Hunglish: a statistical Hungarian–English Machine Translation system

Péter Halácsy*, András Kornai**, László Németh*, András Rung*, István Szakadát*, Viktor Trón***, Dániel Varga*

This paper lays out our plans for a simple Hungarian to English machine translation system and describes our accomplishments so far. In the preparatory stage we collect a parallel corpus (so far, we have collected about a quarter of the planned 100m words) and use this to verify a pre-existing, but in many ways overly broad, Hungarian–English dictionary.

In the simplest version of the planned system, we use the resulting dictionary to create a lattice of translation possibilities, and find the best path through the lattice by Viterbi search. Since Hungarian word order is kept in the translation, the result is not expected to be fully grammatical, let alone idiomatic, English, though it is already expected to be serviceable for cross-language information retrieval, where word order is typically ignored.

More complex versions of the system, which analyze the source in greater detail (NP-level chunks and predicate-argument structure) and transfer this analysis to the target, will be built incrementally on top of the simpler system. As with other NLP projects at the Budapest Institute of Technology Media Research and Education Center, all dictionaries, corpora, and software created in the project will be made available under a non-restrictive (LGPL) open source license.

* Budapesti Műszaki Egyetem Média Oktató és Kutató Központ, {hp, nemeth, runga, szakadat, daniel}@mokk.bme.hu

** MetaCarta Inc., andras@kornai.com

*** International Graduate College, Saarland University and University of Edinburgh, v.tron@ed.ac.uk