

# DARPA calls for MT papers

The Defense Advanced Research Projects Agency (DARPA) of the United States government has issued a call for papers and proposals for funding in the financial year 2001 for systems of statistical machine translation (MT). This could have a major impact on MT research in the next decade. Machine translation in the United States has not benefited from heavy government investment (as distinct from procurement) since the infamous official ALPAC report of the 1960s which dismissed the idea that machine translation could be cost-effective.

"Recent experiments with statistical machine translation", the DARPA memorandum states, "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... New institutions, new research workforce members, and new ideas are needed to address the rapidly-developing need for translation among the World's languages. 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."

The call for proposals is numbered 37 in a list of 38 fields for research announced by DARPA in a recent memorandum. The other fields of research are all for highly specialised technologies in fields such as infrared sensing and detection, coatings, explosives etc.

The principal objective of the new call for MT proposals is stated that of exploring statistical language models, statistical sequence re-ordering techniques, and machine learning approaches for improved, automatic machine translation. There is an additional goal, which is to address the critical shortage of language technology researchers and institutions qualified to investigate natural language understanding using statistical tools.

It is this latter goal which will make many people in the universities and companies developing machine translation systems sit up and take notice.

To back up the call for proposals DARPA experts have attempted to sketch in the background of machine translation, and it is interesting for IJLD readers to see how the US government's defence arm views this area. We therefore quote it in full. The reference to English being the lingua franca of the European Union may raise some eyebrows.

"Attempts to translate human language have a history of more than 50 years of effort. It is the "holy

grail" of natural language understanding since unambiguous translation requires deep understanding of word senses, idioms, and colloquialisms. The need for language translation will not go away anytime soon, in spite of the apparent adoption of English in many contexts as a lingua franca (e.g. the European Union, and most international scientific conferences). Language is the carrier of the cultural knowledge that creates a human mind as much as genetics carries the prescription for making a human body. Differences in culture and language are not likely to ever disappear, nor would it be desirable if they did. Diversity in cultural evolution is as important as it is in biological evolution. Prior art in machine translation exists in several different paradigms, rule-based translation, syntactic transfer, interlingua-based translation, table lookup translation, and statistical example-based translation. None of these approaches has yet produced domain-independent, fully unambiguous translation."

DARPA also lists a number of six areas in which it believes research should be concentrated. Research Concentration Areas: These are improved statistical language modeling techniques, improved techniques for collection and annotation of language corpora suitable for training statistical language processing algorithms, scaling techniques for extending statistical languagemodeling approaches to new domains with minimal training corpora; innovative approaches for speech-to-speech translation other than transcription, translation, and synthesis, machine learning techniques to incorporate human correction into automatic machine translation; and tool kits for rapid development of machine translation in new language pairs.

The research topic chiefs are named as Gary Strong, DARPA, 703-696-2259, gstrong@darpa.mil or Helen Gigley, ONR, 703-696-0407, gigleyh@onr.navy.mil. ■