT systems also create and delete words at will, there is not real control of the number of words needed as it was done by word and phrase penalties in SMT. One could say that an SMT system performs literal translation, while NMT performs free —kind of artistic— translation.

SMT systems allow more control because they follow a pipeline of processes, whereas NMT systems are end-to-end architectures. This feature has pros and cons at the same time. One can improve the output of a module before feeding the next one (e.g. discarding low frequent alignments before phrase extraction or pruning a phrase table before decoding) but also errors in one module are propagated into the others. In standard NMT, error propagation cannot occur but neither can the improvement of specific processes. Even with these problems, NMT systems are nowadays state-of-the-art at least for resource-rich language pairs and trigger new functionalities such as multilinguality and zero-shot translation, that is, translation for language pairs not directly seen in training

but included in other pairs. In the CLuBS project, we have in-doimain parallel data for French–English and German–English but not for French–German for instance, and we could approach that language pair as a zero-shot one. It is worth then trying to adapt the basic architecture to tackle its specific drawbacks.

Regarding the best system for deployment, SMT systems are easier and faster to train and all the process can be done in commodity hardware. On the other hand, NMT systems need at least a week of training time using a GPU. In both cases decoding can be done using CPUs with competitive speeds. As said before, an additional advantage of NMT systems for CLuBS is maintainability, since a single system can be used to translate among the 12 translation directions. Adding new data via transfer learning is another advantage of NMT that would allow the improvement of the translator in our domain as PubPsych adds new documents with parallel content.

Table 1 summarises the pros and cons of each kind of architecture. The top rows show the characteristics to be taken into account for deployment and the bottom rows the quality achieved for different linguistic issues. We have included Rule-based Machine Translation systems (RBMT) for completeness, even if these systems are not adequate for CLuBS due to the languages involved and the amount of time required to develop an in-domain engine. We can see that NMT is easy to mantain (mantainability) because there is a single model for all the language pairs. Besides, new training data can be added just by continuing the training even if for that one needs GPUs. Translations with NMT are state-of-the-art showing translations with a good grammar (Grammaticality) and very few reordering mistakes (Reordering). Coherence and consistence at document level (Consistency) are still an issue though currently better resolved by SMT systems.

2.2 Neural Machine Translation for CLuBS

There are several aspects of NMT systems that should be studied during the course of the project in order to approach specific necessities of our setting. As NMT is a rapidly evolving field, this implies research that can be divided into two main categories:

Architecture. We have at our disposal multilingual thesauri such as MeSH for the domain of psychology. But in NMT, adding external knowledge and forcing translation of a given phrase is not possible in standard systems. We will explore how to implement these features. Two non-trivial topics are approached here: the addition of external resources during training and the transfer of information from source to target.

Data. High quality in-domain parallel data is indispensable for successfully training a system but we do not have data on psychology for all the needed language pairs. We therefore aim to explore the gathering of this data beforehand, and also how to perform an internal auto-data cleaning during training as this would allow to successfully use crawled (low quality) parallel sentences as appropriate.

3 Conclusions

Several approaches can be followed to develop a machine translation engine. CLuBS will focus on variants of SMT and NMT engines, whose characteristics best suit our settings. The final prototype will implement the best performing system in our IR evaluation but, during the project, we will implement at least a variant of the two main architectures. Due to the novelty, higher performance and room for improvement, research will be carried out for NMT and this will provide CLuBS with several system variants.