ACCEPT

COOPERATION

# Work package 9: MT Evaluation

## Description of work

**Description of work**

The study of machine translation has in recent years developed a number of automated-metrics that are well known in the literature. We will use a selection of these metrics including BLUE, GTM, Meteor and TER. These metrics calculate by various mathematical criteria the differences between a machine translation and a reference translation. In this study we will be calculating the performance of the machine translation in reference to the post-edited translation. These automatic metrics will give an estimate of the amount of effort required by the post-editing process to improve the initial machine translation.
However effective automated metrics might be, they are certainly cost-effective and fast to measure and precise, they may or may not reflect the sentiment of the user community.
We will also deploy human evaluation assessments with translators and users; these evaluations ask a group of "typical users" and translators or students in translation to assess the translations for fidelity and comprehensibility. Sample sentences are selected from the corpus and are presented to the evaluator in random order. Evaluators are asked to assess these sentences on a scale, usually 1 to 4 or 1 to 5, where 1 is deemed poor and 5 deemed excellent in terms of comprehensibility. In the case of fidelity, bi-lingual evaluators judge the fidelity of the translation with respect to its source language on a binary scale. Source text is also evaluated in this manner. These human evaluations are used in conjunction with the automated scoring mechanism, the purely subjective, balancing the purely objective, to guide the development of the machine translation technology.

Task 9.1: Evaluate the impact of pre-editing rules on SMT (UNIGE, acrolinx, month 12-18)
Compare the SMT translation with and without various types of pre-editing rules (e.g. rules regarding the style, syntax, orthography) using translator's judgments and automatic metrics (BLUE, GTM, Meteor and TER). We will also examine the quality of SMT when trained on in-domain corpora of various (increasing) sizes in order to determine the impact of domain knowledge and terminology lists availability.

Task 9.2: Evaluate the impact of bilingual and monolingual post-editing rules on translation quality (UNIGE, acrolinx, SYMANTEC, month 12-24)
Evaluate the result of SMT with bilingual and monolingual post-edition using end-users' judgments.

Task 9.3: Assess users' judgments vs. translator's judgments and determine MT task tolerance (UNIGE, month 18-24)
Conduct user surveys to evaluate the improved post-edited SMT with anonymous user's judgments (native target language speakers who have no special knowledge of the source languages) and compare it with translators' judgments in order to assess the reliability of the users' ratings. We will also study the occurring MT errors (both linguistic and non-linguistic) that are tolerable for the particular task in the ACCEPT project.

Task 9.4: Compare post-edited RBMT with post-edited SMT using automated metrics (UNIGE, acrolinx, month 12-30)

First, we will specialise Rule-Based Machine Translation (RBMT) systems such as Systran and Lucy for the ACCEPT project data domains and compare the results with SMT using various automatic metrics (e.g. translation error rate to count the number of required post-edits). Second, we will analyze which automatic metrics correlate with translators' judgment in order to assess their usefulness, adequacy and reliability on predicting the MT ability.

Task 9.5: Statistical analysis and hypothesis testing (UNIGE, month 18-36)
Assess the significance of RBMT vs. SMT performance differences using statistical analysis, significance

Task 9.6: Evaluate how users interact with pre- and post-editing rules (acrolinx, month 24-30)
User decisions to accept or ignore automatically detected errors will be logged and evaluated. Logging of user decisions will make use of the reporting tool of the acrolinx IQ system. This approach will help to find out what kind of errors the users are willing to correct and what kind of changes are seen to be unnecessary and not useful. The results will influence the rule weighting, such that the application of rules is adapted to the user feedback.

| **Deliverables** | **Delivery date** | **Description** |
|---|---|---|
| D 9.1 | PM 12 | Analysis of existing metrics and proposal of a task-oriented metric. |
| D 9.2.1 | PM 18 | Survey of evaluation results – Version 1. |
| D 9.2.2 | PM 24 | Survey of evaluation results – Version 2. |
| D 9.2.3 | PM 30 | Survey of evaluation results – Version 3. |
| D 9.2.4 | PM 36 | Survey of evaluation results – Version 4. |
| D 9.3 | PM 30 | Weighting of pre-editing rules. |