Terminology extraction as a tool for MT output assessment and improvement

Proposition of a new metric method, called the Terminology Recall Index, for evaluating more accurately the edit distance between two translated documents.

Presented by Jean-François Richard
President, Terminotix
jfrichard@terminotix.com
Context and assumptions

• Averagely, between 30 to 40 percent of linguists’ time is dedicated to searching tasks.

• Searches are launched against different sources for all sorts of reasons.

• Concepts and noun groups are the query types that is the most time consuming for linguists to get results for.

• Sentences comparison algorithms (BLEU score, fuzzy matches, post-edit effort, edit distance, etc.) are not concept aware. The real translation effort calculation is theoretically biased.
Term extraction of noun groups

• Statistical term extractors useful for quickly identifying repetitive phrases or sub-segments. Quite noisy.

• Taggers useful for identifying words’ grammar information and for stemming. Needs to be filtered for POS patterns.

• Feed taggers with phrase extractions to obtain a “Semantic” term extraction engine.
TRI calculation

• Sum total occurrences of noun groups.
• Collect occurrences of cross checked noun groups.
• Calculate TRI.
TRI applications

• Comparing a human translated document with a MT translated document.
• For benchmark purposes, allows to identify which MT is best.
• For identifying which translation memory should be used.
• For creating a job glossary.
• For identifying unknown noun groups.
• For determining the domain for automatic task assignments.
• For QA purposes at the end of the translation process.
• In building an ontology model for automatic document archiving.
• Automatic text summarization.
Limitations

- Retains only noun groups comprising two words or more.
- Adding some specific languages could represent a tedious task.
- Tested with a limited set of languages.