The new machine translation: getting blood from a stone

Martin Kay (Stanford University and Universität des Saarlandes)

Lecture at Workshop on Free/Open-Source Rule-Based Machine Translation,
June 2012

Summary

There is now a new sense of excitement in the air about machine translation. After fifty years of unfulfilled promises by linguists, the field has been taken over by computer scientists and reconstructed on scientific principles. A machine translation system requires massive amounts of data. Painstaking work with native informants, and playing examples off against counterexamples, takes far too long and is too unreliable. We now extract the massive amounts of data from massive quantities of naturally occurring text by sophisticated machine-learning techniques. If you doubt the value of this approach, you have only to look at Google Translate. We should be thankful for this new turn of events because massive amounts of data and sophisticated machine-learning techniques have a vital role to play in machine translation. But, as I will show in this talk, they are not enough to finish the job because much of the information required to build a creditable translation system cannot be extracted from examples, even in principle, however massive the number of them that one collects or how sophisticated the techniques one applies. It cannot be extracted because it is not there to be extracted. As my mother would say: "You cannot get blood from a stone".