Machine Translation

Alex Waibel

School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213

Machine Translation was one of the declared highlights and focal points of the Human Language Technology Workshop. Machine Translation, or MT for short, has seen a renaissance in recent years, brought about by the availability of faster and more powerful computing, and several decades of advances in speech and language processing. ARPA now sponsors a machine translation initiative and companies and governments, domestic and overseas, view it as a growth area and devote significant efforts to this problem. Indeed, as nations are growing increasingly intertwined with each other, and as economies, defense, travel, and the media grow increasingly internationalized and globalized, handling and overcoming language barriers effectively, becomes an ever more pressing issue.

The session at the ARPA workshop attempted to do justice to the various blossoming avenues that make up machine translationa as we see it today. First and perhaps foremost is the question of how MT is to be done. Two approaches have attracted considerable scientific interest and debate: the knowledge based and statistical approaches. The knowledge based approach is perhaps the more classical approach, based on linguistic theory and in its most purist incarnation relying on rule based systems for syntactic and semantic analysis and generation. The statistical approach, in contrast, attempts to achieve solutions to machine translation by finding suitable mappings between two languages via statistical analysis based on large corpora. Critcs of the former will argue that a knowledge based approach will lack the ability to make soft decisions, deal with uncertainty and ambiguity and cannot learn. Critics of the latter will see the lack of structure and simplicity of the statistical model as too simplistic and limited, given the intricacies and rich structure of language. The two views were highlighted and well argued by Ed Hovy and Peter Brown in two eloquent, enlightening as well as entertaining tutorials. Both cases were well delivered and the discussion that ensued highlighted perhaps the commonalities more than the differences. Indeed, as knowledge based approaches adopt statistical techniques and as proponents pf statistical MT discover structure in language, the two approaches appear to be growing toward a common middle ground. Learning and handling uncertainty as well as taking advantage of structural universals of human languages will guide progress in years to come. Along he way, better tools, better principles of evaluation and a better understanding of what the ultimate needs in MT would be will drive advances.

Tools, dictionaries and knowledgebases of various kinds make up important parts of the translation task (human or by machine). Short of academic squabbles over the "right" approach, a number of efforts are aiming to improve and expand these supporting technologies to achieve better quality translations more effectively and more efficiently by humans and/or machines. Acknowledging that accurate automatic translation of any unrestricted texts may still be a research item for a while, researchers attempt to develop tools that help the human translator in "Machine Aided Translation" (MAT), to do the job more effectively. Unlike speech recognition, a partial solution here can provide significant help or save costs. A paper by Kevin Knight entitled "Building a Large Ontology for Machine Translation" and by Peter Brown et al. entitled "But Dictionaries are Data too", address ways by which dictionaries and ontologies can be automatically or semi-automatically generated and how they can be applied and used in MT. The papers "LINGSTAT: and Interactive, Machine-Aided Translation System" by Jonathan Yamron et al. and "An MAT Tool and Its Effectiveness" by Robert Frederking et al. address the question of how tools for generating translated documents semi-automatically can improve effectiveness of translation.

In the light of these different streams of activity it is particularly difficult to define commonly useful and accepted evaluation procedures. Since there is no clear definition of a "correct" translation, it is not a simple matter of counting the number correct or error rate. Translation fidelity is subjective in part and is determined by various schemes in which panels of judges decide on the naturalness and intelligibility of translations. No doubt, the cost of performing such evaluations is considerable and different schemes are being discussed. The paper "Evaluation of Machine Translation" by John White et al. addresses this thorny issue and gives evaluation results using current evaluation measures used under the ARPA MT-program.

Finally, two papers on Speech Translation address the questions that arise, when text is not the input medium, but if an input sentence is spoken in one language and should be translated into another. Applications for this kind of MT system abound (telecommunication, media, conferences, etc.). The problem of translation is made harder by the fact that the input to the MT-system is now corrupted by syntactic ill-formedness produced by the speaker, colloquialisms, acoustic noise, and speech recognition error. While a speech translation system may at first glance combine speech-to-text recognition with text based machine translation, its long term viability demands a tighter coupling as translation and recognition need to derive the intended meaning, not a perfect textual transcription and need to involve conextual cues in a cross-lingual dialog. Attempts at answering some of these still open questions are under way and described in two papers: "Recent Advances in Speech Translation" by Monika Woszczyna et al. and "A Speech to Speech Translation

System built from Standard Components" by Manny Rayner et al. They describe currently operational speech translation systems.

In summary, a good number of the outstanding issues in Machine Translation have been touched by the papers presented at the workshop. It is our hope that they pave the way for a rich ongoing debate between proponents of different approaches, applications and deployment considerations to our collective benefit. Indeed, the academic efforts are well warranted by the urgent needs in an increasingly internationalized but linguistically splintered world.

Session 8: Machine Translation Summary

Machine Translation was one of the declared highlights and focal points of the Human Language Technology Workshop. Machine Translation, or MT for short, has seen a renaissance in recent years, brought about by the availability of faster and more powerful computing, and several decades of advances in speech and language processing. ARPA now sponsors a machine translation initiative and companies and governments, domestic and overseas, view it as a growth area and devote significant efforts to this problem. Indeed, as nations are growing increasingly intertwined with each other, and as economies, defense, travel, and the media grow increasingly internationalized and globalized, handling and overcoming language barriers effectively, becomes an ever more pressing issue. The session at the ARPA workshop attempted to do justice to the various blossoming avenues that make up machine translationa as we see it today. First and perhaps foremost is the question of how MT is to be done. Two approaches have attracted considerable scientific interest and debate: the knowledge based and statistical approaches. The knowledge based approach is perhaps the more classical approach, based on linguistic theory and in its most purist incarnation relying on rule based systems for syntactic and semantic analysis and generation. The statistical approach, in contrast, attempts to achieve solutions to machine translation by finding suitable mappings between two languages via statistical analysis based on large corpora. Critcs of the former will argue that a knowledge based approach will lack the ability to make soft decisions, deal with uncertainty and ambiguity and cannot learn. Critics of the latter will see the lack of structure and simplicity of the statistical model as too simplistic and limited, given the intricacies and rich structure of language. The two views were highlighted and well argued by Ed Hovy and Peter Brown in two eloquent, enlightening as well as entertaining tutorials. Both cases were well delivered and the discussion that ensued highlighted perhaps the commonalities more than the differences. Indeed, as knowledge based approaches adopt statistical techniques and as proponents pf statistical MT discover structure in language, the two approaches appear to be growing toward a common middle ground. Learning and handling uncertainty as well as taking advantage of structural universals of human languages will guide progress in years to come. Along he way, better tools, better principles of evaluation and a better understanding of what the ultimate needs in MT would be will drive advances. Tools, dictionaries and knowledgebases of various kinds make up important parts of the translation task (human or by machine). Short of academic squabbles over the right approach, a number of efforts are aiming to improve and expand these supporting technologies to achieve better quality translations more effectively and more efficiently by humans and/or machines. Acknowledging that accurate automatic translation of any unrestricted texts may still be a research item for a while, researchers attempt to develop tools that help the human translator in Machine Aided Translation (MAT), to do the job more effectively. Unlike speech recognition, a partial solution here can provide significant help or save costs. A paper by Kevin Knight entitled Building a Large Ontology for Machine Translation and by Peter Brown et al. entitled But Dictionaries are Data too, address ways by which dictionaries and ontologies can be automatically or semi-automatically generated and how they can be applied and used in MT. The papers LINGSTAT: and Interactive, Machine-Aided Translation System by Jonathan Yamron et al. and An MAT Tool and Its Effectiveness by Robert Frederking et al. address the question of how tools for generating translated documents semi-automatically can improve effectiveness of translation. In the light of these different streams of activity it is particularly difficult to define commonly useful and accepted evaluation procedures. Since there is no clear definition of a correct translation, it is not a simple matter of counting the number correct or error rate. Translation fidelity is subjective in part and is determined by various schemes in which panels of judges decide on the naturalness and intelligibility of translations. No doubt, the cost of performing such evaluations is considerable and different schemes are being discussed. The paper Evaluation of Machine Translation by John White et al. addresses this thorny issue and gives evaluation results using current evaluation measures used under the ARPA MT-program. Finally, two papers on Speech Translation address the questions that arise, when text is not the input medium, but if an input sentence is spoken in one language and should be translated into another. Applications for this kind of MT system abound (telecommunication, media, conferences, etc.). The problem of translation is made harder by the fact that the input to the MT-system is now corrupted by syntactic ill-formedness produced by the speaker, colloquialisms, acoustic noise, and speech recognition error. While a speech translation system may at first glance combine speech-to-text recognition with text based machine translation, its long term viability demands a tighter coupling as translation and recognition need to derive the intended meaning, not a perfect textual transcription and need to involve conextual cues in a cross-lingual dialog. Attempts at answering some of these still open questions are under way and described in two papers: Recent Advances in Speech Translation by Monika Woszczyna et al. and A Speech to Speech Translation System built from Standard Components by Manny Rayner et al. They describe currently operational speech translation systems. In summary, a good number of the outstanding issues in Machine Translation have been touched by the papers presented at the workshop. It is our hope that they pave the way for a rich ongoing debate between proponents of different approaches, applications and deployment considerations to our collective benefit. Indeed, the academic efforts are well warranted by the urgent needs in an increasingly internationalized but linguistically splintered world.