ROOTINDEXING
JOURNAL ABSTRACTING AND INDEXING SERVICE
INTERNATIONAL ADVANCE JOURNAL OF ENGINEERING, SCIENCE AND MANAGEMENT (IAJESM)
January-June 2023, Submitted in January 2023, iajesm2014@gmail.com, ISSN -2393-8048
Multidisciplinary Indexed/Peer Reviewed Journal. SJIF Impact Factor 2023 =6.753

Machine Translation and Assorted Aspects with Real Time Applications

[1]BILAL AHMED, Research Scholar, Department of Computer Science, Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu (Rajasthan), Email- bilal.ahmed.jjn@gmail.com
[2]Dr Abdul Jabbar Khilji, Co Guide, Department of Computer Applications, Govt Engineering College, (Bikaner), Email- dr.jabbarkhilji@gmail.com
[3]Dr. Prasadu Peddi, (Research Guide)Department of Computer Science and Engineering, Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu (Rajasthan), Email: peddiprasad37@gmail.com

Abstract

Machine translation refers to the sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. On a basic level, MT performs mechanical substitution of words in one language for words in another, but that alone rarely produces a good translation because recognition of whole phrases and their closest counterparts in the target language is needed. Not all words in one language have equivalent words in another language, and many words have more than one meaning. Solving this problem with corpus statistical and neural techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies. Current machine translation software often allows for customization by domain or profession (such as weather reports), improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used. It follows that machine translation of government and legal documents more readily produces usable output than machine translation of conversation or less standardised text. Improved output quality can also be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously identified which words in the text are proper names. With the assistance of these techniques, MT has proven useful as a tool to assist human translators and, in a very limited number of cases, can even produce output that can be used as is (e.g., weather reports). The progress and potential of machine translation have been much debated through its history. Since the 1950s, a number of scholars, first and most notably Yehoshua Bar-Hillel, have questioned the possibility of achieving fully automatic machine translation of high quality.

Keywords: Machine Translation, Machine Translation and Use Cases, Machine Translation and Scenarios

Introduction

Machine translation can use a method based on linguistic rules, which means that words will be translated in a linguistic way – the most suitable (orally speaking) words of the target language will replace the ones in the source language. It is often argued that the success of machine translation requires the problem of natural language understanding to be solved first.

Generally, rule-based methods parse a text, usually creating an intermediary, symbolic representation, from which the text in the target language is generated. According to the nature of the intermediary representation, an approach is described as interlingual machine translation or transfer-based machine translation. These methods require extensive lexicons with morphological, syntactic, and semantic information, and large sets of rules [1].

Given enough data, machine translation programs often work well enough for a native speaker of one language to get the approximate meaning of what is written by the other native speaker. The difficulty is getting enough data of the right kind to support the particular method. For example, the large multilingual corpus of data needed for statistical methods to work is not

necessary for the grammar-based methods. But then, the grammar methods need a skilled linguist to carefully design the grammar that they use.

To translate between closely related languages, the technique referred to as rule-based machine translation may be used.

Rule-based

The rule-based machine translation paradigm includes transfer-based machine translation, interlingual machine translation and dictionary-based machine translation paradigms. This type of translation is used mostly in the creation of dictionaries and grammar programs. Unlike other methods, RBMT involves more information about the linguistics of the source and target languages, using the morphological and syntactic rules and semantic analysis of both languages. The basic approach involves linking the structure of the input sentence with the structure of the output sentence using a parser and an analyzer for the source language, a generator for the target language, and a transfer lexicon for the actual translation. RBMT's biggest downfall is that everything must be made explicit: orthographical variation and erroneous input must be made part of the source language analyser in order to cope with it, and lexical selection rules must be written for all instances of ambiguity. Adapting to new domains in itself is not that hard, as the core grammar is the same across domains, and the domain-specific adjustment is limited to lexical selection adjustment [2].

Figure 1 : Neural Machine Translation

Transfer-based machine translation

Transfer-based machine translation is similar to interlingual machine translation in that it creates a translation from an intermediate representation that simulates the meaning of the original sentence. Unlike interlingual MT, it depends partially on the language pair involved in the translation.

Interlingual

Interlingual machine translation is one instance of rule-based machine-translation approaches. In this approach, the source language, i.e. the text to be translated, is transformed into an interlingual language, i.e. a "language neutral" representation that is independent of any language. The target language is then generated out of the interlingua. One of the major advantages of this system is that the interlingua becomes more valuable as the number of target languages it can be turned into increases. However, the only interlingual machine translation system that has been made operational at the commercial level is the KANT system (Nyberg and Mitamura, 1992), which is designed to translate Caterpillar Technical English (CTE) into other languages [3].

Dictionary-based

Machine translation can use a method based on dictionary entries, which means that the words will be translated as they are by a dictionary.

Statistical

Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora, such as the Canadian Hansard corpus, the English-French record of the Canadian parliament and EUROPARL, the record of the European Parliament. Where such corpora are available, good results can be achieved translating similar texts, but such corpora are still rare for many language pairs. The first statistical machine translation software was CANDIDE from IBM. Google used SYSTRAN for several years, but switched to a statistical translation method in October 2007. In 2005, Google improved its internal translation capabilities by using approximately 200 billion words from United Nations materials to train their system; translation accuracy improved. Google Translate and similar statistical translation programs work by detecting patterns in hundreds of millions of documents that have previously been translated by humans and making intelligent guesses based on the findings. Generally, the more human-translated documents available in a given language, the more likely it is that the translation will be of good quality [4, 5, 6]. Newer approaches into Statistical Machine translation such as METIS II and PRESEMT use minimal corpus size and instead focus on derivation of syntactic structure through pattern recognition. With further development, this may allow statistical machine translation to operate off of a monolingual text corpus. SMT's biggest downfall includes it being dependent upon huge amounts of parallel texts, its problems with morphology-rich languages (especially with translating into such languages), and its inability to correct singleton errors [6, 7].

Example-based

Example-based machine translation (EBMT) approach was proposed by Makoto Nagao in 1984. Example-based machine translation is based on the idea of analogy. In this approach, the corpus that is used is one that contains texts that have already been translated. Given a sentence that is to be translated, sentences from this corpus are selected that contain similar sub-sentential components. The similar sentences are then used to translate the sub-sentential components of the original sentence into the target language, and these phrases are put together to form a complete translation [7, 8].

Hybrid MT

Hybrid machine translation (HMT) leverages the strengths of statistical and rule-based translation methodologies. Several MT organizations claim a hybrid approach that uses both rules and statistics. The approaches differ in a number of ways:

Rules post-processed by statistics: Translations are performed using a rules based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine. Statistics guided by rules: Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalization. This approach has a lot more power, flexibility and control when translating. It also provides extensive control over the way in which the content is processed during both pre-translation (e.g. markup of content and non-translatable terms) and post-translation (e.g. post translation corrections and adjustments) [8, 9, 10].

More recently, with the advent of Neural MT, a new version of hybrid machine translation is emerging that combines the benefits of rules, statistical and neural machine translation. The approach allows benefitting from pre- and post-processing in a rule guided workflow as well as benefitting from NMT and SMT. The downside is the inherent complexity which makes the approach suitable only for specific use cases [10, 11].

Neural MT

A deep learning-based approach to MT, neural machine translation has made rapid progress in recent years, and Google has announced its translation services are now using this technology in preference over its previous statistical methods.

To address the idiomatic phrase translation, multi-word expressions, and low-frequency words (also called OOV, or out-of-vocabulary word translation), language-focused linguistic features have been explored in state-of-the-art neural machine translation (NMT) models. For instance, the Chinese character decompositions into radicals and strokes have proven to be helpful for translating multi-word expressions in NMT.

Conclusion

Machine translation is the process of automatically translating content from one language (the source) to another (the target) without any human input. Translation was one of the first applications of computing power, starting in the 1950s. Unfortunately, the complexity of the task was far higher than early computer scientists' estimates, requiring enormous data processing power and storage far beyond the capabilities of early machines. It was only in the early 2000s that the software, data, and required hardware became capable of doing basic machine translation. Early developers used statistical databases of languages to "teach" computers to translate text. Training these machines involved a lot of manual labor, and each added language required starting over with the development for that language.

References

[1] Budiansky, Stephen (December 1998). "Lost in Translation". Atlantic Magazine. pp. 81–84.

[2] Albat, Thomas Fritz. "Systems and Methods for Automatically Estimating a Translation Time." US Patent 0185235, 19 July 2012.

[3] Bar-Hillel, Yehoshua (1964). Language and Information: Selected Essays on Their Theory and Application. Reading, Massachusetts: Addison-Wesley. pp. 174–179.

[4] Madsen, Mathias Winther (2009). The Limits of Machine Translation (MA thesis). University of Copenhagen. p. 5. Archived from the original on 17 October 2021.

[5] J. Hutchins (2000). "Warren Weaver and the launching of MT". Early Years in Machine Translation (PDF). Semantic Scholar. Studies in the History of the Language Sciences. Vol. 97. p. 17. doi:10.1075/sihols.97.05hut. ISBN 978-90-272-4586-1. S2CID 163460375. Archived from the original (PDF) on 28 February 2020.

[6] Wolfgang Saxon (28 July 1995). "David G. Hays, 66, a Developer Of Language Study by Computer". The New York Times. Archived from the original on 7 February 2020. 7 August 2020. wrote about computer-assisted language processing as early as 1957.. was project leader on computational linguistics at Rand from 1955 to 1968.

[7] Farwell, David; Gerber, Laurie; Hovy, Eduard (29 June 2003). Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas, AMTA'98, Langhorne, PA, USA, October 28–31, 1998 Proceedings. Berlin: Springer. p. 276. ISBN 3540652590.

[8] SPIEGEL ONLINE, Hamburg, Germany (13 September 2013). "Google Translate Has Ambitious Goals for Machine Translation". SPIEGEL ONLINE. Archived from the original on 14 September 2013. 13 September 2013.

[9] "Machine Translation Service". 5 August 2011. Archived from the original on 8 September 2013. 13 September 2013.

[10] Google Blog: The machines do the translating Archived 23 March 2006 at the Wayback Machine (by Franz Och)

[11] "Geer, David, "Statistical Translation Gains Respect", pp. 18 – 21, IEEE Computer, October 2005". Ieeexplore.ieee.org. 27 September 2011. doi:10.1109/MC.2005.353. S2CID 7088166.