Machine Translation and Multilingual Technologies

Machine Translation and Multilingual Technologies

Machine translation is an attractive interdisciplinary task relying on computer science, linguistics, statistics and mathematical modeling, with specific challenges to software engineering – as the volume of the data processed typically reaches tens of billions of words. At the Institute of Formal and Applied Linguistics, several prototypes of machine translation systems have been created and implemented. Two approaches have been adopted.

First, phrase-based statistical translation which uses machine learning based on very large text corpora; second, a deep linguistic analysis of whole sentences is used. Combination of both approaches, e.g. in a form of grammatically-based corrections of the phrase-based translation output, seems highly promising. In addition to text-to-text machine translation, cross-language search which operates on multimodal data and multilingual dialogue systems enabling natural human-computer interaction are being developed. Translation systems developed at the Institute focus especially on translation between the Czech language and other European languages (esp. English, German, French and Spanish); however, a phrase-based translation system can be quickly adapted for any language pair for which a sufficient amount of language data can be obtained. Further development of machine translation methods and systems is being pursued in collaboration with many academic and commercial partners within several European, U.S.-based and national projects as well as in direct cooperation with industry.

Research in Machine Translation and related text and speech analysis areas is supported by a Research Infrastructure LINDAT/CLARIN (http://lindat.cz), which is part of the European Clarin ERIC network. The infrastructure collects and prepares open language resources necessary for all areas of research in computational linguistics and natural language processing. It also provides open tools and services for both fundamental as well as applied research in this dynamic area.

Selected outputs
