Automated Translation

Kashif Hussain

1/11/2012

Contents

1 Preface

Translation from one language to another has always been a focal point of human history from centuries past. Translating languages in order to understand them has always been difficult, but with the invention of computers, the hope of being able to understand different dialects has been elevated once more.

2 The History of Automated Translation[1]

One of the biggest difficulties in translating a language has been the differing alphabets and sentence structures of the language you want to translate to and the language you are translating from. Machine translation in the early days tried to learn how to translate by translating material written in a simple and predictable style on a limited subject matter. However, the computers which had machine translation capabilities all had one thing in common: they were horrible at translating. Accuracy for translators, at the time, was meant to describe how many sentences came out in a re-constructible form as no translator could translate a sentence from one language into an intelligible sentence in another.

The process which these computers used involved parsing[1] a sentence from one language which helped select the most natural reading of each word in the sentence. Once a sentence from the language you want to translate from has been broken down as a result of parsing, it had to be built up again into the language you wanted to translate to using the respective languages parser backwards.In between this process, however, comes the transfer stage which involves replacing the words from the language you want to translate from, to the language you want to translate to and there were two ways of doing this.

The most sophisticated way involved an intermediate language instead of listing a languages translation for each meaning of every word from the language you which to translate from. This intermediate language was based on Esperanto[1] called interlingua[2] This allowed the system to be expanded so that the language could be translated into a third language. A dictionary would be needed to convert the third language to and from the intermediate language. However, using the sophisticated way resulted in the problem growing as the number of languages increased which this added to the complexity of translation.

The other method involved the sentence-analysis. As the sentence-analysis translation required a lexicon, extra languages could be built into the lexicon

so they were also translatable, but this meant the lexicons usually became extremely long.

Automated translation in general required combining advanced machine translation, voice-recognition and voice-synthesis equipment all in order to reach the ultimate goal: simultaneous translation of unrestricted speech.

3 Current Automated Translation Capabilities

Over time, many technological improvements have been made to achieve the goal of simultaneous translation of unrestricted speech as well as new translation paradigms being discovered.

3.1 Statistical Machine Translation

"Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora[4]"[8] The notion of statistical machine translation was introduced by Warren Weave in 1949 using the idea of applying Claude Shannons information theory.[8] However, it played no part in the advancement of automated machine translation until 1991 where researchers at IBMs Thomas J. Watson Research centre re-introduced the concept which has made significant strides in the interest of machine learning. The benefits of the statistical machine translation over traditional paradigms was most often cited as the following[8]:

3.1.1 Benefits

- Better use of resources
- There is a great deal of natural language in machine-readable format
- Generally, SMT systems are not tailored to any specific pair of languages.
- Rule-based translation systems require the manual development of linguistic rules, which can be costly, and which often do not generalize to other languages.
- More natural translations

As great as statistical machine translation paradigm is, it still has many shortcomings such as the following:

3.1.2 Shortcomings

- Statistical machine translation do not work well between languages that have significantly different word orders (e.g. Japanese and European languages).
- The benefits are overemphasized for European languages.

The ideas behind statistical machine translation come out of information theory. Essentially, the document is translated on the probability p(e—f) that a string e in native language (for example, English) is the translation of a string f in foreign language (for example, French). Generally, these probabilities are estimated using techniques of parameter estimation. One approach to modelling p(e—f) is applying Bayes Theorem which would give you the best translation possible from SMT.

3.1.3 Word-based translation

In word-based translation, translated elements are words. Typically, the number of words in translated sentences are different due to compound words, morphology and idioms. The ratio of the lengths of sequences of translated words is called fertility, which tells how many foreign words each native word produces. Simple word-based translation is not able to translate language pairs with fertility rates different from one. To make word-based translation systems manage, for instance, high fertility rates, the system could be able to map a single word to multiple words, but not vice versa. For instance, if we are translating from French to English, each word in English could produce zero or more French words. But there's no way to group two English words producing a single French word.

3.1.4 Phrase-based translation

In phrase-based translation, the aim is to reduce the restrictions of word-base translation by translating whole sequences of words to sequences of words, where the lengths can differ.[8] The sequences of words are called, for instance, blocks or phrases, but typically are not linguistic phrases but phrases found using statistical methods from the corpus. Restricting the phrases to linguistic phrases has been shown to decrease translation quality. By 2006 statistical machine translation became the most widely-studied machine translation paradigm[2] and still is to this day.

3.2 Modern Translation Utilities

3.2.1 Google Translate

Machine translation has come a long way, not just through the reintroduction of past paradigms, but also through the rapid expansion of the web and the web bore fruit to a massive internet search engine which we know today as "Google". Google translate is one of the many systems that google has pioneered.

"Google Translate is a free translation service that provides instant translations between dozens of different languages. It can translate words, sentences and web pages between any combination of our (current 71) supported languages."[3]

Google translate uses the statistical machine translation paradigm. With it, it is able to produce translations of text, documents and web pages of dozens of languages promptly at the press of a button.

We know that a computer can learn a foreign language by referring to vocabulary and a set of rules, but languages are complicated and there are exceptions to almost any rule. When you try to capture all of these exceptions, and exceptions to the exceptions, in a computer program, the translation quality begins to break down. Google Translate takes a different approach. Instead of trying to teach their computers all the rules of a language, they let their computers discover the rules for themselves (through their learning algorithm). They learn by analysing millions and millions of documents that have already been translated by human translators. These translated texts come from books, organizations like the UN and websites from all around the world. Their computers scan these texts looking for statistically significant patterns - that is to say, patterns between the translation and the original text that are unlikely to occur by chance. Once the computer finds a pattern, it can use this pattern to translate similar texts in the future. When you repeat this process billions of times you end up with billions of patterns and one very smart computer program: Google Translate

However, programs like Google Translate arent perfect. For some languages, they have fewer translated documents available and therefore fewer patterns that their software has detected and so translation quality will vary by language and language pair

4 The Future

Whilst automated machine translation has come from what it was in the last century, it still has room to grow but the future looks promising.

Whilst machine translation can currently translate text of some of the more popular languages and vice versa, it cannot do so for all languages. The translation capability, as described with google translate, varies for each language due to the number of texts available which have been translated by human translators.

However, machine translation is gaining ground especially in the speech translation department where speech can be almost simultaneously translated from one sentence to another. Microsoft's research team who worked on the speech translation software is one of the pioneers in this area.

The speech translations software is dedicated to provide a "real-time automatic translation system that can handle the informality and grammatical lapses that characterize our everyday chats."[4] The software preserves intonation and cadence so that the translated speech still sounds like the original speaker Microsoft used statistical models that did a better job of capturing the range of human vocal ability. Using these statistical models, Microsoft were able to "cut error rates to about 15%"[5] Microsoft demonstrated this prototype technology by speaking in English and having the software automatically translate the speaker's voice into Chinese and reordering the Chinese text on-screen so that they made sense. The speakers voice was then piped to a text-to-speech system which read out the text sounding like the speaker.

NTT Docomo, Japan's largest mobile service provider, has developed an app that allows people to translate their calls almost immediately. "After speaking into the phone, the app will translate what you have said into the receiver's language – after a slight pause – and then it provides a voice readout as well as screen text version of your message to your phone."[6] Whilst the translations aren't perfect, the basic message is conveyed.

Many different technological companies including the likes of AT&T and google have similar projects underway. With Google innovating new technologies such as google glass, Google's Vice President of Android stated, "Google is now in the early stages of creating real-time translation software which is hopes to perfect in several years."[7]

References

[1] Economist 14th November 1987 Science and Technology pg 97 - 98
http://studentnet.cs.manchester.ac.uk/ugt/year2/readingWeekSources/translation.pdf

[2] Translation Directory September 2008 Statistical Machine Learning
http://www.translationdirectory.com/articles/article1684.php

[3] Inside Google Translate How our translations work
http://translate.google.com/about/intl/en_ALL/

[4] Microsoft Research - How Technology Can Bridge Language Gaps
http://research.microsoft.com/en-us/research/stories/speech-to-speech.aspx

[5] BBC News Technology 9th November 2012 Microsoft demos instant English-Chinese translation
http://www.bbc.co.uk/news/technology-20266427

[6] CNN Travel Article 22nd October 2012 Whatever my phone said: NTT DoCoMo's real time speech translation app
http://travel.cnn.com/explorations/life/tech-trend-speech-translation-apps-smartphone-conversations-085231

[7] Techradar July 26th 2013 Google is evidently working on real-time mobile translation tech
http://www.techradar.com/news/world-of-tech/future-tech/google-is-working-on-real-time-translator-phones-1169019

[8] Wikipedia Article Statistical Machine Translation http://en.wikipedia.org/wiki/Statistical_machine_translation

[1]Parsing involves resolving (a sentence) into its component parts and describe their syntactic roles.

[2]Interlingua is an artificial language, devised for machine translation, that makes explicit the distinctions necessary for successful translation into a target language, even where they are not present in the source language

[3]Text Corpora: In linguistics, a corpus or text corpus is a large and structured set of texts