Achievements of AI in Linguistics

Anna Sågvall Hein

anna.sagvall-hein@convertus.se

Summary

AI aims at simulating human intelligence by computer. Language is one of the primary expressions of human intelligence. It has been claimed that the acquisition language is the greatest intellectual achievement of man. Natural language processing is, naturally so, one of the major fields of AI, and to a large extent it overlaps with Computational Linguistics, an important and continuously growing field of linguistics.

The fundamental problem of computational linguistics is the modelling of the basic linguistic processes – comprehension, production and learning of language. It includes central AI problems such as perception, communication, knowledge, planning, reasoning and learning. On the application side, information retrieval including text mining and machine translation, currently, seem to dominate.

By tradition, the term Computational Linguistics refers to written language, whereas Speech Technology is used for the analysis and synthesis of spoken language. In later years, Language Technology has emerged as a common denominator of both modes of language. The division into Speech Technology and Computational Linguistics was mainly based on different scientific traditions and methods. Whereas statistical methods dominated in speech technology, so did symbolic processing in computational linguistics. However, the statistical methods originally developed for the analysis of speech were found to be more or less directly applicable to the fairly new paradigm of data-driven machine translation. Below, machine translation will be focused.

Intuitively, translation may be understood as comprehending a source language text and producing an equivalent text in the target language. In terms of AI, or CL, it would imply modelling the comprehension process and the production process, respectively.

Assuming that the comprehension process would result in a complete, language independent meaning representation, an interlingua, this representation might be the starting point of the production process, and there would be no need for a specific translation step. In such an approach, knowledge becomes a central issue. How to identify it and how to represent it? This ideal approach to machine translation, the interlingua approach, has been explored by many researchers with interesting findings, but no viable translation system, as a result. The Google translation service is as far from the interlingua model that one may get, since it does not use any refined knowledge of language at all.

AI researchers in the 70-ies, such as Roger Schank, Terry Winograd and Yorick Wilks, made substantial and innovative contributions to the exploration of language comprehension and production, illustrating the relevance of knowledge, in particular world knowledge, planning, and reasoning in the use of language. Their approaches were exclusively symbolic, excluding, basically,. machine learning of language. Currently, the data-driven approach dominates the field of machine translation, and attempts are made to combine or complement it with rule-based methods in order to overcome inherent limitations with the approach.

Basic problems of machine translation are due to lexical ambiguity and variation, and grammatical differences in morphology and word order. In the rule-based paradigm, the problems are approached by linguistic analysis sorting out the different uses, hence translation options, of the words. The linguistic analysis also is also the back-bone in creating an appropriate word order. The data-driven, or statistical approach, relies on reuse of large amounts of previously translated text, parallel data.

A problem with the rule-based approach, RBMT, is the coverage of language resources, dictionaries and grammars. Typically, coverage is insufficient and methods have to be found to handle text outside dictionaries and grammars. This is where statistical methods may help.

In statistical machine translation, SMT, the choice of translation alternatives is based on a comparison of the source text with large amounts of parallel data. The fundamental problem is access to sufficiently large amounts of parallel data. The amount needed depends, among other things, on the language-pair in focus. Integrating linguistic data aims at reducing the amount of parallel data needed for quality translation.

Translation quality is a critical issue in machine translation. However, demands on quality vary with the purpose of the translation. Browsing quality, understandability, may be sufficient in certain contexts, whereas publishing quality must approach the quality of human translation. In addition to human evaluation, costly and time-consuming, different metrics for the automatic evaluation of machine translation have been proposed. They assume access to reference data in terms of previous translations of high-quality, typically human translations.

Success factors of machine translation in practice concern adaptation and customisation of the machine translation system to the needs of the user, pre-processing of the source text (language checking, controlled language), and post-processing of the target text. A research issue of great concern is automating the post-processing process.

As an example of machine translation in use, the *Convertus Syllabus Translator*, CST, will be presented. CST translates course syllabi from Swedish to English and is in use at six Swedish universities. The presentation will include the following aspects:

- Background and development
- Methodology
- Translation dictionary
- Spell-checking
- Post-processing interface and translation memory function
- Automatic post-processing
- Scenarios of use
- User feed-back
- Future development and deployment

Finally, further development and prospects of machine translation will brought up for discussion.