A Linguistically Enriched Translation Memory

Gábor Hodász

Faculty of Information Science
Pázmány Péter Catholic University
Budapest

hodasz@morphologic.hu

Tamás Grőbler

MorphoLogic Ltd.

Budapest

grobler@morphologic.hu

Keywords: translation memory (TM), machine translation (MT), computer-aided translation (CAT)

Translation memories (TMs) offer translations based on previously stored translation units. Traditionally, translation units are sentence pairs, and the search method is based on character-level similarity. Most existing systems recognise "do not translate" elements such as dates or numbers and are equipped with terminology databases. Current TM systems, however, fail to handle morphological differences or syntactically similar sentences that are very different on the character level. We propose a TM system that applies the machinery of a rule-based machine translation (RBMT) system to compose the target sentence from the stored sub-sentential translation units.

MetaMorpho TM, a linguistically enriched TM system, uses MorphoLogic's fine-grained language technology in both languages to yield more translations that are also more exact matches to the source sentence. In the current version of the system, translation units are noun phrases (NPs) and the sentence skeletons that incorporate them. The stored NPs consist of morphologically analysed terminal symbols. Sentence skeletons contain the NP slots and the rest of the sentence as analysed terminal symbols. The engine is being integrated with a word processor to allow translators to select the best match and prepare the translation with minimal interaction.

Database lookups are performed by the MetaMorpho machine translation system. Translation is performed at the same time as syntactic parsing. When applied to the translation memory, MetaMorpho interprets translation units as highly lexicalised rule pairs and attempts to compose the target sentence from them. As the representation of the MT and the TM samples are compatible, our RBMT system also benefits from using the memory building mechanisms in-house.

The actual memory is implemented as a relational database. To further increase the number of matches, the database is also indexed using a "linguistic similarity" measure. Automatic addition of new items to the database is available both during translation and from a parallel corpus. The latter will be supported by the aligner software to be developed as a separate module.

In the paper, we show examples to demonstrate how sentences are translated, how new samples are added to the database, and how they are reused in another translation.