Machine Translation Techniques

for

Advanced Natural Language Processing (968G5): Assessed Coursework 2

Prepared by Andreas Kyratzis

SCHOOL OF MATHEMATICAL AND PHYSICAL SCIENCES
University of Sussex

May 2019

Machine translation (MT) is considered as the use of computer's software to automatically translate natural languages without human interaction. Correct use of machine translation requires the meaning of the text translated not being altered after the translation. The objective of this computer-aided procedure is to breach the language gap of people without the need of a translator being present. By correct use of language translation, people across the world can also use literature or any kind of information from across the world through the world wide web. While people across the world today can speak and understand more language than ever before individually; there is still a large gap between people due to the number of languages that exist. Therefore, machine translation is the closest thing to an individual's understanding of different languages or the context of the text or speech that the individual needs to translate.

There are several different techniques being used for machine translation but the most commonly used are Rule-Based Machine Translation (RBMT) and Statistical-Based Machine Translation (SBMT). The first relies on rules as it name states and its introduction was numerous decades ago. It is the pioneer in machine translation and has been used for a vast number of translations over the years. Technically, this approach works by taking the initial text's sentences once at a time and trying to break the sentence down into words while also analysing the structure of the sentence. After this initial analysis, the system converts the text to the specified language by using rules of language translation linguistic specialists. The rules which are used are executing similarity checks from one language to the other in order to find patterns which could enable easier translation techniques while also using standard rules of grammar.

The latter of the two techniques analysed in this report, Statistical-Based Machine Translation, uses corpora of multiple languages' text and the translation of those text sources while also using large amounts of text from various languages. This technique is based on the statistical connections and correlations of the text being translated and the corpora in order to build a model which can translate from one language to the other. This technique not only searches for statistical connections on document level and corpora text but also on small phrases and corpora text which enables it to gain a better understanding of several different meanings of word sequences. As this technique doesn't use rules to translate text, it provides a mapping of how likely the translation is correct in the language translated to. SBMT is used today by services such as Google Translate and Bing Translator. SBMT as analysed by Brown et al., (1990) when using sentence T, T derived from the target language, the aim is to find sentence S which S derived from source language is the source used to translate the text. By choosing the most plausible sentence S, the chances of error of translation are diminished and therefore the probability of getting the correct translation is maximised.

Taking on a critical evaluation of the two techniques used for MT, looking at it from a performance-based perspective; the technique used should be fast and require a small amount of computer memory. Consequently, RBMT, which uses rules to translate text, has higher performance and is more robust as rules used do not use as much disk storage as SBMT due to not being a corpora-based approach. On the other hand, the development of the system is equally important as performance. Thus, creating a system for a given pair of languages requires a certain amount of labour and time. Since RBMT uses rules by various well-respected linguistic experts, it requires a lot of hard human labour and more importantly is very time-consuming. SBMT however only uses various corpora of text from translations of text on the pair of languages that are translated therefore getting a hold of the material that is going to be used is the only part of development that requires labour but still way less intensive labour than for RBMT.

Taking on quality of translation as a part of the evaluation, RBMT provides reliable and expected quality translations as grammatical rules are used as part of the translation procedure. As SBMT uses probabilistic approaches to translate natural languages, the results of the translation are often erratic and quality is not always a given as probabilistic approaches provide a hit or miss result when considering natural languages.

Quality of translation is not only a factor of providing good translations for a given domain of language though. When it comes down to creating a system that is able to handle different domains of language (e.g. financial papers and interviews of sports athletes), RBMT clearly can handle the translation with more ease as rules apply to most domains and are likely to get quality results given any domain. However, SBMT produces inferior quality translations when is out of the domain used in the corpora collection it uses.

As natural languages' forms and rules rarely change, RBMT mostly produces regularly at a high standard while when considering SBMT and its' uses of corpora collections, when the collection changes, so will the results of the translations as the corpora collection is the only thing that in SBMT can translate natural languages.

As fluency is a big factor of translation, seeing the difference between RBMT and SBMT when considering the better translation technique is essential. By using rules, which may be incorrect for certain source language, fluency is not something that can be achieved with RBMT to a high degree while SBMT due to its use of corpora collections from source and target languages, provides great fluency to the target language. Adding to the above statement, rules can only go so far with natural language understanding; meaning that by not adding all kinds of linguistic structures of languages into the rules of an RBMT, translation can be out of context and not worthy of use. This can relate with slang language and how slang language doesn't use most rules that are used in regular natural language; same can be said for any use of metaphorical terms as their meaning may be misleading. With ambiguities such as the above, SBMT has a ruling power over RBMT as it lacks rules, therefore, all exceptions to rules do not matter to the model being used.

SBMT is cost-effective to maintain as publicly available text has seen a rise due to the internet age while RBMT requires fine-tuning by linguistic experts and thus is not cost-effective. SBMT doesn't need any more adjusting to do when forming new pairs of languages to translate apart from getting a hold of the corpora of text that is to be used and is therefore the technique used for most not common translation languages pairs as text is publicly available while rules aren't as easy to come across to or to write them easily. Building a corpora of languages today is easier than it used to be a few decades ago but still getting hold of bilingual corpora text can be a resource-intensive strategy as domain is a concern when doing any kind of domain-specific translation which could have different meanings for certain words or phrases than any other part of language domain use.

Considering all of the above statements about each of the two main techniques of Machine Translation, as (Kantan MT, 2014) states, SBMT has grown in popularity over the last few years due to the advances of technology and the use of higher calibre computers for such procedures. Adding on to the verdict of above literature; as the data that is being used to build the corpora of SBMT is more accessible today, the application of MT through this technique becomes a more viable option. Another article which focuses on using both techniques on getting better English to Indian translations and Indian to Indian translations (Sreelekha, 2019), has uncovered that the quality of SBMT translations is higher compared to translations based on RBMT. The paper also acknowledges that RBMT systems require a lot of resources and labour. Another discovery outlined in the paper is that due to RBMT having a morph analyser, it was able to successfully translate words that had inflected suffixes whereas SBMT failed top translate those words correctly.

Costa-Jussa et al. (2012) analyses RBMT and SBMT on Spanish-Catalan translations and evaluates both methods using services that provide the two techniques. Given certain tests and evaluation techniques, they have found that SBMT provides better semantic level performance and RBMT provides translations with fewer orthographical and morphological errors than of SBMT. They also found that considering a fixed domain of natural language, results can vary as a certain RBMT system may have an expertise of rules in that domain or an SBMT system having a corpora of that domain-specific natural language. When systems are trained on a particular domain or have rules specifying the language used in the domain, they achieve better predictions.

As the most commonly used MT system, Google Translate; which uses SBMT, is being trained on an enormous amount of data from all different kinds of domains, can achieve very good results considering the amount of trained data of that domain in that language. The correlation between trained data and prediction scores is very high for any SBMT system as the results vary based on how much trained data has been used.

To evaluate the above literature, I believe the most obvious choice would be to use SBMT out of the two as it is being treated with a large amount of data to work with meaning that the prediction scores of correct translations are higher than of RBMT. As the likelihood of achieving a correct translation using RBMT is solely based on rules that may not be up to date or may not provide translations for certain parts of language use such as slang which has seen a rise in use in the last few years due to the internet era; I stand by using SBMT for any language pair translation that has a big corpora of data in that language pair.

As SBMT is not the most novel technique due to some certain difficulties it has with translations, a Hybrid machine translation (HMT) technique has been developed which uses two or more different techniques of MT to increase the level of accuracy on providing the correct translations. As RBMT are better at keeping translations orthographically and morphologically correct while SBMT is better at keeping fluency and semantic level performance high; it is a more novel approach to use both techniques to achieve the higher level of accurate predictions. Such systems exist and bring together the better characteristics of each MT technique. An example of hybridisation of MT techniques uses corpora to build an RBMT system which in return reduces the need of linguistic experts while becoming cost and time effective (Habash, Dorr and Monz, 2009; Costa-jussà and

Fonollosa, 2015). Most HMT systems are based on either of the two techniques analysed and discussed in this paper (RBMT & SBMT) and use in parallel techniques which are better at dealing with what the basic technique cannot do to a high degree due to its restrictions. Such systems (HMT) have proved to be working better on speech translation and bilingual information retrieval which are tasks that the two pre-discussed techniques have not had exceptional translations in. All in all, as natural languages are languages of humans, it is extremely difficult for a computer to achieve what a human can achieve when doing a translation of a natural languages pair. Advances in technology are what is needed for the task of MT being a precise tool to use for natural language translations.

Bibliography

[1] Andovar. (n.d.). Machine Translation - Andovar. [online] Available at: https://www.andovar.com/machine-translation/ [Accessed 3 May 2019].

[2] Brown, P., Cocke, J., Pietra, S., Pietra, V., Jelinek, F., Lafferty, J., Mercer, R. and Roossin, P. (1990). A STATISTICAL APPROACH TO MACHINE TRANSLATION. [online] Available at: http://delivery.acm.org/10.1145/100000/92860/p79-brown.pdf?ip=5.64.183.22&id=92860&acc=OPEN&key=4D4702B0C3E38B35

[3] Costa-jussà, M. and Fonollosa, J. (2015). Latest trends in hybrid machine translation and its applications. Computer Speech & Language, 32(1).

[4] Costa-Jussa, M., Farrus, M., Marino, J. and Fonollosa, J. (2012). STUDY AND COMPARISON OF RULE-BASED AND STATISTICAL CATALAN-SPANISH MACHINE TRANSLATION SYSTEMS.

[5] Garr, B. (2017). Why is Machine Translation Soooo Hard?. [Blog] Medium. Available at: https://medium.com/@brianlgarr/why-is-machine-translation-soooo-hard-a0a4983084fd [Accessed 5 May 2019].

[6] Habash, N., Dorr, B. and Monz, C. (2009). Symbolic-to-statistical hybridization: extending generation-heavy machine translation. Machine Translation, 23(1), pp.23-63.

[7] Kantan MT (2014). RBMT vs SMT. [Blog] Kantan MT Blog. Available at: https://kantanmtblog.com/2014/02/13/rbmt-vs-smt/ [Accessed 3 May 2019].

[8] Sreelekha, S. (2019). Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective. [online] Available at: https://arxiv.org/ftp/arxiv/papers/1708/1708.04559.pdf [Accessed 4 May 2019].

[9] Weeds, J. (2019). Machine Translation 1-2.