Statistical MT from TMI-1988 to TMI-2007: What Has Happened?

Hermann Ney
Professor of Human Language Technology and Pattern Recognition,
RWTH Aachen University
Ahornstr. 55
DE-52056 Aachen, Germany

ney@informatik.rwth-aachen.de

When Peter Brown of IBM research presented a statistical approach to French—English MT at TMI 1988 at CMU, the audience was shocked because this approach was a slap in the face for the then received MT theories. At the time of TMI 2007, nearly two decades later, the statistical approach seems to be the mainstream approach in MT research.

Since the first approach to statistical MT had been worked out by IBM for French—English translation, many attempts have been made to push the state of the art and to improve the translation accuracy.

Statistical MT systems are now able to translate across a wide variety of language pairs and translation tasks. The statistical approach forms the basis for many recent and ongoing large-scale MT projects like the EU-funded TC-Star project and the US-DARPA-funded GALE project. In both projects, statistical MT is extended from text input to speech input.

Today, a typical state-of-the-art statistical MT system has the following four components:

1. Training: For each sentence pair in the training data, an alignment matrix is computed, typically by using the set of IBM-1 to IBM-5 alignment models and a Hidden Markov model.

2. Phrase extraction: From the alignment matrices of all training sentence pairs, source-target fragments are excised and used to define the so-called phrase tables.

3. Definition of the log-linear model: For each source-target phrase pair in the phrase table, so-called scoring functions are defined. Based on the training data, these scoring functions compute a probabilistic score of the hypothesis that the source fragment and the target fragment under consideration are translations of each other. These scoring functions are complemented with a word and/or phrase re-ordering model. All these scoring functions are combined in a so-called log-linear model. The weight of each scoring function is tuned for optimal translation quality or a related criterion.

4. Generation or search: For the given source sentence, the goal is to select the target sentence with the highest probabilistic score in the log-linear model. To

this purpose, the search algorithm has to generate and score hypotheses along various dimensions: unknown segmentation of the source sentence, unknown target phrases and unknown order of these phrases in the target sentence.

This talk will review the details of these components and the progress that the field has made so far and will also compare the statistical approach with example- and memory-based approaches.