Reliving the History: The Beginnings of Statistical Machine Translation and Languages with Rich Morphology

Jan Hajič

Institute of Formal and Applied Linguistics, School of Computer Science
Charles University, Prague, Czech Republic
hajic@ufal.mff.cuni.cz

Abstract. In this two-for-one talk, first some difficult issues in morphology of inflective languages will be presented. Then, to lighten up this linguistically and computationally heavy issue, a half-forgotten history of statistical machine translation will be presented and contrasted with current state-of-the art (in a rather non-technical way).

Computational morphology has been on and off the focus of computational linguistics. Only few of us probably remember the times when developing the proper formalisms has been in such a focus; a history poll might still find out that some people remember DATR-II, or other heavy-duty formalisms for dealing with the (virtually finite) world of words and their forms. Even unification formalisms have been called to duty (and the author himself admits to developing one). However, it is not the morphology itself (not even for inflective or agglutinative languages) that is causing the headache – with today's cheap space and power, simply listing all the thinkable forms in an appropriately hashed list is o.k. – but it's the disambiguation problem, which is apparently more difficult for such morphologically rich languages (perhaps surprisingly more for the inflective ones than agglutinative ones) than for the analytical ones. Since Ken Church's PARTS tagger, statistical methods of all sorts have been tried, and the accuracy of taggers for most languages is deemed pretty good today, even though not quite perfect yet.

However, current results of machine translation are even farther from perfect (not just because of morphology, of course). The current revival of machine translation research will no doubt bring more progress. In the talk, I will try to remember the "good old days" of the original statistical machine translation system Candide, which was being developed at IBM Research since the late 80s, and show that as the patents then filed gradually fade and expire, there are several directions, tweaks and twists that have been used then but are largely ignored by the most advanced systems today (including, but not limited to morphology and tagging, noun phrase chunking, word sense disambiguation, named entity recognition, preferred form selection, etc.). I hope that not only this will bring some light to the early developments in the field of SMT and correct some misconceptions about the original IBM system often wrongly labeled as "word-based", but perhaps also inspire new developments in this area for the future – not only from the point of view of morphologically rich languages.

H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.): IceTAL 2010, LNAI 6233, p. 1, 2010.