

Combining Outputs from Machine Translation Systems

By Fahim A. Salim

Supervised by: Ing. Zdenek Zabokrtsky, Ph.D

Institute of Formal and Applied Linguistics, Charles University in Prague 2010.

Abstract:

Due to the massive ongoing research there are many paradigms of Machine Translation systems with diverse characteristics. Even systems designed on the same paradigm might perform differently in different scenarios depending upon their training data used and other design decisions made. All Machine Translation Systems have their strengths and weaknesses and often weakness of one MT system is the strength of the other. No single approach or system seems to always perform best, therefore combining different approaches or systems i.e. creating systems of Hybrid nature, to capitalize on their strengths and minimizing their weaknesses in an ongoing trend in Machine Translation research.

But even Systems of Hybrid nature has limitations and they also tend to perform differently in different scenarios. Thanks to the World Wide Web and open source, nowadays one can have access to many different and diverse Machine Translation systems therefore it is practical to have techniques which could combine the translation of different MT systems and produce a translation which is better than any of the individual systems. Since output combination is an additional step over actual translation, therefore it should be very resource and time efficient to be practically usable, and these techniques should also work only on individual system output without bothering much about how the translation was generated because such information is usually not available.

This thesis investigates system output combination techniques. The focus is on techniques which are individual system and language pair independent and are efficient enough to be usable in variety of application scenarios.