JEFF ALLEN WRITES:

What about statistical MT?

I have been asked to specifically address in this column the issue of statistical Machine Translation (MT) with regard to an article entitled "Machine Translation: DARPA calls for MT papers" that appeared in the IJLD issue 5, June 2000, p. 20.

The Defense Advanced Research Projects Agency (DARPA) is one of the important funders of MT projects throughout the United States. Cited in that article from a DARPA memorandum is the statement that "Recent experiments with statistical machine translation ... have suggested that, with new computer power and recent techniques, it may be time to revisit statistical machine translation as an approach that may scale to new language pairs depending on availability of linguistic data" and "Modern statistical techniques are beyond the capabilities of many professional linguistics researchers but offer the best hope for addressing the need to communicate among the world's 6,700 languages" (IJLD 5, p. 20). In essence, DARPA would like to focus on statistical-based approaches of MT and natural language processing for future projects.

I am not convinced that Statistical MT is the solution to breaking the language barrier for the majority of the world's unsupported languages. Myself, having worked on an MT project that included the statistical approach for several languages, of which some were considered to be minority, less-prevalent, sparse-data languages, I would like to comment on this approach.

The first of my comments on this topic has to do with the majority of the world's languages which according to DARPA are considered to be low-density languages. This term (and related terms like sparse-data) simply reflects the fact that the majority of languages currently have little or no electronic data with which to work. One of the main sources of electronic language data that any sparse-data language researcher wants to get hold of is the Bible. According to public reports, the Summer Institute of Linguistics (http://www.sil.org) is currently working on 1000 different languages (SIL, 2000). The result of this linguistic work is that the Suriname Javanese New Testament was recently celebrated as the 500th translation that Wycliffe Bible translators (Wycliffe, 2000a) has helped to complete. These organisations began in 1934, and currently with nearly 6000 short and career staff members (Wycliffe, 2000b), it will be a while before 1000, 2000, or even 3000 languages have complete electronic versions of the New Testament. As for other statistics, the Jesus Film Project recently noted that there are currently 606 languages with audio and / or visual (DVD, film, video) of the film (Jesus Film Project, June 2000), that to my knowledge is based on the Gospel of Luke, of which the majority of the languages are considered to be minority or neglected languages. Again, this represents 10% of the world's languages.

Jeff Allen
postediting@aol.com

Taking a look at the data itself, the entire Haitian Creole Bible is 4.2 megabytes of plain ascii text data. So, the New Testament alone basically represents about 1 megabyte of textual data per language. With respect to speech and transcribed texts, the Jesus Film soundtrack for the book of Luke for one language (Haitian Creole) contains just over 0.5 megabytes of .wav format files. Since my team of Haitian linguists performed the transcription validation work of the speech files of the Jesus Film for Haitian Creole, I know that the transcribed speech files represent 50 kilobytes (Kb) of textual data. Yet, when large corporations conduct textmining and datamining statistical analyses on their abundant technical documentation, they often start with approximately 50 megabytes (Mb) of textual data per language. In order words, the majority of the languages spoken in the world today do not even have an entire electronic version of the Bible for the language, and even if they do, it only represents 1/10 of the textual data that is needed to conduct basic natural language processing statistical testing. The limiting factor in the equation still remains the fact that there is a significant lack of electronic data for the majority of the 6,000+ languages of the world. It is simply not valid to run statistical methods on data that does not exist or barely exist at all.

Secondly, statistical MT is exactly what the name indicates: it is a strategy used to make word, phrase, and sentence choices through a statistical analysis of the data that is studied. Statistical methods imply that there are numerous occurrences of each of the items likely to be chosen and that there is a high probability of finding threads of commonalities and similarities that can be used to detect the most probable items in order to make default choices. Yet, with such a limited amount of data for any given sparse-data language, there are usually not enough occurrences for the majority of entries in order to make proper statistical analyses of them. Let us look at a case study. In a 13,000 entry word-frequency database for Haitian Creole that was derived from a 1.2 million word text database taken from 13 different textual sources, 9,600 word entries occur more than 3 times of which only 4,400 of the

Continued on page 42

Jeff Allen
postediting@aol.com

word entries occur more than 10 times each. Also, nearly 3,500 of the word entries occur less than 2 times. This information indicates that using statistical methods really only helps for possibly 1/3 of the items in the word frequency list that was extracted from a database.

The design, creation, and compilation of any language database is not an easy task, and it is certainly more difficult to undertake for sparse-data languages. I remind IJLD readers of the issues that I brought up in a previous column article (Allen, 2000). Some comparative time tables for various international languages and sparse-data languages are also provided in Lenzo et al (1998).

A third point to consider is that it is often tempting to see a new method or technology as the language technology solution. MT was seen as this for several decades, but now we know that Translation Memory (TM), which basically stems out of Example-based MT methodologies, has become one of the primary productivity boosters for the translation and localisation sectors over the past 10 years. With each new technology, it is easy to swing from one extreme to the other. Yet, nearly all players in the industrial and corporate sectors who implement translation technologies have been advised to combine and integrate both TM and MT technologies into their processes. I have even seen cases where companies implement multiple competing TM tools and MT systems in order to render their translation processes as cost-efficient as possible. In other words, it is best not to put all of your eggs into a single basket, but rather to strategically emphasise the strong points of each of the component tools. This has been referred to as Multi-Engine MT (MEMT) by developers at the Centre for Machine Translation of Carnegie Mellon University (Hogan and Frederking, 1998). Based on this project, and many similar projects, the combination of different MT approaches seems to be the most favourable for the rapid development and rapid deployment of MT systems, especially for sparse-data languages. I do not expect the statistical MT approach to be the sole beneficial way of dealing with the majority of the world's "sparse-data" languages in the medium-term.

In summary, I would say that statistical MT is one possible approach, and that, when combined with other types of MT systems, it can provide improved results. However, the hindering fact is that most of the languages in the world today lack electronic data upon which to test and train systems. Thus, statistics-based methods as a primary approach should be reserved for the international languages that have sufficient data with which and from which to work. Only data collection and data compilation efforts with a significant amount of invested human resources could eventually allow the world's less-supported languages to also benefit more amply from the statistical MT approach.

Jeff Allen can be contacted by e-mail at <posteditting@aol.com>.

References:

- Allen, Jeffrey. 2000. The risks of spelling variation and reform. In IJLD 4, April 2000, pp. 41-42.
- Lenzo, Kevin, Christopher Hogan, and Jeffrey Allen. 1998. Rapid-Deployment Text-to-Speech in the DIPLOMAT System. Poster presented at the International Conference on Spoken Language Processing. 30 November - 4 December 1998, Sydney, Australia.
- Hogan, Christopher and Robert Frederking. 1998. An Evaluation of the Multi-Engine MT Architecture. In Farwell et al (eds.) Machine Translation and the Information Soup, proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA), Philadelphia, 28-31 October 1998.
- Jesus Film Project. 29 June 2000. Update on Languages available. Available online at: http://www.jesusfilm.org/updates/languages.html
- Summer Institute of Linguistics (SIL), 2000. Overview of Translation in SIL. Available online at: http://www.sil.org/translation/TransinSIL.htm
- Wycliffe Bible translators. 2000a. 500th Translation a Testament to God's Power, Faithfulness. Feature story available online at:
  http://www.wycliffe.org/features/500thNT/home.htm
- Wycliffe Bible translators. 2000b. Frequently Asked Questions. Available online at: http://www.wycliffe.org/cgi-bin/jumpbox.pl

- IJLD. 2000. Machine Translation: DARPA calls for MT papers. In IJLD issue 5, June 2000, p. 20.