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 1

Workpackage n° 10	Name: Dissemination and Exploitation
Deliverable n° 10.6	Name: Exploitation Plan Update 1
Due date: 31 December 2012	Submission date: 21 December 2012

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 1

This deliverable updates and refines the exploitation plan presented in March 2012 in Deliverable D10.5. This updated plan is valid for the period M13-24. The high-level goals of the original plan are still valid. A number of major steps described in the previous plan have been taken to reach these goals.

Introduction

We look forward to the day where information can be provided in the natural language of the various global users of our products. This challenge is addressed by enabling ease of editing of machine translated material using three processes: lightly controlling source, tuning MT engines and developing best practices for editors. We will enable any individual or organisation to post-edit machine translation with increased ease and confidence, and demonstrate how to edit the translated output without direct reference to the source content.

The project intends to provide technology and services to a wide range of users. These users would fall into three general classes:

- Third parties as yet unknown to the project,
- European citizens and their information providers,
- The ACCEPT partners themselves.

Third parties would include a range of individuals including but not limited to self-help groups with translation needs, through non-government organisations and charities, to commercial enterprises, who often have large amounts of information available in the source language(s) they use, but cannot readily provide translated materials in other languages.

Ordinary European citizens will act as the ultimate recipient of these services in that an increased amount of information will be available in a greater range of languages. Information providers in general, including the commission itself, will have an opportunity to leverage the technology as it develops.

The ACCEPT partners plan to address a number of specific goals. Academics have the opportunity to push back the frontiers of their chosen specialisation, and their results will be readily accessible to the global science community. The commercial partners will be able to advance their service offerings, while encouraging their peers to assimilate the experience and adopt the technologies. As the commercial parties support the charitable goals and activity of Translators without Borders, and the needs of their client-NGOs, these goals are shared by the whole consortium.

Target Groups, Potential Partners and Other Stakeholders

The exploitation coordinator in ACCEPT is Andrew Bredenkamp.

Symantec is an active member of the Centre for Next Generation Localisation (CNGL) based in Dublin (Ireland). Their association with other industry partners in this organization gives us a platform to disseminate information on the ACCEPT project.

We will build a Special Interest Group (SIG) who will help us test the portal and its component software elements. The membership of the SIG will consist of technically 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 will serve to guide the later stages of development of the project. We expect to grow the use of the portal from a few members in the first year to a group of informed and supportive members in the final year.

Progress at M12

The portal prototype (www.accept-portal.eu) has been made available and after extensive feed-back from the consortium can be made available to the initial list of special interest group members for further testing and use.

Exploitable Knowledge in the Project

Pre-Editing Rules

Description

The project will develop pre-editing rules for Machine Translation purposes for English and French. These rules will be based on a standard corpus analysis.

Exploitable Results

Rules will be implemented using the Acrolinx software tools and provide markings and suggestions to Symantec user forum authors.

Textual rule descriptions will be exploitable as open-source knowledge, in order to enhance the discussion on improvements to Statistical Machine Translation (SMT).

Progress at M12

A reduced rule set has been made available through the plug-in on the Norton Forum board used by the gurus. They have supplied cogent feedback which has been reflected in later versions of the tools.

Post-Editing Rules

Description

The project will define rules for monolingual post-editing of SMT output, for English, French, German and Japanese. These rules will be automatically applicable to the output and support human post-editors as well as automatic ranking of SMT results.

Exploitable Results

These rules will also be implemented inside Acrolinx and provide markings and suggestions to posteditors at Symantec and Lexcelera.

Textual rule descriptions will be available as open-source knowledge.

Text Classification Rules for Forum and NGO Context

Description

The project will develop rules for automatic classification of forum and NGO context. Classifications can for example be questions and answers, or matching simple categories.

Exploitable Results

Rules will be implemented inside Acrolinx and used in the Symantec and Translation Without Borders (TWB) environments.

Sentiment Analysis

Description

The project will develop rules for automatic detection of content with negative or positive sentiment.

Exploitable Results

Rules will be implemented inside Acrolinx and used by Symantec and TWB to guide translations (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?)

Domain Adaptation Methods for SMT

Description

The project will explore and develop novel domain adaptation methods.

Exploitable Results

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.

Linguistic Back-Off for SMT

Description

The project will develop methods for the translation of morphologically rich languages with little in-domain parallel data, using linguistic methods.

Exploitable Results

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

Exploitation of Usage Data

Description

We will exploit the usage of human editors and post-editors, in order to find out where the MT system failed.

Exploitable Results

Results will be used to enhance the Moses SMT system and the post-editing and pre-editing rules.

ACCEPT API

Description

The ACCEPT API will be a middleware API that abstracts and unifies the functionality provided by engines such as Acrolinx.

Progress at M12

The ACCEPT API has been deployed both on the portal and directly onto the Symantec Norton forum Lithium based technology stack.

ACCEPT Portal

Description

A portal for hosting the evaluation environment, simulators and administration functions for configuring solutions.

Exploitable Results

An implementation of the ACCEPT Portal will be released.

Progress at M12

The ACCEPT portal has been released to the consortium members.

ACCEPT Enabled Plug-in

Description

This project will provide a web plug-in that can integrate with existing web edit controls to deliver a content-rich quality assurance experience (grammar, style, spelling) to users. The plug-in will use the ACCEPT API for customized content analysis.

Exploitable Results

Once released, the plug-in may be used by Symantec and TWB authors and by Special Interest Group (SIG) members in their projects (as long as they have the appropriate Acrolinx licenses).

Progress at M12

The plug-in has been released to the consortium members.

Post-Editing Environment

Description

The project will develop a Post-Editing environment, where monolingual and bilingual user edits can be captured. This environment will be initially based on the functionality of Symantec's existing community collaboration portal.

Exploitable Results

Once released, the environment may be used by the Symantec forum and TWB communities. It may also be made accessible to SIG members.

Progress at M12

The first prototype of the post-editing functionality is available to consortium members via the portal for experimentation.

Evaluation Environment

Description

The project will provide an evaluation portal where user ratings can be collected to assess the quality of source, machine translated and post-edited content. This portal will be based on Symantec's existing evaluation portal.

Exploitable Results

Once released, the environment will be available for use by the Symantec forum and TWB communities. The portal may also be made available to SIG members who will access it through their individual accounts.

Progress at M12

The evaluation functionality is available to consortium members via the portal.

Community Development

Description

We will build communities, where the members use their native language and subject matter expertise to edit machine translated texts.

Exploitable Results

Communities will initially be organised around the Symantec forum and TWB/NGO translation activities.

Seminar on Pre-Editing Rules

Description

We will generate a set of presentations on pre-editing rules that can be delivered either by a practitioner or virtually, i.e. without active human intervention.

Exploitable Results

The presentations will be available to community members to guide their editing efforts.

Progress at M12

Deployed on the Norton community site for evaluation.

Seminar on Monolingual Post-Editing Rules

Description

We will generate a set of presentations on post-editing rules that can be delivered either by a practitioner or virtually, i.e. without active human intervention.

Exploitable Results

The presentation will be available to community members to guide their post-editing efforts.

Additional Information in the Monolingual Translator Scenario

Description

Machine Translation output will be enriched with information, such as alternate translations, in order to support monolingual post-editing.

Exploitable Results

Results will be directly part of the post-editing environment, in that post-editors will get enriched information.

Evaluation of Impact of Pre-Editing Rules on SMT

Description

The project will compare SMT translation with and without various types of pre-editing rules.

Exploitable Results

Results of this evaluation will be fed into the design, configuration and implementation of the pre-editing rules.

Evaluation of Impact of Bilingual and Monolingual Post-Editing Rules on Translation Quality

Description

The project will evaluate the result of SMT with bilingual and monolingual post-edition using end-users' judgments.

Exploitable Results

Results of this evaluation will be fed into the design, configuration and implementation of the post-editing rules.

Determination of MT Task Tolerance

Description

We will conduct studies investigating which types of MT errors (both linguistic and non-linguistic) are tolerable for specific tasks.

Exploitable Results

Results of these studies will lead to improved methods for configuring the Moses system.

Evaluation of How Users Interact with Pre- and Post-Editing Rules

Description

User decisions to accept or ignore automatically detected errors will be logged and evaluated. We will also conduct usability studies and observe how users perceive the overall authoring support during pre- and post-editing.

Exploitable Results

The user decisions will influence rule weightings, so that the application of rules is adapted to user feedback. The results from the usability studies will be fed back into the design and documentation of Acrolinx rules, and into their presentation within the ACCEPT plug-in.

Commercial Exploitation by Project Partners

The business models associated with this technology depend on the actors involved. On the one hand, aspects of these 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.

The improvements in SMT will flow directly into the Moses system, which will considerably improve the performance of Moses in comparison to other MT systems which do not have an integrated concept of editing and MT. Increased take-up of Moses in the translation industry, with all the competitive advantages this brings, will be a significant result of the success of the project.

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 than optimally productive. 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.

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. Symantec forums run on commercial social web technology stacks. The functionality of ACCEPT would be available to these technology stacks through APIs and JavaScript plug-in modules.

Progress at month 12

Initial work in both pre-editing and post-editing is being carried out through the Symantec forums and TWB community.

Knowledge Transfer

The world of customer and end-user engagement is changing rapidly. Over a few years, the means of communication has shifted from the traditional printed page to include a variety of digital media supported through websites, blogs and forums. Even the source of the information is changing, as users communities generate and propagate information and sentiment.

The ACCEPT project has a number of deliverables which, while including traditional printed reports, invited speaking engagements and conferences also includes video, blog 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 practical everyday examples of best practice.

The ACCEPT enabled plug-in can be integrated into “any” web based system that uses text input with minimal effort.

Community Building

Development of the project will proceed in three phases.

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 in this iterative phase. The emphasis is on ensuring that we are identifying the practical and significant variables to optimise. The community of practice and that of organisation will have to be motivated and sustained by positive direction and acknowledged involvement.

3) **Reporting:** Gathering together the central threads to generate best practice in service, and new directions for research. Also, the installation of a service infrastructure that would allow the community to continue to benefit from the new technology being developed.

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 the communities of practice and organisation 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.

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 specifics 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 failure analysis of SMT will help to improve Moses MT results and to concentrate on errors that have a negative effect on user acceptance.

Contributions to Standards

Where appropriate we are using XLIFF as a standard for data exchange (e.g. for the Evaluation environment).

ACCEPT Portal Access

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/JQUERY plug-in) which can be deployed into other web systems.

This ability gives businesses, whether large and small, technically adept or technically naïve, an even playing field in assessing ACCEPT functionality. This diversity is reflected in the consortium membership, where Symantec will be able to deploy functionality in its own web stack, whereas the NGOs represented by Lexcelera will avail themselves of the portal to evaluate their content. In this way, we hope to encourage the broadest usage of the developing technologies and hopefully glean rapid and relevant feedback for the latter development stages.

Table 1: Exploitation deliverables – taken from our original deliverable list.

Deliverable	Description	Access	Month
D2.1	Definition of pre-editing rules for English and French	PU	12
D2.2	Definition of pre-editing rules for English and French	PU	18
D2.3	Definition of post-editing rules for English, French, German and Japanese	PU	20
D2.4	Definition of post-editing rules for English, French, German and Japanese	PU	30
D3.1	Taxonomy of forum content and rules for automatic classification	PU	16
D3.4	Report of evaluation results	PU	36
D4.1	Baseline machine translation systems	PU	3
D4.2	Report on robust machine translation: domain adaptation and linguistic back-off	PU	24
D4.3	Report on improved machine translation by exploiting post-editing data	PU	36
D5.3	Adapted evaluation portal prototype to allow for the collection of user ratings	PU	9
D5.4	Browser-based client demonstrator used to access acrolinx IQ	PU	18

D5.5	Adapted Post-Editing Environment prototype	PU	24
D5.6	Adapted evaluation portal prototype	PU	24
D5.7	Monolingual and Bilingual Post-Editing Environment demonstrator	PU	36
D5.8	Evaluation portal demonstrator	PU	36
D6.1.1	Seminar Material on Pre-Editing - Edition 1	PU	6
D6.1.2	Seminar Material on Pre-Editing - Edition 2	PU	12
D6.1.3	Seminar Material on Pre-Editing - Edition 3	PU	18
D6.2.1	Seminar Material on Post-Editing - Edition 1	PU	6
D6.2.2	Seminar Material on Post-Editing - Edition 2	PU	12
D6.2.3	Seminar Material on Post-Editing - Edition 3	PU	18
D7.1.1	Data and report from user studies - Year 1	PU	12
D7.1.2	Data and report from user studies - Year 2	PU	24
D7.1.3	Data and report from user studies - Year 3	PU	36
D7.2	Report on assistance	PU	24
D8.1.1	Data and report from user studies - Year 1	PU	12
D8.1.2	Data and report from user studies - Year 2	PU	24
D8.1.3	Data and report from user studies - Year 3	PU	36
D8.2	Report on comparison of bilingual and monolingual editing	PU	36
D9.1	Analysis of existing metrics and proposal of a task-oriented metric	PU	12
D9.2.1	Survey of evaluation results – Version 1	PU	18
D9.2.2	Survey of evaluation results – Version 2	PU	24
D9.2.3	Survey of evaluation results – Version 3	PU	30
D9.2.4	Survey of evaluation results – Version 4	PU	36
D9.3	Weighting of pre-editing rules	PU	30
D10.1	Project Logo	PU	1
D10.2	Dissemination plan	PU	3
D10.3	Dissemination plan update	PU	12
D10.4	Dissemination plan update	PU	24
D10.5	Exploitation plan	PU	3
D10.6	Exploitation plan update	PU	12
D10.7	Exploitation plan update	PU	24
D10.8	Project Website	PU	1

D10.9	Video clip 1	PU	20
D10.10	Video clip 2	PU	20
D10.11	List of published papers in ACCEPT	PU	36

Table 2: Exploitation Goals – Measures and trends we can fit into our exploitation report(s)

Exploitation Goals	**2012 Target**	**2013 Target**	**2014 Target**
Sources examined	50.000 words	70.000 words	100.000 words
MT docs scored	30	100	500
Special Interest Group members	13	7	10
Links to ACCEPT pre-editing rules	2	10	30