STATISTICAL MACHINE TRANSLATION

Dr. Vijay Bhardwaj

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. The possibility of using fewer but appropriately selected training sets, depending on the textual variety of the documents that need to be translated case by case - has not been extensively explored as one of the main trends is to rely as much as possible on already available large collections of data, even when they do not fit quite well specific translation tasks in terms of relatedness of content.

Statistical Machine Translation (SMT) deals with automatically mapping sentences in one language (for example English) into another language (such as Marathi). The first language is called the source and the second language is called the target. This process can be thought of as a stochastic process. 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).Motivation and Background Machine Commercial, military, and political applications for translation are numerous. 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 SMT techniques, is an excellent example of this.

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

Introduction

The idea of building Machine Translation (MT) systems first emerged around 60 years ago (Weaver, 1955) and it has seen a remarkable growth during the last decades, when MT systems started being developed both in the academic field and in the private sector, becoming a widely employed technology by users as well. The emergence of MT even led many people to seriously consider that MT may soon take over and substitute human translation - even claiming that human translators would be left unemployed because of that. But MT is far from being a 100% reliable fully- functioning multipurpose technology, and it still requires a certain degree of human interpretation of its output.

As said by Koehn (2010, 20), the possibility of having fully-automatic high quality machine translation can be considered at the moment nothing more than a holy grail of MT, since so far it has been possible to develop fully-automatic MT systems only for a limited amount of specific (and of very codified) communicative situations, e.g. weather forecast, summaries of sports events, multinational companies documentation. This means that translation could be difficult, sometimes impossible, to be performed completely automatically in most cases. So, rather than aiming at the quite unfeasible target of building a fully reliable all-purpose MT system, it may be possible to improve the performances of MT approaching the problem from alternative points of view, like the possibility to carry out topic-specific MTtasks.

In Statistical Machine Translation (SMT) it is possible to create translation systems providing a certain quantity of bilingual (and monolingual, in the target language) texts as training data to an SMT engine, so in order to obtain good performances for a specific SMT task it is crucial to employ (and where possible select) those training data which are most suitable for the text(s) one wants to translate. The main trend is to employ large quantities of parallel data in order to maximise the coverage of translation possibilities (Bloodgood&Callison-Burch, 2010). In many cases most of the data employed may be out-of-domain and the translation performance is then adjusted tuning SMT systems towards specific translation tasks. 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.

The advantage of this approach is twofold: in one way, tailored MT systems may yield better translations; in another way, using less but focused data means fewer time to train the SMT systems themselves. Such operational benefits would be very valuable when thinking of possible implementations of the strategy here described in actual scenarios like the translation industry, where companies may have limited amounts of time to carry out specific translationtasks.

Corpus

A collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research. (Oxford Dictionary)

A large collection of writings of a specific kind or on a specific subject.(The Free Dictionary)

A collection of written or spoken material stored on a computer and used to find out how language is used (Cambridge Dictionary)

Multilingual Corpora

In linguistics, a corpus (plural *corpora*) or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.

A corpus may contain texts in a single language (*monolingual corpus*) or text data in multiple languages (*multilingual corpus*).

Multilingual corpora that have been specially formatted for side-by-side comparison are called *aligned parallel corpora*. There are two main types of parallel corpora which contain texts in two languages. In a *translation corpus*, the texts in one language are translations of texts in the other language. In a *comparable corpus*, the texts are of the samekind and cover the same content, but they are not translations of each other.

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.

Multilingual web-corpora

Freely accessible parallel corpora available on the web have some limitations. However on the Internet it is possible to find a large amount of bilingual/multilingual websites/pages in a variety of language pairs, published for disparate purposes. They can be collected and processed, extracting their plain text and aligning translated content at the sentence level, in order to build new parallel corpora. But there is a surplus of difficulties compared to the traditional monolingual corpus collection from the web, mainly concerning how to find and pair multilingual webpages, added to the usual “web as corpus” issues such as text quality, copyright matters etc. Several strategies to collectparallel corpora from the web have been developed during the last 15 years, but most of them are not available to the community for various reasons: some of them were based on now deprecated technology, contractual constraints, the authors choice not to publicly release them etc. Moreover the majority of these contributions do not provide wide information about the genres/domains of the retrieved parallel data, whereas it may be important to know the nature of possible training data with regards to their composition in terms of text types. Based on them a system able to collect parallel corpora from the web has been set up for this project, providing two new corpora in a variety of genres and domains, and their composition has been analysed and so doing provided an overview about the most common typologies of multilingual websites on the web for the consideredlanguages.

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 unpredictable but sentences are morefluid
5. SMT can be free opensource
6. SMT makes more fluidsentences
7. SMT and RBMT are matched for languages like French andSpanish
8. SMT can handle over 50 languages (Google, Bing andMicrosoft)
9. 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 a great example of how cutting edge research and world class infrastructure come together with Google. Google strives towards developing statistical translation techniques that improve more data and generalize well to new languages. The large scale computing infrastructure allows the translators and research scholars to rapidly experiment with new models trained on web-scale data to significantly improve translation quality.

This research backs the translations served at translate.google.com, allowing the users to translate text, web pages and even speech. Deployed within a wide range of

Conclusion

We have presented an in-depth study of machine translation consistency, using state-of- the-art SMT systems trained and evaluated under various realistic conditions. Our analysis highlights a number of important, and perhaps overlooked, issues regarding SMT consistency. First, SMT systems translate documents remarkably consistently, even without explicit knowledge of extra-sentential context. They even exhibit global consistency levels comparable to that of professional human translators. Second, high translation consistency does not correlate with better quality: as can be expected in phrase-based SMT, weaker systems trained on less data produce translations that are more consistent than higher-quality systems trained on larger more heterogeneous data sets. However, this does not imply that inconsistencies are good either: inconsistently translated phrases coincide with translation errors much more often than consistent ones. In practice, translation inconsistency could therefore be used to detect words and phrases that are hard to translate for a given system. Finally, manual inspection of inconsistent translations shows that only a small minority of errors are the kind of terminology problems that are the main concern in human translations. Instead, the majority of errors highlighted by inconsistent translations are symptoms of other problems, notably incorrect meaning translation, and syntactic or stylistic issues. 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 plan to explore the potential of consistency for confidence estimation 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

Baroni, M., Chantree, F., Kilgarriff, A. &Sharoff, S. (2008).Biber, D. (1988). Variation acrossspeech and writing . Cambridge University Press.28

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

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.

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

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

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

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

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

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