ACCEPT

SEVENTH FRAMEWORK PROGRAMME

THEME ICT-2011.4.2(a)

Language Technologies

ACCEPT

Automated Community Content Editing PorTal

www.accept-project.eu

Starting date of the project: 1 January 2012

Overall duration of the project: 36 months

Definition of post-editing rules for English, French, German and Japanese (draft version)

Workpackage n° 2	Name: Definition of Pre- Editing and Post- Editing rules
Deliverable n° 2.3	Name: Definition of post-editing rules for English, French, German and Japanese (draft version)
Due date: 31 August 2013	Submission date: 26 August 2013

Dissemination level: PU

Organisation name of lead contractor for this deliverable: Acrolinx

The research leading to these results has received funding from the European Community's Seventh Framework Programme (*FP7/2007-2013*) under *grant agreement n° 288769.*

SEVENTH FRAMEWORK PROGRAMME

Contents

Definition of Post-Editing Rules for English, French, German and Japanese (Draft Version)

Objectives of the Deliverable

The main goals of this deliverable are to describe the expected requirements and specifications for the development of post-editing rules in *Task 2.2: Post-editing rules for MT*. This task started in month 19 and will continue to the end of the project.

Post-Editing Rules

The goal of *Task 2.2: Post-editing rules for MT* is to create post-editing rules that support the community activity of improving the quality of machine-translated user-generated content. These rules should be created for and used by the Acrolinx software.

Post-editing rules thus complement pre-editing rules, which already improve the input text to reduce out-of-domain problems such as wrong or uncommon spelling. While pre-editing rules prepare the input to make it more similar to the SMT training corpus, post-editing rules aim at improving the quality of the MT output by directly addressing the shortcomings of SMT translations.

Requirements

This section lists requirements for the post-editing rules that we plan to develop.

Addressing Quality Issues

The purpose of a post-editing rule is to help authors improving the MT output's comprehensibility and correctness. While the Acrolinx software is good at improving the writing style and finding concise formulations, the challenge is that the output of SMT systems is often deficient in other ways. In particular, we expect three types of errors:

1. Tokenization issues, including: use of upper and lower case, punctuation, hyphenation;
2. Lexical/syntactic issues, including: word order, agreement issues, use of wrong word in the given context, untranslated terminology;
3. Semantic problems, including: input text parts that are completely missing in the translation, change of polarity from negative to positive or vice versa.

While some of these issues can be corrected using standard Acrolinx spelling and grammar rules, we expect that many of them cannot be defined in terms of linguistic patterns in the translated output. Instead, we expect them to be highly dependent on the probabilities of the translation and language models. A rule-based post-editing approach would thus need to take the source text into account, and cannot solely rely on the translation.

Reducing the Post-editing Bottleneck

Traditional approaches to post-editing are not particularly helpful in the community context, since the need for bilingual expertise is a major bottleneck. Experts who could potentially correct machine-generated content in areas such as healthcare or engineering are not necessarily bilingual, and translators with the critical subject-matter expertise and knowledge of two languages are generally in short supply.

To meet the pressing need to leverage MT effectively for this kind of content, we need to develop post-editing rules whose application does requires little or ideally no understanding of the source text, thus significantly increasing the pool of candidates for editing.

Learning from Post-editing Data and Giving Feedback to MT System

For the development of post-editing rules, our starting-point should be the behaviour of human post-editors. This includes both deriving patterns from manual post-editing samples, and also improving rules by gathering data on how they are applied in specific contexts.

The gathered data should also provide feedback to the MT system. Commonly mistranslated words, for example, should be collected to improve the MT system directly.

Development and Evaluation of Post-editing Rules

This section describes the two lines of work that will be pursued by the WP participants.

Learning Rules from Post-editing Data

Having identified the challenge that many required post-editing actions are lexical and do not follow larger linguistic patterns, we will follow the approach of learning rules from existing (manual) post-editing data.

Our initial experiments have been based on small samples of post-edited text created in the course of evaluation exercises such as the one described in Gerlach et al. (2013). We extract local editing patterns by taking differences between translated sentences and the variant forms produced by post-editing, if necessary combining elementary editing operations (insertions, deletions and substitutions) into larger operations such as exchanges. Each editing operation is associated with a surrounding context, which at present is a window of 0 to 2 words in each direction. The set of <edit-operation, context> pairs is then sorted to find high-frequency elements, where frequency is measured both in the small post-edited corpus and in a large monolingual corpus created by translating the source-language training corpus into the target language. The highest-frequency elements are then reviewed by human judges.

This method is certainly able to find at least a few useful post-editing rules; the human reviewing process is quick and efficient. The small amount of post-edited data however makes it difficult to form a clear idea of how much can reasonably be expected. Our first priority is thus to collect larger amounts of data. We will do this using the Amazon Mechanical Turk (AMT). The enhancements required to interface AMT to the ACCEPT post-editing portal have been implemented under WP5, and a pilot study carried out to find AMT workers capable of carrying out FR-EN post-editing. Our next goal will be to use AMT to collect a post-editing corpus of at least 5K sentences of FR-EN Symantec Forum data. At a later stage, we will also use the TWB data collected in WPs 8 and 9; our impression from the initial studies is that it is considerably easier to learn rules from the Symantec data, partly because it is a more constrained domain.

In parallel with data collection, we will enhance the initial version of the rule-extraction code. The most important enhancement will be to make it possible for rules to include a source-language context, which will be derived from SMT alignment information. We will also experiment with including POS information derived from word tagging, making it possible to generalize rules to tag-patterns; this technique will however be used sparingly, given the focus on lexical rules. Another obvious possibility is to try using automatic metrics like BLEU to identify promising rules (Kjellin 2012).

When we have acquired enough data to be able to carry out more substantial experiments, we will also feed the information back into the SMT training process (WP 4) and compare the results with the ones we obtain from rule-based post-editing. This work will continue the strand we have begun with comparisons of rule-based and statistical methods in pre-editing (Rayner et al. 2012; Bouillon et al. 2013).

This line of work will be led by the University of Geneva.

Translation QA

In previous work in the work package, we occasionally found fundamental quality problems in the SMT output because of parts of the input text that "got lost" during a translation. Two common patterns are:

- The SMT system fails to generate a long-distance dependency present in the source, such as between an auxiliary and a main verb when translating from English to German.
- The polarity of a sentence reverses because negation words or particles get lost or are introduced in the translation.
- Content-carrying phrases, such as proper nouns and domain-specific vocabulary, are not translated into the recommended terminology, or not translated at all.

Naturally, the correction of such issues should precede the more cosmetic post-editing actions. It is not useful to improve the writing style of the MT output if parts of the meaning of the sentence got lost. At the same time, such issues are among the most cumbersome to correct, as the post-editor needs to understand the source text to find out which part of the meaning has changed. Therefore, providing support in these cases may lead to big gains in post-editing effectiveness.

To address these issues, we plan to use Acrolinx rules to collect linguistic information from both the source and the target text, such as terminology, verbs, and negation words. We will then relate these linguistic features to find out whether certain parts are missing in the translation. This could include terms that are not translated according to accepted bilingual terminology, verbs that do not exist in the translation, or a mismatch in the number of negative-polarity words.

The post-editor is then pointed towards the respective source text phrases that may have been translated incorrectly. The aim is to help the post-editor identify and fix the most fundamental issues first, without the need to understand the source text deeply.

This line of work will be led by Acrolinx.

Evaluation of New and Existing Post-editing Rules

To guide the development process, the developed post-editing rules will be accompanied by small-scale evaluations that measure their effectiveness in terms of reduction of post-editing effort. These evaluations will include a manual inspection of the effectiveness of the rules.

We will carry out larger evaluations at the end of the task in collaboration with WP9. The evaluations will include both automatic metrics and manual judgements. For the automatic metrics, we use the reference translations that we also used to find effective pre-editing rules. For manual judgements, we will conduct internal evaluations, as well as rely on Amazon Mechanical Turk experiments, as we did for pre-editing rules. Moreover, we will be gathering feedback from communities in collaboration with WPs 7 and 8.

At the same time, a smaller study will be carried out to investigate to what extent existing Acrolinx rules reduce the post-editing time. Although we do not expect these rules to correct many issues due to the reasons described above, we may nevertheless be able to identify a subset of rules that provide reliable results, such as rules for punctuation, hyphenation and casing issues.

Deviation: English-Japanese Post-editing Rules

Due to unforeseen personal changes within the project (departure of Dr. Melanie Siegel), the project has found itself without the resources and expertise needed to develop EN-JP post-editing rules and evaluate their impact on Japanese MT output. We have consequently decided to focus instead only on pre-editing rules in the EN-JP pair, which require far less Japanese expertise. A student working for Acrolinx is performing this work within Task T2.2, and the results will be published in deliverable D2.4.

Timetable

The following timetable shows when the planned work will be carried out. Note that while task 2.2 continues to the end of the project (month 36), the final deliverable of this work package (D2.4) is already due in month 30. The main development work will therefore finish before that deliverable.

M21-M26:

- Learning rule from post-editing data
- Translation QA
- English-Japanese pre-editing rules

M27-M30:

- Manual and automatic evaluation of new and existing post-editing rules

M31-M36:

- Adjustment of developed post-editing rules for use in practice.

References

- Pierrette Bouillon, Johanna Gerlach, Ulrich Germann, Barry Haddow and Manny Rayner: Two Approaches to Correcting Homophone Confusion in a Hybrid Machine Translation System.
  In *Proceedings of Second ACL Workshop on Hybrid Approaches to Translation (HyTra*), Sofia, Bulgaria, 2013.
- Johanna Gerlach, Victoria Porro, Pierrette Bouillon and Sabine Lehmann:
  Combining pre-editing and post-editing to improve SMT of usergenerated content.
  In *Proceedings of 2nd Workshop on Post-Editing Technologies and Practice*, Nice, France, 2013.
- Martin Kjellin:
  Automatic Generation of Post-Editing Rule Sets from Parallel Corpora.
  Bachelor's Thesis, Uppsala University, 2012.
- Manny Rayner, Pierrette Bouillon and Barry Haddow:
  Using Source-Language Transformations to Address Register Mismatches in SMT.
  In *Proceedings of AMTA,* San Diego, California, 2012.