[From: *Ezine Articles*, submitted 8 May 2009]

Machine Translation - How it Works, What Users Expect, and What They Get

By Neil Coffey

Machine translation (MT) systems are now ubiquitous. This ubiquity is due to a combination of increased need for translation in today's global marketplace, and an exponential growth in computing power that has made such systems viable. And under the right circumstances, MT systems are a powerful tool. They offer low-quality translations in situations where low-quality translation is better than no translation at all, or where a rough translation of a large document delivered in seconds or minutes is more useful than a good translation delivered in three weeks' time.

Unfortunately, despite the widespread accessibility of MT, it is clear that the purpose and limitations of such systems are frequently misunderstood, and their capability widely overestimated. In this article, I want to give a brief overview of how MT systems work and thus how they can be put to best use. Then, I'll present some data on how Internet-based MT is being used right now, and show that there is a chasm between the intended and actual use of such systems, and that users still need educating on how to use MT systems effectively.

How machine translation works

You might have expected that a computer translation program would use grammatical rules of the languages in question, combining them with some kind of in-memory "dictionary" to produce the resulting translation. And indeed, that's essentially how some earlier systems worked. But most modern MT systems actually take a statistical approach that is quite "linguistically blind". Essentially, the system is trained on a corpus of example translations. The result is a statistical model that incorporates information such as:

- "when the words (a, b, c) occur in succession in a sentence, there is an X% chance that the words (d, e, f) will occur in succession in the translation" (N.B. there don't have to be the same number of words in each pair);
- "given two successive words (a, b) in the target language, if word (a) ends in -X, there is an X% chance that word (b) will end in -Y".

Given a huge body of such observations, the system can then translate a sentence by considering various candidate translations-- made by stringing words together almost at random (in reality, via some 'naive selection' process)-- and choosing the statistically most likely option.

On hearing this high-level description of how MT works, most people are surprised that such a "linguistically blind" approach works at all. What's even more surprising is that it typically works better than rule-based systems. This is partly because relying on grammatical analysis itself introduces errors into the equation (automated analysis is not completely accurate, and humans don't always agree on how to analyse a sentence). And training a system on "bare text" allows you to base a system on far more data than would otherwise be possible: corpora of grammatically analysed texts are small and few and far between; pages of "bare text" are available in their trillions.

However, what this approach does mean is that the quality of translations is very dependent on how well elements of the source text are represented in the data originally used to train the system. If you accidentally type he will returned or vous avez demander (instead of he will return or vous avez demandé), the system will be hampered by the fact that sequences such as will returned are unlikely to have occurred many times in the training corpus (or worse, may have occurred with a completely different meaning, as in they needed his will returned to the solicitor). And since the system has little notion of grammar (to work out, for example, that returned is a form of return, and "the infinitive is likely after he will"), it in effect has little to go on.

Similarly, you may ask the system to translate a sentence that is perfectly grammatical and common in everyday use, but which includes features that happen not to have been common in the training corpus. MT systems are typically trained on the types of text for which human translations are readily available, such as technical or business documents, or transcripts of meetings of multilingual parliaments and conferences. This gives MT systems a natural bias towards certain types of formal or technical text. And even if everyday vocabulary is still covered by the training corpus, the grammar of everyday speech (such as using tú instead of usted in Spanish, or using the present tense instead of the future tense in various languages) may not.

MT systems in practice

Researches and developers of computer translation systems have always been aware that one of the biggest dangers is public misperception of their purpose and limitations. Somers (2003)[1], observing the use of MT on the web and in chat rooms, comments that: "This increased visibility of MT has had a number of side effets. [...] There is certainly a need to educate the general public about the low quality of raw MT, and, importantly, why the quality is so low." Observing MT in use in 2009, there's sadly little evidence that users' awareness of these issues has improved.

As an illustration, I'll present a small sample of data from a Spanish-English MT service that I make available at the Español-Inglés web site. The service works by taking the user's input, applying some "cleanup" processes (such as correcting some common orthographical errors and decoding common instances of "SMS-speak"), and then looking for translations in (a) a bank of examples from the site's Spanish-English dictionary, and (b) a MT engine. Currently, Google Translate is used for the MT engine, although a custom engine may be used in the future. The figures I present here are from an analysis

of 549 Spanish-English queries presented to the system from machines in Mexico[2]-- in other words, we assume that most users are translating from their native language.

First, what are people using the MT system for? For each query, I attempted a "best guess" at the user's purpose for translating the query. In many cases, the purpose is quite obvious; in a few cases, there is clearly ambiguity. With that caveat, I judge that in about 88% of cases, the intended use is fairly clear-cut, and categorise these uses as follows:

- Looking up a single word or term: **38%**
- Translating a formal text: **23%**
- Internet chat session: **18%**
- Homework: **9%**

A surprising (if not alarming!) observation is that in such a large proportion of cases, users are using the translator to look up a single word or term. In fact, 30% of queries consisted of a single word. The finding is a little surprising given that the site in question also has a Spanish-English dictionary, and suggests that users confuse the purpose of dictionaries and translators. Although not represented in the raw figures, there were clearly some cases of consecutive searches where it appeared that a user was deliberately splitting up a sentence or phrase that would have probably been better translated if left together. Perhaps as a consequence of student over-drilling on dictionary usage, we see, for example, a query for cuarto para ("quarter to") followed immediately by a query for a number. There is clearly a need to educate students and users in general on the difference between the electronic dictionary and the machine translator[3]: in particular, that a dictionary will guide the user to choosing the appropriate translation given the context, but requires single-word or single-phrase lookups, whereas a translator generally works best on whole sentences and given a single word or term, will simply report the statistically most common translation.

I estimate that in less than a quarter of cases, users are using the MT system for its "trained-for" purpose of translating or gisting a formal text (and are entering an entire sentence, or at least partial sentence rather than an isolated noun phrase). Of course, it's impossible to know whether any of these translations were then intended for publication without further proof, which definitely isn't the purpose of the system.

The use for translating formal texts is now almost rivalled by the use to translate informal on-line chat sessions-- a context for which MT systems are typically not trained. The on-line chat context poses particular problems for MT systems, since features such as non-standard spelling, lack of punctuation and presence of colloquialisms not found in other written contexts are common. For chat sessions to be translated effectively would probably require a dedicated system trained on a more suitable (and possibly custom-built) corpus.

It's not too surprising that students are using MT systems to do their homework. But it's interesting to note to what extent and how. In fact, use for homework includes a mixture of "fair use" (understanding an exercise) with an attempt to "get the computer to do their

homework" (with predictably dire results in some cases). Queries categorised as homework include sentences which are obviously instructions to exercises, plus certain sentences explaining trivial generalities that would be uncommon in a text or conversation, but which are typical in beginners' homework exercises.

Whatever the use, an issue for system users and designers alike is the frequency of errors in the source text which are liable to hamper the translation. In fact, over 40% of queries contained such errors, with some queries containing several. The most common errors were the following (queries for single words and terms were excluded in calculating these figures):

- Missing accents: **14%** of queries
- Missing punctuation: **13%**
- Other orthographical error: **8%**
- Grammatically incomplete sentence: **8%**

Bearing in mind that in the majority of cases, users where translating from their native language, users appear to underestimate the importance of using standard orthography to give the best chance of a good translation. More subtly, users do not always understand that the translation of one word can depend on another, and that the translator's job is more difficult if grammatical constituents are incomplete, so that queries such as hoy es día de are not uncommon. Such queries hamper translation because the chance of a sentence in the training corpus with, say, a "dangling" preposition like this will be slim.

Lessons to be learnt...?

At present, there's still a mismatch between the performance of MT systems and the expectations of users. I see responsibility for closing this gap as lying in the hands both of developers and of users and educators. Users need to think more about making their source sentences "MT-friendly" and learn how to assess the output of MT systems. Language courses need to address these issues: learning to use computer translation tools effectively needs to be seen as a relevant part of learning to use a language. And developers, including myself, need to think about how we can make the tools we offer better suited to language users' needs.

Notes

[1] Somers (2003), "Machine Translation: the Latest Developments" in The Oxford Handbook of Computational Linguistics, OUP.
[2] This odd number is simply because queries matching the selection criteria were captured with random probability within a fixed time frame. It should be noted that the system for deducing a machine's country from its IP address is not completely accurate.
[3] If the user enters a single word into the system in question, a message is displayed beneath the translation suggesting that the user would get a better result by using the site's dictionary.

The ESPANOL-INGLES web site offers various resources for English-speaking learners of Spanish and vice versa, including a Spanish dictionary, Spanish phrases section with audio recordings, plus grammar information and on-line word games.

Article Source: http://EzineArticles.com/?expert=Neil_Coffey