International Journal of Advanced Technology in Engineering and Science
Volume No.06, Issue No. 04, April 2018
www.ijates.com
ijates
ISSN 2348 - 7550

A Short Overview to the Statistical Machine Translation

Dr. Amit Tuteja[1], Dr. Madhuchanda Rakshit[2]

[1,2]Guru Kashi University, Talwandi Sabo

Abstract

Statistical Machine Translation (SMT) is currently the most promising and widely studied paradigm in the broader field of Machine Translation, with researchers constantly exploring ways to improve its performance and find solutions to its current flaws, such as the scarcity of large bilingual corpora in a variety of domains or genres to be used as training data. One of the main trends is to rely as much as possible on already available large collections of data, sometimes when they do not fit specific translation tasks well in terms of content relatedness. The possibility of using fewer but more appropriately selected training sets, depending on the textual variety of the documents that need to be translated case by case - has not been explored as much as it should have been. SMT (Statistical Machine Translation) is a technique for translating statements from one language (for example, English) into another (such as Marathi). The source language is referred to as the source, whereas the target language is referred to as the target. This procedure may be described as a stochastic procedure. Depending on how translation is represented, there are a variety of SMT versions. Some methods employ a string-to-string mapping, others use trees-to-tree models, while yet others use tree-to-tree models. All of them share the core principle of automated translation, using models derived from parallel corpora (source-target pairs) as well as monolingual corpora (examples of target sentences). Background and Motivation Commercial, military, and political applications for machine translation are many. Non-English speakers, for example, are increasingly using the Internet and reading non-English pages. Machine learning advances, such as maximum-margin algorithms, are commonly used in translation studies. SMT systems have matured to the point where they can be used in production systems. Google's online English-Hindi translation, which is built on SMTtechniques, is an excellent example of this.

Key Words: *corpus, data, language, linguistics, multilingual, statistical machine translation, technology, text, variables, words, web.*

Introduction

The concept of developing Machine Translation (MT) systems was initially proposed over 60 years ago (Weaver, 1955), and it has had rapid expansion in recent decades, with MT systems being created both in academia and in the corporate sector, and becoming a widely used technology by consumers. Many individuals have speculated that MT would soon take over and replace human translation, saying that this will result in the loss of jobs for human translators. However, MT is far from becoming a completely functional, 100 percent dependable multifunctional technology, and it still requires some human interpretation of its output.

According to Koehn (2010, 20), the possibility of having fully automated high-quality machine translation is currently nothing more than a holy grail of MT, because fully automated MT systems have only been developed for a limited number of specific (and highly codified) communicative situations, such as weather forecasts, sports event summaries, and multinational company documentation. This means that in the vast majority of circumstances, fully automated translation may be challenging, if not impossible. Rather of attempting the near-impossible goal of developing a totally dependable all-purpose MT system, it may be possible to enhance MT performance by attacking the problem from different angles, such as the ability to conduct topic-specific MTtasks.

It is possible to create translation systems in Statistical Machine Translation (SMT) that provide a certain quantity of bilingual (and monolingual, in the target language) texts as training data to an SMT engine; however, in order to achieve good results for a specific SMT task, it is critical to use (and where possible select) those training data that are most suitable for the text(s) to be translated. The current tendency is to use massive amounts of parallel data to maximise the range of translation options (Bloodgood&Callison-Burch, 2010). In many circumstances, the majority of the data used is out-of-domain, and the translation performance is then tuned to specific translation jobs by adjusting SMT systems. Recent developments, however, have proven that relying on less quantities of meaningful training data is achievable. Given the textual diversity of specific texts to be translated, it may be beneficial to train MT systems on a

case-by-case basis, utilising tiny quantities of carefully selected training data. To learn how to choose the best data for each given translation circumstance, as well as if it makes sense to use much smaller training sets than are often used in SMT.

This method has two advantages: on the one hand, customised MT systems may provide better translations; on the other hand, utilising less but more concentrated data requires less time to train the SMT systems. Such operational gains would be extremely important when considering how the technique presented here may be implemented in real-world circumstances, such as the translation sector, where organisations may have limited time to complete certain translation projects.

Corpus

A collection of machine-readable written or spoken information gathered for the aim of linguistic study. (From the Oxford Dictionary)

A huge collection of texts of a certain type or on a particular topic.

(From the Oxford English Dictionary)

A computer-based collection of written or spoken information used to investigate how language is utilised (Cambridge Dictionary)

Multilingual Corpora

A corpus (plural corpora) or text corpus is a vast and organised series of texts in linguistics (nowadays usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, as well as to check for occurrences and validate linguistic rules within a certain language region.

Texts in a single language (monolingual corpus) or text data in numerous languages can be found in a corpus. (*multilingual corpus*).

Aligned parallel corpora are multilingual corpora that have been particularly structured for side-by-side comparison. Parallel corpora, which include texts in two languages, are divided into two categories. The texts in one language in a translation corpus are translations of texts in the other language. The texts in a similar corpus are of the same type and contain the same material, but they are not translations.

For analysis of a parallel text, some type of text alignment identifying comparable text

segments (sentences or paragraphs) is required. Machine translation algorithms for translating between two languages are frequently taught utilising parallel segments that include a first language corpus and a second language corpus that is an element-for-element translation of the first.

Additional organised layers of analysis have been applied to some corpora. A number of smaller corpora, in particular, can be fully parsed. Treebanks or Parsed Corpora are common names for such corpora. Because it is difficult to ensure that the entire corpus is annotated thoroughly and consistently, these corpora are often smaller, comprising one to three million words. Annotations or morphology, semantics, and pragmatics are all possible layers of language structured analysis.

In corpus linguistics, corpora are the major knowledge basis. In computational linguistics, voice recognition, and machine translation, corpora analysis and processing are frequently used to generate hidden Markov models for part of speech tagging and other reasons. Language training can benefit from corpora and frequency lists built from them. Corpora can be thought of as a type of foreign language writing aid because the contextualised grammatical knowledge acquired by non-native language users through exposure to authentic texts in corpora allows learners to grasp the manner of sentence formation in the target language, allowing them to write effectively.

Multilingual web-corpora

Parallel corpora that are freely available on the internet have several restrictions. On the other hand, a great number of bilingual/multilingual websites/pages in a range of language pairs, produced for various reasons, may be found on the Internet. They may be gathered and analysed, with plain text extracted and translated content aligned at the sentence level, to create new parallel corpora. However, compared to standard monolingual corpus collection from the online, there are additional challenges, namely in finding and pairing multilingual webpages, as well as the normal "web as corpus" concerns such as text quality, copyright issues, and so on. Several strategies for collecting parallel corpora from the web have been developed over the last 15 years, but the majority of them are not available to the public for a variety of reasons, including the use of now-deprecated technology, contractual constraints, and the authors' decision not to publicly release them.

Furthermore, the bulk of these contributions do not include extensive information on the genres/domains of the recovered parallel data, despite the fact that knowing the nature of prospective training data in terms of text kinds may be significant. Based on them, a system for collecting parallel corpora from the web has been set up for this project, resulting in the creation of two new corpora in a wide range of genres and domains, and their composition has been evaluated, providing an overview of the most common typologies of multilingual websites on the web for the considered languages.

Variables, Symbols and Operations used in SMT

Vsrc	source languagevocabulary
Vtgt	target languagevocabulary
e,eI 1	source sentence, i.e. a sequence of source language words
f, fJ 1	targetsentence, i.e. a sequence of target language words
a,aJ1	alignment sequence
t(f|e)	word-to-word translation probability, i.e. probability f is generated frome
t(f|e,c)	probability f is generated from e in contextc
a(j|i,I,J)	probability of emitting target word in position j from source word in position i under Models 1 and 2
a(i|i ′,I)	probability of moving from state i ′ to state i under HMMmodel
hm(e,f)	feature function for log-linearmodel
λm	featureweight
h	vector of featurefunctions
Λ	vector of featureweight

Advantages of SMT

1. SMT is better for User Generated Content and broad domain material such aspatents
2. SMT may translate softwaretags
3. SMT isn't expensive like Rule-based Translationsystem
4. SMT is better suited to on-the-fly translations of short-shelf-lifecontent
5. SMT will use the most likely term, but not necessarily the one youwanted

6. SMT is unpredictable but sentences are morefluid
7. SMT has longer updating cycles (once or twice a year istypical)
8. SMT can be free opensource
9. SMT is heavy on processingresources
10. SMT makes more fluidsentences
11. SMT can handle bad grammar, and doesn't improve much with controlledauthoring
12. SMT is the only choice for minoritylanguages
13. SMT and RBMT are matched for languages like French andSpanish
14. SMT can handle over 50 languages (Google, Bing andMicrosoft)
15. SMT may need millions of bilingual and monolingual segments but engines maybe pre-trained for a particulardomain

Shortcomings of SMT

1. Corpus creation can be costly.
2. Specific errors are hard to predict andfix.
3. Results may have superficial fluency that masks translationproblems.
4. Statistical machine translation usually works less well for language pairs withsignificantly different wordorder.

Google Translation for Research Scholars

Machine Translation is an excellent illustration of how Google brings together cutting-edge research with world-class infrastructure. Google is working to enhance statistical translation approaches so that they can better more data and be used to additional languages. The large-scale computer infrastructure enables translators and researchers to quickly test novel models trained on web-scale data in order to enhance translation quality.

This study supports the translations available at translate.google.com, which allows users to translate text, web pages, and even speech. Google Translate is a high-impact, research-driven tool that bridges the language barrier and allows you to explore the multilingualweb. It's available in a variety of Google services, including GMail, Books, Android, and web search. GoogleTranslate now supports 103 languages at various levels and serves over 200 million people daily as of August 2016. Google is pursuing human-quality translation and developing

machine translation systems for new languages, which presents exciting research challenges.

Conclusion

We've published an in-depth analysis of machine translation consistency, utilising cutting-edge SMT systems that were trained and assessed under a variety of realistic scenarios. Our research reveals a number of critical, yet often neglected, aspects of SMT consistency. First, even without explicit knowledge of extra-sentential context, SMT algorithms translate materials with remarkable consistency. They even have global consistency levels that are equivalent to those of human translators. Second, high translation consistency does not imply higher quality: as predicted in phrase-based SMT, lower-quality systems trained on smaller, more diverse data sets generate more consistent translations than higher-quality systems trained on larger, more heterogeneous data sets. This does not, however, mean that inconsistencies are desirable: inconsistently translated phrases are associated with translation mistakes far more frequently than consistent phrases. In practise, translation inconsistency might be used to identify difficult-to-translate words and phrases for a certain system. Finally, physical assessment of inconsistencies in translations reveals that just a tiny percentage of mistakes are the type of terminology issues that are the most common cause of worry in human translations. Instead, most faults indicated by inconsistencies are indicators of other disorders, such as faulty meaning translation and syntactic or stylistic flaws. These issues can occur with both consistent and inconsistent translations. While maintaining translation consistency in MT may be advantageous in some cases, our research shows that the phrase-based SMT systems under consideration would benefit more from addressing the underlying - and admittedly more complicated - problems of meaning and syntactic mistakes. We intend to enhance our analysis in the future by expanding our diagnosis algorithms and taking into account more data situations and genres. We also want to see how consistency may help with confidence estimate and error detection. In a nutshell, machine translation may be used for minor, non-critical tasks where a complete translation isn't required, but simply a broad understanding is required - for example, internal reasons. Human translation is crucial and much more dependable for key projects that will be seen by a worldwide audience to ensure a high-quality job and to ensure that the message you want to send is correctly understood by everyone.

References

[1]. Baroni, M., Chantree, F., Kilgarriff, A. &Sharoff, S. (2008).Biber,

[2].D. (1988). Variation acrossspeech and writing . Cambridge University Press.28

[3].Brown, P.F., Della Pietra, V.J., Della Pietra, S.A. & Mercer, R.L. (1993). The mathematics of statistical machine translation: parameter estimation. Comput.Linguist., 19, 263–311. 17

[4].Dasaradhi, K. Usage of Statistical Machine Translationin Textual Translation, Smart Moves Journal IJELLH, ISSN 2582-3574, Volume 6, Issue 7, July 2018, P. 1089 – 1100.

[5].Forsyth, R. &Holmes, D. (1996).Feature-finding for text classification. Literary and Linguistic Computing, 11,163–174. 32

[6].Koehn, P. (2010). Statistical machine translation . Cambridge University Press, Cambridge. 9, 16, 19, 21

[7].McCallum, A.K. (2002). MALLET: A Machine Learning for Language Toolkit. 39

[8].Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Comput. Surv., 34, 1–47.31

[9].Upton, G. & Cook, I. (2008). The Oxford Dictionary of Statistics . Oxford University Press, Oxford.67

[10]. Weaver, W. (1955).Translation. In W.N. Locke & A. Booth, eds., Machine Translation of Languages , MIT Press, Cambridge, MA, USA. 9, 16