>>> revista tradumàtica
Traducció i Tecnologies de la informació i la Comunicació
Número 4 : traducció automàtica : desembre 2006 : http://www.fti.uab.cat/tradumatica/revista :
ISSN 1578-7559

Computational Linguistics and Machine Translation Research and Development

David Farwell
Computing Research Laboratory, New Mexico State University, USA

Introduction

It is not always easy to distinguish between research and development or demonstration system and product in the area of Machine Translation (MT) however, it is useful to distinguish between fully automatic machine translation, human-assisted machine translation and machine-assisted human translation all of which fall within the purview of MT. In this article the focus is exclusively on the first of these, fully automatic machine translation.

The mid-1980's on into the early 1990's marked the zenith of linguistically inspired, rule-based approaches to MT. Major projects included Eurotra (Europe, 1982-1993; Durand, et al. 1991), the fifth generation initiative (Japan, 1981-1993; Nagao 1989) and various small experimental systems in the US (e.g., Carnegie Mellon University – Goodman & Nirenburg 1991; New Mexico State University – Farwell & Wilks, 1991; University of Maryland – Dorr 1993 among others) which culminated in the Pangloss-Mikrokosmos Spanish-English Knowledge-Based MT system (Nirenburg 1995). But, in no small measure it was the US government funded MT initiative from 1991 through 1995 through which Pangloss was funded which led to a profound shift in focus for MT R&D (and not simply for MT but for all NLP tasks). That same initiative also funded the development of Candide (Brown, et al. 1993), a French-English statistics-based MT system developed at IBM which to this day is the paradigm for statistics-based approaches. Equally importantly, the initiative funded the development of a framework for evaluating and comparing the performances of different systems over a "comparable" task (White & O'Connell 1994). The upshot of this initiative was that Candide out-performed Pangloss on the evaluation task and, in fact, almost performed as well as Systran French-English, the leading commercial system that participated in the evaluation exercise, a rule-based system that had some 15 years of development behind it. Thus, almost more impressive than the translation results was the fact that those results were achieved by a relatively small research team (5 or so members) in a rather short time (roughly three to five years).

This initiative had a profound effect on MT R&D. By the year 2000 there was a handful of groups around the world working on statistical MT (IBM Yorktown Heights., ISI, CMU, University of Aachen, and Karlsruhe University). A year earlier there was a summer workshop for researchers at Johns Hopkins where the participants were able to assemble an infrastructure and develop a respectable working system. By 2002 there were at least a dozen research groups around the world working on statistical MT and by 2005 at least one system, from Language Weaver, a spin-off from the ISI research group in California, had been installed at the CIA and incorporated into the workflow of intelligence analysts of documents in unfamiliar source languages.

Current MT R&D

There are two major corpus-based approaches to Machine Translation that have become the focus of the research community. Obviously, given what has been presented thus far, there is a good deal of energy being expended on improving the state of the art of statistical systems. But there is also a fairly high degree of activity aimed at developing example-based MT. In addition, as has always been the case in this field, evaluation is receiving a good deal of attention.

Statistical MT

A prototypical stochastic system consists of three statistical models: alignment, translating (or decoding) and target language modeling for improving fluency. The first two are built by looking at parallel texts, lots of parallel texts, millions of words of parallel text if possible. The text is generally broken down into short units (such as sentences). The alignment model essentially provides statistics to answer questions such as the following: given the third word of the source language string, what is the likelihood its counterpart is the first word of the target language string, the second word of the target language string, the third word, and so on. Given the fourth word of the source language string, what is the likelihood that its counterpart is the first word of the target language string, the second word, and so on. Then by simply counting the number of times each case is true in a huge corpus and dividing by the number of strings altogether, the results is a set of likelihoods for each alignment. The translation model essentially provides answers to the following: given a word “s” (possibly in a specific context), what is the likelihood that its target language equivalent (the aligned counterpart) is “x,” is “y,” is “z,” and so on? The answer can be provided by simply looking at the aligned counterparts of all the occurrences of “s” (possibly in specific context) in the source language corpus and dividing by the total number of occurrences. Now when these two statistical models are applied to a sequence of words in a novel source language string, they will suggest a sequence of words in the target language. Actually, they will suggest any number of possible strings having different words (translations) in different orders (alignments). To select the most promising, the third model is applied, the target language model. Looking only at the target language text, this model essentially provides statistics to answer the following question: given words “a” and “b” in that order, what is the likelihood that the next word in the sequence is “c,” what is the likelihood that the next word is “d,” and so on for all the words of the language. To calculate these statistics, every sequence of “a” followed by “b” followed by “x” is inspected and, for each different word “x,” the number of occurrences and divide by the total number of “a b x” sequences in the corpus. Returning to the different translation suggested by the alignment and translation models, the target language model is applied to each suggestion in order to calculate which is the most probable.

Generally this approach is better and better the larger the parallel corpus there is for training the models because the smaller the corpus, the more likely novel combinations of words will be encountered for which there is insufficient statistical information to make choices. On the other hand, there are so many possibilities to calculate that even modern CPU and storage capacities are such that it might take months or years to actually carry out all the necessary calculations. Thus we arrive at what is capturing the interest of statistical MT researchers. The basic issues are:

- How can approximate a complete calculation be approximated so that the statistics are reliable but at the same time the calculation is possible within constraints of time and memory,
- How can we improve each of the models (in particular, what parameters can we

use beside word form, word form sequences and word form alignments that might improve the estimates.
In the former case, the expectation is that larger corpora (such as the world wide web) will inevitably lead to improved results. In addition, there have been experiments with different alignment techniques which focus on “segments” of sentences (e.g., Deng, et al. 2004) or bilingual parallel “segments” extracted from non parallel texts (e.g. Munteanu, et al. 2004). In the latter case, even such naïve additions word stems, part-of-speech or morphological information (case, number, noun-adjective agreement, verb-nominal agreement, etc.) have sometimes lead to improved performance (but not necessarily!). Yamuda and Knight (2001), for instance, describe a technique for developing syntactic transfer systems using aligned corpora in which at least one of the languages is syntactically annotated. In the near future there is sure to be experimentation using additional linguistically motivated morphosyntactic and possibly semantic parameters.

Example-based MT

A prototypical example-based system also is corpus based but it approaches the corpus with different assumptions and different goals. In this case the idea is to break the parallel corpus down into repeating translation units, constituent-sized templates generally, sequences of words with possible variables interspersed (such as “fishing license” – “permiso” or “licencia de pesca” vs “drivers license” – “carnet” or “permiso de conducer”) but including sequences of constituents as well (such as “level a building” – “arrasar un edificio” vs “level the score” – “igualar el marcador” vs “level charges against” – “hacer acusaciones en contra”) and on up to full sentence templates (such as “in for a penny, in for a pound” – “de perdidos, al rio”). These equivalence units are then stored in a large database which can later be used to support the translation process or such equivalences may be detected on the fly as part of the process of translation. In either case, during translation, the input text is first matched against the templates on the source language side of the example base and, if a match is found, the corresponding target template is available for generation. If no match is found, or if there is untranslated material corresponding to a template variable, then the text may be translated using a traditional rule-based system. In some sense, the equivalences in the example base are similar to the expressions recorded in a translation memory except, they may be so basic that any translator would think as too obvious to warrant recording and they may be so “literal” (e.g. it might include “level a building”, “level a barn”, “level a skyscraper”, and so on) as to not merit the time to record them. This approach was initially developed in the 1980’s by the Kyoto University MT research group working on a grammar-based approach to Japanese-English translation (Nagao 1984). It has three principle advantages. First, it allows for the treatment of discontinuous constituents such as “figure out” in English which often appears with intervening material as in “figure the answer out.” Second, it allows the translation system to deal with idiosyncratic collocational phenomena (which all of the examples above reflect). Finally, it allows the translation system increased potential to generate natural fluent as well as colloquial target language text. More recently, it also provides an additional advantage. It can be incorporated more easily into a rule-based MT system (as opposed to the combination of rule-based and statistics-based systems.

As example-based approaches try to increase the use of the examples during processing (as opposed to applying a general grammar-based MT analyzer/generator) they appear to be slowly converging with statistical approaches (which conversely appear to be moving from a word-level focus to a constituent level focus). More recently, interest in example-based MT approach

has focused on trying to skip the construction of an example base and instead attempt to use the parallel corpus directly as a source of examples during the translation process. For this exercise to work, corpus alignment is extremely important, as it is for statistical approaches, although there is perhaps more concern for establishing a constituent level alignment as opposed to a word or string level alignment (e.g., Owczarzak, et al. 2006). In addition, there is a good deal interest in improving the matching process between source text and the source corpus of the parallel aligned corpus and long with merging target language text segments together to form fluent output (Brown, et al. 2003).

Evaluation

As mentioned earlier, a major outcome of the US government funded MT initiative in the early 1990's was the development an evaluation methodology that could compare system performance on comparable tasks, the translation of texts in a common genre (namely, news articles) and of a similar length. That methodology however relied on human evaluators who assessed such factors as fidelity and fluency and as a result was both expensive and time consuming, especially from the perspective of MT developers. As a result, in the late 1990's an automatic evaluation technique was developed at IBM which, while not especially useful as a diagnostic, was shown to correlate with human judgments of relative quality. BLEU (Papineni, et al. 2002), as the methodology is referred to, is a statistical metric which provides a score based on the number and length of text segments in an output translation which match text segments of one or more reference (human generated) translations. It is widely used at this point and helps developers by indicating whether a more recent version of their system performs better, on a par with or worse than a prior version as well as telling them how the performance of their system compares with others over a common test set.

But BLEU has it draw backs not the least of which is the fact that test set have to be developed generally by human translators. It is not very insightful. It does not recognize categories of errors nor the strengths and weaknesses in some broad sense of different systems and so is not useful as a diagnostic tool. It is entirely focused on a throughput, that is, the relationship of the input and output texts. As a result, other evaluation methodologies have been proposed and there has been at least one effort, FEMTI (King, et al. 2003), to systematically analyze the objectives of an evaluation and to suggest a range of metrics based on objectives.

Integrating fully automatic MT systems into translation process

Statistical MT systems are fully automatic translators and cannot actually be integrated into the work stream of a particular translator. Rather, they are used replace the translator. But the quality, while much improved, is not especially good. Thus such systems are generally used to support document filtering for assimilation tasks such as information analysis, email and chat specifically for texts in languages unknown to a "customer." In this case the system provides its translation such as it may be and the customer must decide whether the document appears to be worth closer investigation in which case it is passed to a human translator. There are some cases of applying fully automatic systems to dissemination tasks (e.g., job descriptions) especially if boring, repetitive, closed domain translation is involved. In these cases translations are automatically generated and then passed to a human, ideally monolingual, post-editor who produces a fluent version of the translation. In fact, one area of growing interest in the MT research community is in developing automatic (statistical) post-editors. In any case, the key here is that the task involve a domain-limited repetitive translation task and that the automatic translation are sufficiently high quality to

make post-editing more efficient (and lees expensive) than human translation.

Using fully automatic MT for second language learning or translator training

Using translation in language second language learning has been controversial for some time but for those who find it useful, fully automatic MT could conceivably be (and no doubt already have been) incorporated into on line reading, writing and translation exercises. Whether the system provides high quality translations or merely hints at the content of the source text, it might be used (obviously accordingly) to assist in understanding or producing texts in the language being acquired as well as to provide materials which need to be edited using knowledge of the language to be acquired. But this topic is outside the area of expertise and is best left to the interested reader.

One area of potential benefit to both translator training and MT, however, would be activities that promote the development of corpora consisting of multiple translations (in a given target language) of a given set of source language documents, especially if the translation were annotated with linguistic information (morphological, syntactic and semantic information). For translators the central activity would be to compare and contrast translations, identifying wherever translation vary whether the variation is the result of an error (classification being useful), a non meaning impacting variation (i.e., essentially paraphrases communicating the same information content), or meaning bearing variations permissible within the set of possible (rational) interpretations of a text. The resultant corpus would benefit MT for both training and evaluating MT systems and presumably benefit developing translators by sensitizing them to the enormous range of plausible interpretations (and therefore translation) a text may have as well as providing an interesting methodology for evaluating a translator's level of proficiency and improvement over time.

References

Brown, P. F., Della Pietra, S.A., Della Pietra, V.J., and Mercer, R.L. 1993. "The mathematics of statistical machine translation: parameter estimation." *Computational Linguistics* 19 (2), 263-311.

Brown, R., R. Hutchinson, P. Bennett, J. Carbonell, and P. Jansen. 2003. Reducing Boundary Friction Using Translation-Fragment Overlap", in *Proceedings of the Ninth Machine Translation Summit,* New Orleans, USA, pp. 24-31.

Deng, Y., S. Kumar, and W. Byrne. 2004. Bitext Chunk Alignment for Statistical Machine Translation. *CSLP Tech Report,* Johns Hopkins University.

Dorr, B. J. 1993. *Machine translation: a view from the lexicon*. MIT Press, Cambridge, Mass.

Durand, J., P. Bennett, V. Allegranza, F. Van Eynde, L. Humphreys, P. Schmidt & E. Steiner. 1991. The Eurotra Linguistic Specifications: an overview, In: *Machine Translation* 6, Kluwer, Dordrecht, pp. 103-147.

Farwell, D., and Y. Wilks. 1991. ULTRA: A Multilingual Machine Translator. *Proceedings of the Machine Translation Summit III*, 19-24.

Goodman, K. and Nirenburg, S. (eds.) 1991. *The KBMT project: a case study in*

knowledge-based machine translation. San Mateo, CA: Morgan Kaufmann.

King, M., Popescu-Belis, A. and Hovy, E. 2003. "FEMTI: creating and using a framework for MT evaluation" In: *AMTA* (2003), 224-231.

Munteanu, D., A. Fraser, and D. Marcu. 2004. Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. *Proceedings of HLT/NAACL*.

Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In: Elithorn, A. & Banerji,R.(eds.) *Artificial and human intelligence* (Amsterdam: North-Holland)

Nagao, M.1989. *Machine translation: how far can it go?* (Oxford: Oxford University Press)

Nirenburg, S. (ed.) 1995. The Pangloss Mark III Machine Translation System. A Joint Technical Report by NMSU CRL, USC ISI and CMU CMT. Issued as CMU tech report CMU-CMT-95-145 (Also available as HTML from NMSU).

Owczarzak, K., B. Mellebeek, D. Groves, J. Van Genabith and A. Way. 2006. Wrapper Syntax for Example-based Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Boston, MA., pp.148—155.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. 2002. "BLEU: a method for automatic evaluation of machine translation." In: *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics,* Philadelphia, July 2002; 311-318.

White, J.S. and T.A. O'Connell. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*.

Yamada, K., and K. Knight. 2001. A syntax-based Statistical Translation Model. *Proceedings of ACL*, 523-530, Toulouse, France.

Desembre 2006