Keynotes

Is Machine Translation ripe for EU translators ?

Josep Bonet, DG Translation of the European Commission

Or, conversely, are EU translators ready for MT? MT has been in use in the EU for almost 20 years. Among the 28 language pairs available around a dozen can be utilised to one or another extent. But the rapid increase in the number of official languages excluded MT as an option ... until new data-drive systems made surface. The Google effect has generated enormous interest among (an increasing number of) translators. End-users of translations are even more excited about MT at times when translation needs grow exponentially and provision of high-quality human translation is capped by budgetary constraints. How can translators help the end-user get a better service while helping themselves is a challenge to be addressed. Can translators accept good enough as the result of their work? Is it possible to move from computer-assisted human translation to human-assisted computer translation? In this presentation, such questions will be debated and the roadmap chosen by the European Commission's Directorate-General for Translation to re-introduce MT to cover all official languages will be described.

Hierarchical Phrase-based Translation with Weighted Finite State Transducers

William Byrne, Cambridge University

I will present recent work in statistical machine translation which uses Weighted Finite-State Transducers (WFSTs) to implement a variety of search and estimation algorithms. I will describe HiFST, a lattice-based decoder for hierarchical phrase-based statistical machine translation. The decoder is implemented with standard WFST operations as an alternative to the well-known cube pruning procedure. We find that the use of WFSTs in translation leads to fewer search errors, better parameter optimization, and improved translation performance. We also find that the direct generation of target language lattices under Hiero translation grammars can improve subsequent rescoring procedures, yielding further gains with long-span language models and Minimum Bayes Risk decoding.

Resources for adding semantics to machine translation

Jan Hajič, Charles University in Prague

Current (Statistical) Machine Translation systems rarely go beyond morphology, lemmatization, phrases or syntax. One of the possible ways to direct research in the

near future is use semantics in one way or the other, whether as semantics features or factors within the successful phrase-based or hierarchical systems, or in hybrid systems, or otherwise. However, semantic features have to be learnt from annotated data, at least until unsupervised learning can replace all the expensive annotation projects. In the talk, I will present the basics of the family of Prague dependency treebanks (currently available for Czech, English and Arabic), which to various extents provide combined manual annotation of syntax and semantics based on the dependency framework, but general enough to be used in systems of all types, including the classical non-hierarchical SMT systems where only word-based features can be incorporated into the model. One of the corpora available is specifically aimed at machine translation, since it is a parallel, fully manually annotated Czech-English corpus, which consists of the Penn Treebank texts (preserving also the original annotation) and its professional translation to Czech. Specific resources aimed at spoken language analysis will also be presented, even though no parallel version exists yet. These are based on the "speech reconstruction" idea by Fred Jelinek and his students, which was incorporated into a dialog corpus of Czech and English that was then developed at Charles University.

The Quaero program: Multilingual and multimedia technologies

Jean-Luc Gauvain, LIMSI/CNRS

The goal of the Quaero programme is to promote research and industrial innovation for multilingual multimedia content processing and information management and access. The Core Technology Cluster (CTC) groups the research activities in Quaero which aim improve the state-of-the-art in automatic multimedia document structuring and indexing, and to develop and evaluate the core technologies. The core technologies cover text processing, translation, audio and speech processing, image and video processing, audio and video fingerprinting, cross-modal processing, and search and navigation methods. Generic technologies are being developed that can be applied to a wide range of documents, along with tools to enhance portability. This talk will overview the core technologies being developed and evaluated in Quaero, with an emphasis on multilingual language technologies (NE, Q&A, MT, STT, SPKR). The talk will conclude with some examples cases of the CTC research being incorporated in application prototypes. The main applications areas are general search engines for multimedia documents; portals for personal multimedia document management; online services to access audiovisual archives and digital libraries; advanced tools for the content production and management chain; and highly personalized content-on-demand services.