

WHITE PAPER

SMT by Lingtech and LanguageLens: Technical Background

The SMT system presented by LanguageLens and Lingtech performs automatic translation, using statistical information. The statistical approach constitutes a major breakthrough in translation technology, characterized by:

- low cost
- high quality
- high speed
- easy portability to new languages and text types

Setting Up the System

The system requires a large amount of translated material, preferably produced by skilled translators. Typically, the translation data is taken from a specific text type, and consists of several million words. The system examines each word sequence in the translation data, and compiles a *translation list* containing all the translations observed in the data. The translation list is similar to a traditional bilingual dictionary – it differs, however, in two important ways. First, while a traditional dictionary consists primarily of individual words, the translation list includes a large number of multi-word sequences. Secondly, the translation list computes a probability score for each translation choice, based on the number of occurrences in the translation data. These probabilities are the key to the success of the system – they allow translations to mimic the subtle, complex preferences observed in the translation data.

Translating with the SMT System

The SMT system is accessed via a simple web-based interface. The user uploads the document to be translated, and, in seconds, a link to the complete translation appears. The system accepts both Microsoft Word documents and web pages, and the resulting translation preserves all the formatting information that was in the file to be translated. In addition to the translation itself, the system provides a *Term Consistency* report, identifying all relevant words sequences that received two or more different translations at different parts of the document. This supports the user in finding potential inconsistencies, while producing the final, corrected translation.

How the System Works

To translate a sentence, the system consults the translation list to determine all possible translations for each sub-part of the sentence. This gives the system a large *search space* of possible translations – the system examines thousands of options for each sentence, consistently arriving at a high-quality translation in a matter of milliseconds.

A key factor in the success of the SMT system is the flexibility of its translation lists. These lists correlate not only individual words, but also multi-word sequences of varying sizes. This effectively liberates the system from the plodding word-by-word approach thought to be characteristic of automatic translation. Another essential feature of the system is that its choices are relativized to context: a potential translation will often be rejected because it doesn't fit smoothly with what comes before or after. This flexibility and context-sensitivity enables the system to produce translations that are remarkably faithful to its translation data, not only in terms of grammatical correctness, but also in terms of the style and idioms of a given text type.