D5.1.1: DRUPAL MT TRAINING MODULE

Karl Fritsche, Stephan Walter (Cocomore)

Distribution: Consortium Internal Report

MultilingualWeb-LT (LT-Web)
Language Technology in the Web

FP7-ICT-2011-7

Project no: 287815

Document Information

Deliverable number:	5.1.1
Deliverable title:	Drupal MT Training Module
Dissemination level:	CO
Contractual date of delivery:	30th September 2013
Actual date of delivery:	30th September 2013
Author(s):	Karl Fritsche, Stephan Walter (Cocomore)
Participants:	Cocomore
Internal Reviewer:	Cocomore
Workpackage:	WP5
Task Responsible:	Karl Fritsche
Workpackage Leader:	Clemens Weins

Revision History

Revision	Date	Author	Organization	Description
1	17/09/2013	Karl Fritsche, Stephan Walter	Cocomore	Draft

Contents

1. Executive Summary

This document describes the Drupal module developed by Cocomore to send aligned original and translated data with ITS 2.0 markup to a machine translation (MT) provider for data driven creation or optimization of machine translation engines or models.

The most common use case will be to train or tune a statistical MT model based on the aligned data and give special consideration on top of the standard techniques to the knowledge that is encoded in the ITS 2.0 markup. But other use cases, like the systematic identification of problematic cases for manual adjustment of a rule based MT system are also conceivable..

While ITS aware MT training was explored in more detail in D 5.2, the scope of this deliverable is the extraction of annotated and aligned bilingual data from the Drupal CMS. This process is based on the ITS 2.0 capabilities added to Drupal as described in the deliverables for WP 3. It was successfully tested in the context of the business case described in these deliverables (translation of VDMA press releases). Based on 141 press releases that were translated from German to French and Chinese. we could provide a three-way parallel annotated corpus of some 12.000 sentences.

2. Workflow

The Drupal MT Training Module described in this deliverable works on top of the ITS 2.0 aware Drupal translation workflow implemented in WP3. It therefore requires that the modules documented in D 3.1.1 are installed and a sufficient amount of CMS content has been annotated and translated within this workflow.

If these conditions are fulfilled extracting suitable MT training data from Drupal consists of the following two steps:

1. Collect source and translated content in files
2. Encode alignment of content items within these files

Both steps are carried out by custom-implemented command line tools. Step 1 can make use of the fact that within the translation workflow XHTML files for the annotated content (source and translations) are generated as part of a fail-over mechanism. These files are stored in a directory specific to the respective language service provider connector (e.g. the Drupal directory under “sites/all/linguaserve” for the Linguaserve connector created in WP3). These files are guaranteed to be complete and in-sync with respect to the corresponding content in the CMS database. However they need to be copied and re-named systematically in order to identify and handle them schematically by file-name and to prevent unintended interactions with their original failure recovery purpose.

Alignment between source and translated text is represented at paragraph level. To encode the alignment all paragraphs an paragraph are assigned identifiers, with identical Ids being used for aligned items in the different language versions of the same content. During the assignment of these Ids there is also a check for differences between ITS2.0 annotations. Assuming that e.g. all phrases which are marked up as non-translatable in the source text should still have the same markup in the target, warnings are issued if this is not the case (note that this check only covers the markup itself, not the content).

After these two steps the aligned files together with the problem report from the markup check are packed as an archive than can be sent to the MT provider.

Currently the data-specific part of the logic incorporated in the tools reflects the properties to the use-case investigated in the MultilingualWeb-LT project (e.g. texts are translated without removing or adding paragraphs, paragraphs are encoded as *p* and not as *div* in XHTML). Nevertheless they can be easily extended to cover further types of content that do not conform to the current assumptions on data formatting.

3. AVAILABILITY

[We are currently determining the details of where and how to make the software available.]