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

Exploitation Plan Update

Workpackage n° 10	Name: Dissemination and Exploitation
Deliverable n° 10.7	Name: Exploitation Plan Update
Due date: 31 December 2013	Submission date: 19 December 2013

Dissemination level: PU

Organisation name of lead contractor for this deliverable: Lexcelera

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.*

Contents

Exploitation Plan Update

Objectives of the Deliverable

This deliverable updates and refines the exploitation plan presented in December 2012 in deliverable D 10.7, namely it defines the project's exploitable results, target groups, commercial and academic exploitation, and sustainability.

Month 24 Exploitable Results

Pre-Editing Rules

The pre-editing rules are implemented using the Acrolinx software. Two types of rules have been designed; one which improves the source text and one for MT only, whose changes are automatic and not visible to users (see Deliverable 2.2).

In order to improve and adapt the pre-editing rules that have been developed, we have put in place an infrastructure for collecting user feedback. This enables us to monitor the way rules are used and the extent to which they are helpful in current and future integrations. Rule application is adapted to user feedback in two ways: 1) users can disable rules that do not work for them, 2) in the future, we will investigate how users acted on rules in order to improve rule precision. The results from the usability studies are now fed back into the design and documentation of Acrolinx rules. More concise rules for help texts (documentation) have been created and the presentation in the plug-in has been substantially improved.

We will treat the specifications for the pre-editing rules which we have identified as useful for MT as Open Source community content. Textual rule descriptions will be exploitable as Open Source knowledge, in order to make it available in the context of discussions around improvements to Statistical Machine Translation (SMT). The source code of the actual implemented rules will not be published.

We have generated a set of tutorials on pre-editing rules suitable both for live presentations and for self-study over the Web. The presentations are available to community members to guide their editing efforts. More details can be found in deliverable D 2.2 Definition of pre-editing rules for English and French.

Post-Editing Rules

The project is in the process of defining rules for post-editing SMT output in English, French and German. These rules will be applied automatically to the SMT output and support human post-editors as well as automatic ranking of SMT results.

These rules will also be implemented in Acrolinx and provide markings and suggestions to users. The users can then either accept or reject the suggestions. Some changes can also be applied automatically if the underlying rules are deemed highly reliable. As with the pre-editing rules, we will open-source textual rule descriptions for post-editing rules which we have identified as useful for

community content, but not the source code for their implementation in Acrolinx. We will also generate a similar set of tutorials on post-editing rules, which will again be available to community members.

Text Classification Rules for Forum and NGO Context

The project will develop rules for automatic classification of forum and NGO context. Classes can for example be questions and answers, or simple semantic categories. This also includes the development of rules for automatic detection of content with negative or positive sentiment (Is sentiment preserved in target language sentence? Is sentiment too strong to translate this sentence? Does sentiment lead to problems in translation? Should someone be informed about strong sentiment?)

Completed project tasks on text classification have focused on topic identification. As topics are open-domain entities, we used unsupervised learning algorithms on keywords. There are several ways to exploit the results of topic classification, such as:

- adapting pre-editing and post-editing rules to specific topics, and using them for documents with the respective topic
- training different SMT systems for different topics, and tuning them for documents with the respective topic
- forum community: asking users to edit "relevant" forum posts
- forum community: providing users with relevant answers to their questions
- translator community: assigning translation tasks to translators who are most conversant with a specific area.

Current efforts focus on rules to detect sentiment. As with the pre-editing rules, we will open-source textual rule descriptions of the classification rules which we have identified as useful for community content, but not the implemented rules.

Moses Enhancements and Extensions

The enhancements to Moses focus on domain adaptation methods and linguistic back-off for SMT.

The project explores and develops novel domain adaptation methods. The project also aims to improve the translation of morphologically rich languages with little in-domain parallel data, using linguistic methods.

Results on promising methods will be published in appropriate academic forums and integrated in software form into the Moses SMT system. Some of the development and test sets may be released for other research teams to conduct experiments.

The developed methods will be part of the Moses SMT system, and will mean that translations in cases of sparse in-domain data may be improved.

Furthermore to the above highlights, we will conduct studies investigating which types of MT errors (both linguistic and non-linguistic) are tolerable for specific tasks. Results of these studies will lead to improved methods for configuring the Moses system. The results will also be used to enhance the post-editing and pre-editing rules.

A third type of SMT system improvement is the enrichment of machine translation output with information, such as translation alternatives, in order to support monolingual post-editing once in production. These annotations will be integrated into the post-editing environment, so that post-editors will have immediate access to this additional information.

So far research has produced results on the following (see deliverable *D 4.2 Report on robust machine translation: domain adaptation and linguistic back-off*):

- Domain adaptation using modified Moore-Lewis filtering
- Training data weighting
- Training of mixture models using Pairwise Rank Optimization
- Improvements to the experimental pipeline for lattice translation
- Improved modelling with sparse features

All the Moses engine enhancements developed in ACCEPT will be either incorporated into the Moses core or be made available as separate optional packages, under the same open-source license as the rest of the Moses system.

ACCEPT-Enabled Plug-ins

The project has already delivered a pre-editing plug-in and will deliver a second one for post-editing the results of automatic translation. The plug-ins will be open-sourced and be available for web-based applications. They will be compatible with most website software as well as support platforms. The ACCEPT enabled plug-ins can be integrated into most web based system that uses text input with minimal effort.

The plug-ins are configured to be used with links to an Acrolinx server. This fact may limit the usage of the plug-ins for organizations that do not have Acrolinx licenses. Nonetheless, there is a configuration setting to disable the language checking functionality in the post-editing plug-in. This means the post-editing plug-in is still usable even if the user does not have an Acrolinx server connected to the ACCEPT API. For the pre-editing plug-in, the ACCEPT architecture should allow users to connect to another language checker. Obviously, some changes would be required, but this option might be appealing to non-Acrolinx users.

At the end of year 2 of the project, the plug-ins were released to the consortium members as well as SIG members. The latter were surveyed on their usage of the tool and were asked to rate its usability. Results are encouraging, the rating going from “useful” to “very useful”.

Community Development

We are building communities where the members use their native language and subject matter expertise to edit machine translated texts whether in a bilingual or monolingual environment.

The main objective for these communities is to provide feedback on development efficiency. Their feedback is used for measuring impact of the pre-editing rules – impact in terms of productivity, impact in terms of output quality determined through the correction categorization – as well as measuring the impact of Moses extensions on raw output and measuring impact of post-editing rules.

Feedback is collected through several user studies. The first year user study focused on tool adaptation, the second year user study focused on impact of pre-editing and the third year will focus on post-editing rules and Moses enhancements.

Besides the main objective of collecting feedback, several exploitable results are arising from the community.

In the first place, from the community feedback we would be able to identify best practices in terms of post-editing methods, particularly of monolingual post-editing methods.

In the second place, the project will result in defining guidelines and best practices for community management in a multilingual environment, particularly on methods for rewarding contributors for achieving tasks that are not linked to their main reason for being part of the community.

At the end of year 2 of the project, the best practices on post-editing have already been fine-tuned. The plug-ins have also evolved to take into account the post-editors' feedback from both communities.

Target Groups, Potential Partners and Other Stakeholders

The target groups are those that share online information with user communities, whether the content is user-generated or linked to knowledge bases. These groups cover all kinds of organisations: commercial corporations, academic institutions, NGOs, supra-governmental entities and non-profit organizations.

Communication Plan

We have established a plan to communicate with target groups in three phases. Currently, we are in the second phase.

1. Invitation: Baseline generation will coincide with generating interest in the project and collecting problem statements.

At the launch of the project, there has been a concerted effort to build a community of practitioners interested in the technical objectives of the project, to set baselines and find activists to participate in the practical aspects of the study. This core community is supplemented by building a community of potential technology adopters. The organisations in the larger community range from NGOs of charitable status and academic groups to commercial enterprises, who would engage in trials, discussion and review of the developing technologies.

2. Iteration: Recording improvement and the use and review of the technology by third parties.

The project is designed to have three iterations of technology development in this iterative phase. The emphasis is on ensuring that we are identifying the practical and significant variables to optimise. Both communities will have to be motivated and sustained by positive direction and acknowledged involvement.

3. Reporting: Gathering the central threads to generate best practice in service, and new directions for research. Also, the installation of a service infrastructure which would allow the community to continue to benefit from the new technology is being developed and will be issued either in WP5 or by one of the ACCEPT parties after the project. This service infrastructure would consist in plug-ins and APIs that are adapted to the user environment. In the final stages of the project there are three imperatives. First, to record the scientific advancements; second, to ensure that the science is reflected in the published training material; and third, to ensure that both communities are well integrated and the technology platform not only has a continued life but that it has taken root within the membership of the community of organisations.

Special Interest Group

To put the above plan into practice, we have established a Special Interest Group (SIG). The members of the SIG consist of tech-savvy institutions, both commercial and non-profit, who wish to test the technology and deploy it on their own social software stack, and smaller groups or companies who opt to use the portal to test the technology. Social software stacks are forum software and community platforms. The feedback from these diverse groups serves to guide the later stages of development of the project.

The feedback we received so far is very positive. SIG users declare needs for 8 languages, 3 of which are covered in the ACCEPT project. All of them, no matter what type of organization they represent, have forum content to translate. They set up such forums mainly to have a better understanding of people's needs and expectations. Their main focus when dealing with user content is the comprehension of such content whatever the language, with a priority on terminology.

Regarding the portal, they rated its usefulness from useful to very useful for both the pre-editing and post-editing features. Nonetheless they also showed preference to have access to APIs that can be installed on their own webpages rather than a portal.

On the legal side, to allow the SIG group to use the portal and related APIs, we defined Terms & Conditions for usage. On the same topic, a privacy policy has been drawn up for the use of the portal.

Communities

A similar plan has been carried out in the communities examined in the project. We have carefully rolled-out the ACCEPT technology to an increasing group of Norton community users, and have invited TWB translators to use ACCEPT technology in their translation workflow. As detailed in the previous chapter, the community feedback has been collected to improve the quality of the rules and the usability of the editing plug-ins.

Commercial Exploitation

The business models associated with this technology depend on the participants involved.

The ACCEPT APIs will be middleware APIs that abstract and unify the functionality provided by engines such as Acrolinx. Some of the APIs are currently not of that much use without access to a working Acrolinx server or something similar. Nonetheless, as described above, it is relatively easy to make this change in the next months. The plug-ins will be open-sourced, which means that these deliverables will not be commercialized as such, but may be used for consultancy purposes.

The open source status of these deliverables will encourage further development and widen the reach of the project. Interested parties may take these prototypes and develop their own tuned engines.

Aspects of the developments will lead to new product features which Acrolinx can exploit in their commercial offerings. Specifically, the pre-editing and post-editing strategies and the associated linguistic software will fit well with the existing product. Half of the SIG members are also Acrolinx customers, which simplifies the marketing of these new features.

We also expect the results of the ACCEPT project to flow directly into the relevant production settings at Symantec product forums with active communities, especially those where translation has a potential. Non-native speaker forums creating high-value information (i.e. around new products, for instance) would be most relevant, but ultimately this technology would be rolled out across all forums at Symantec, and other corporations with similar forums.

Lexcelera, as a Language Services Provider and expert in Machine Translation, will be able to improve its commercial offering through technologies and processes that result from this project. Several types of services can be identified such as pre-editing consultancy and support for its customer portfolio. For that purpose, Lexcelera and Acrolinx are investigating a SaaS-model for Lexcelera to use an Acrolinx server for their customers to check the MT-readiness of their content. Lexcelera will also be able to widen its post-editing services offering, whether through use of rules set through an Acrolinx server, but also by offering monolingual tasks for contents other than forums. Lexcelera has already foreseen the possible exploitation of the ACCEPT processes for providing new services to its customers around multilingual competitive intelligence and multilingual e-reputation. The community management experience also invites Lexcelera to improve its resource/vendor management model that could lead to more efficient and reliable linguistic teams. This community management experience can also be leveraged on customer management, to invite them as active participants.

Academic and Charitable Exploitation

Educational Sphere

The ACCEPT project has a number of deliverables which include not only traditional printed reports, invited speaking engagements and conferences, but also videos, blog posts and forum activity. These training materials will improve the richness and appeal of the research findings by

embedding the abstract science developed by the project into everyday examples of best practice.

This material can be beneficial for academic institutions that may widely and freely use the guidelines in trainings and seminars of post-editing (bilingual and monolingual).

Moreover, our analysis of existing automatic metrics and their suitability to user-generated content could be exploited by the scientific community, in particular by those researchers working with UGC.

Charitable

Lexcelera is committed to scaling up the operations of Translation Without Borders from millions of words per year to tens or even hundreds of millions of words. This level of scalability, and the enormous benefits it brings in giving more people access to important information, can only be achieved by more automation and by addressing bottlenecks around editing which currently make the use of MT less productive than it could be.

The ACCEPT project results could be reused to translate huge medical knowledge bases maintained by NGOs, whether these databases are online (e.g., http://www.msf.fr/activites/rougeole) or offline. This would lead to a larger and quicker spreading of risks, precautions and treatments of hundreds of recurrent diseases.

Contributions to Standards

The improvements in SMT will flow directly into the Moses system. As these improvements are being open-sourced, they can also be of benefit to other SMT systems which do not have an integrated concept of editing and MT yet. Increased up-take of SMT in the translation industry, with all the competitive advantages this brings, will be a significant result of the success of the project.

The public-facing portal is the testbed where technology iteration and process development are observed and harvested. The portal not only facilitates experimentation among the consortium partners and acts as a demonstrator for the user communities but also serves as an agent allowing dispersion of its own technology. The editor components will be available as a series of modules (e.g. Javascript/jOuery plug-in) which can be deployed into other web systems. In terms of standards, the ACCEPT Post-Editing API provides a reference implementation of the XLIFF (1.2) standard within the context of a post-editing scenario. The current instance of the portal in use in the project may be operated differently once the project is over.

Sustainability

Linguistic adaptations such as dictionary additions to the Acrolinx software are necessary to extend the coverage on forum and medical data. These adaptations will be part of the product, and will make it possible for language specific of forum and medical data to be processed by the Acrolinx software.

The pre-editing and post-editing rules will be part of the Acrolinx linguistic portfolio. They will thus be included in the Acrolinx software, further extended and provided to users who want to control their MT input and output. Text classification and sentiment detection rules will also be included in the extended software provided to Acrolinx users.

The Moses system will be generally improved based on the ACCEPT project results. Domain adaptation methods that will be developed will be accessible as part of the Moses software. The intensive error analysis of SMT will help to improve Moses results and to concentrate on errors that have a negative effect on user acceptance.

Appendix A. Relevant Reports Referred to in the Exploitation Plan

Description	Deliverable
Definition of pre-editing rules for English and French	D 2.2
Taxonomy of forum content and rules for automatic classification	D 3.1
Seminar material on pre-editing	D 6.1.2
Seminar material on post-editing	D 6.2.2
Analysis of existing metrics and proposal of a task-oriented metric	D 9.1
Dissemination plan	D 10.3
Project website	D 10.8

Table 1: Exploitation deliverables