D5.2: METADATA-AWARE MT TRAINING TOOLS

Ankit K. Srivastava, Declan Groves, John Judge
Dublin City University (DCU)

Distribution: Public

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

FP7-ICT-2011-7

Project no: 287815

Document Information

Deliverable number:	5.2
Deliverable title:	Report on Metadata-Aware Machine Translation Training Tools
Dissemination level:	PU
Contractual date of delivery:	30th September 2013
Actual date of delivery:	31st October 2013
Author(s):	Ankit K. Srivastava, Declan Groves, John Judge
Participants:	Dublin City University (DCU)
Internal Reviewer:	TCD
Workpackage:	WP5
Task Responsible:	Dublin City University (DCU)
Workpackage Leader:	Dublin City University (DCU)

Revision History

Revision	Date	Author	Organization	Description
1	27/09/2013	Ankit K. Srivastava	DCU	Draft Version
2-final	30/10/2013	Ankit K. Srivastava	DCU	Final Version (Incorporated Comments from Internal Review)

CONTENTS

1. Executive Summary

The goal of this deliverable was to showcase the ability to make use of LT-web metadata for the training of MT system components.

In this Work Package (WP5 Task 5.2), we demonstrate ITS 2.0-tagged localized content stored in CMSs can be leveraged as parallel text to train new and re-train pre-existing Machine Translation (MT) engines.

This document is structured as follows:

- **Motivation** Highlight via a workflow diagram the in-house (DCU) MT engine *MaTrEx* along with extensions and modifications for translating ITS 2.0 tagged documents (developed in WP4, cf. D4.1.2)
- **Task Overview** Describe the extension modules (pre-process and post-process wrapper scripts) which enable training of MaTrEx on ITS 2.0 metadata-tagged parallel content
- **Web Service** Screenshots and webpage link to the metadata-aware MT training service
- **ITS 2.0 Data Categories** Describe the data (obtained from Cocomore and Linguaserve) used in MT training experiments and the ITS 2.0 data categories explored in order to test the metadata-aware MT training functionality
- **Conclusions** Summarise the findings of this deliverable

2. MOTIVATION

The DCU MT system MaTrEx (http://www.openmatrex.org/) is a Statistical Machine Translation (SMT) system developed in-house using the open-source Moses decoder.

The current functionality of MaTrEx (developed in WP4 Online MT Systems) is represented by the following workflow:

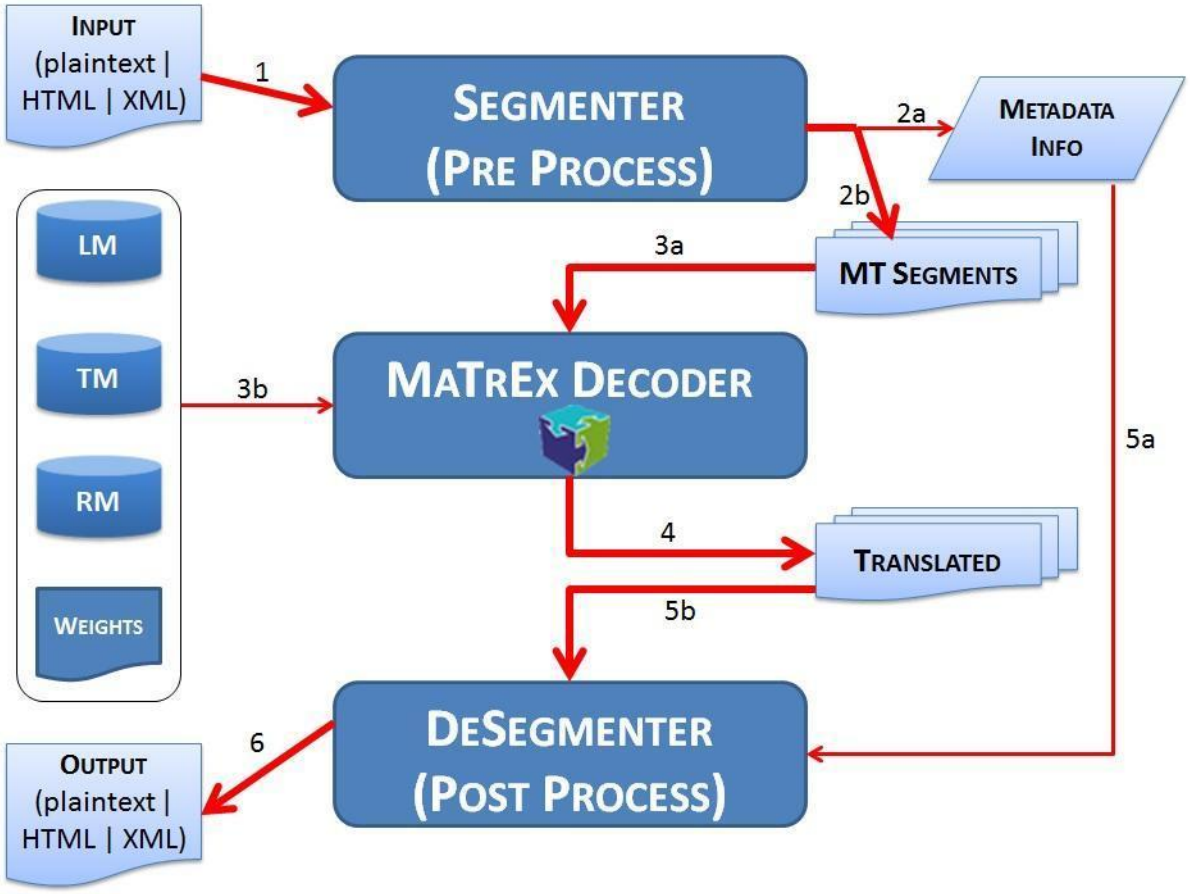


Figure 1. Workflow of MaTrEx Translation Module for ITS 2.0 metadata (WP4)

1 Input source language text in any one of the formats

plaintext | ITS2-HTML | ITS2-XML | ITS2-XLIFF

2 SEGMENTER - Pre-Process Module: Parses the ITS 2.0 tagged input document and generates
   a Metadata Wrapper Information {2a}
   b Segments to be translated (source language) {2b}

3 DECODER – MT Module: Translates the segments {3a} with the help of

{3b} SMT modules (Translation Model (TM), Language Model (LM), Reordering Model (RM), and feature weights)

4 Translated segments are generated (target language), also contains additional information like MT Confidence scores, etc.

5 DESEGMENTER – Post-Process Module: Takes as input the {5a} MetaData Wrapper Info (*from 2a*) and the {5b} translated segments (*from 4)* to merge and concatenate into one document

6 Output target language text in the same format as input

plaintext | ITS2-HTML | ITS2-XML | ITS2-XLIFF

The MaTrEx MT Engine (online interface at http://srv-cngl.computing.dcu.ie/mlwlt/) takes as input segments or a document of segments in source language (annotated with ITS 2.0 metadata), parses it to extract text to be translated, feeds the plain text to the MaTrEx decoder for translation, merges the ITS 2.0 metadata with the translated content, and generates the translated segment or document of segments in target language (annotated with ITS 2.0 metadata).

Thus the DCU MaTrEx MT system (Figure 1, developed and tested in WP4 Online MT Systems demo-ed with Linguaserve & Lucy Software [Deliverable D4.1.2]) is currently capable of translating an ITS 2.0-tagged document with the help of pre-processing and post-processing wrapper scripts.

In this task (WP5 Task 5.2), we seek to investigate whether the metadata contained in the IT 2.0 tagged documents can be utilised to train some of the MT system components like the translation model and the language model.

3. Task Overview

The key idea behind ITS 2.0-aware MT training is to take as input corpora of ITS 2.0-tagged documents in both source and target languages (in the simplest scenario, sentence-aligned parallel content). The pre-process module (Segmenter) parses these to extract the text from ITS 2.0 metadata, and uses it re-train a pre-existing SMT system (generate new translation models (TMs), reordering models (RMs) and language models (LMs)). Thus MT components / models are trained on ITS 2.0-aware parallel content.

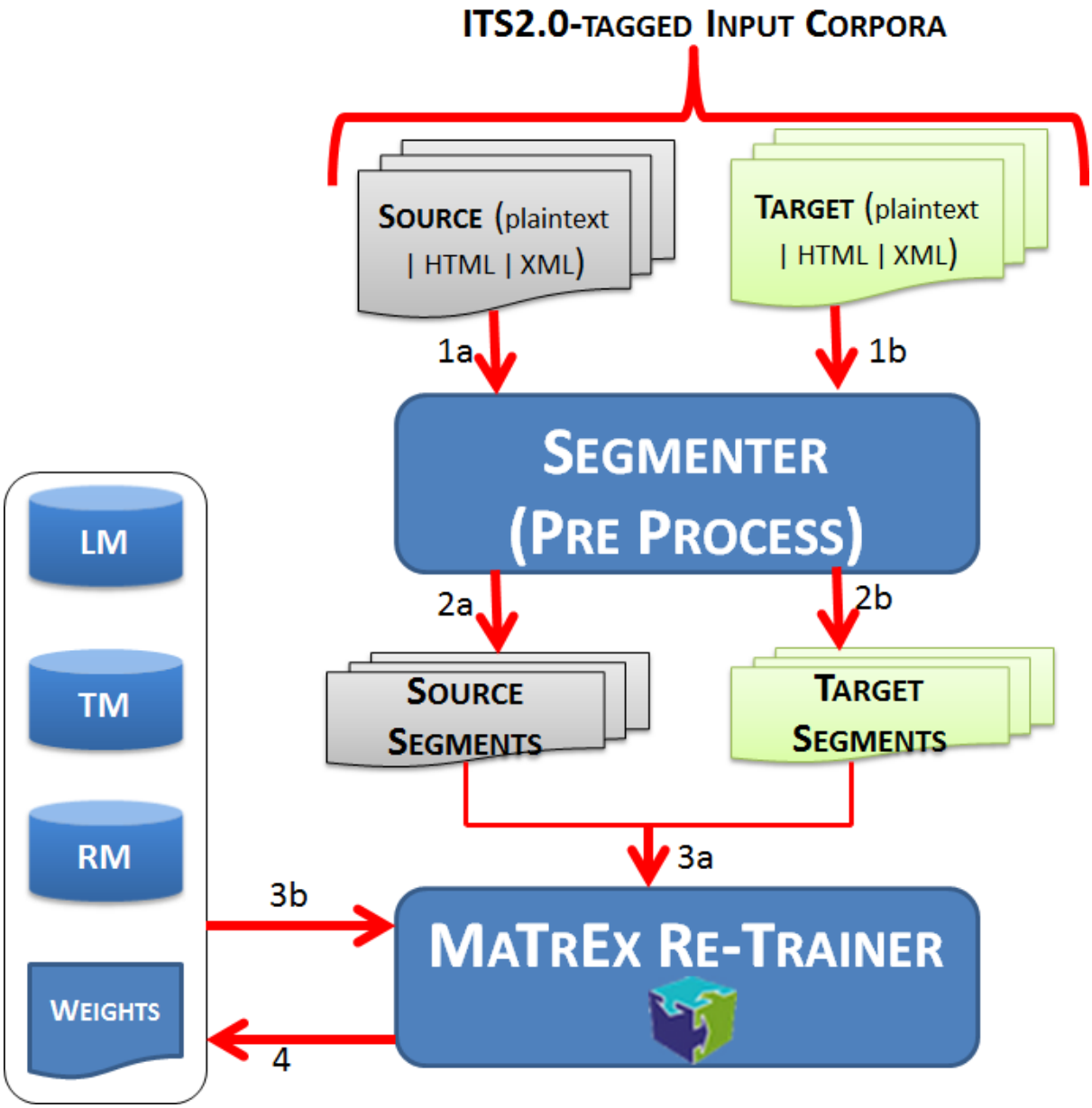


Figure 2. Workflow of MaTrEx Re-Training Module for ITS 2.0 metadata (WP5)

The workflow (illustrated in Figure 2) is as follows:

1. Input corpora (sentence-aligned parallel content) in one of the formats: plaintext | ITS2-HTML | ITS2-XML | ITS2-XLIFF
    a. Content in Source Language
    b. Content in Target Language
2. SEGMENTER - Pre-Process Module: Parses the ITS 2.0 tagged input corpora (both source and target) and extracts [some of the parsing capabilities has already been developed as part of the work for WP4]
    a. Segments in source language ({2a}
    b. Corresponding Segments in target language {2b}
3. RETRAINER – Main Module: Processes the bilingual segments {3a} to generate new training data to augment pre-existing {3b} SMT models (Translation Model (TM), Language Model (LM), Reordering Model (RM), and feature weights)
4. Retrained SMT models ((Translation Model (TM), Language Model (LM), Reordering Model (RM), and feature weights) are produced and replace the old versions.

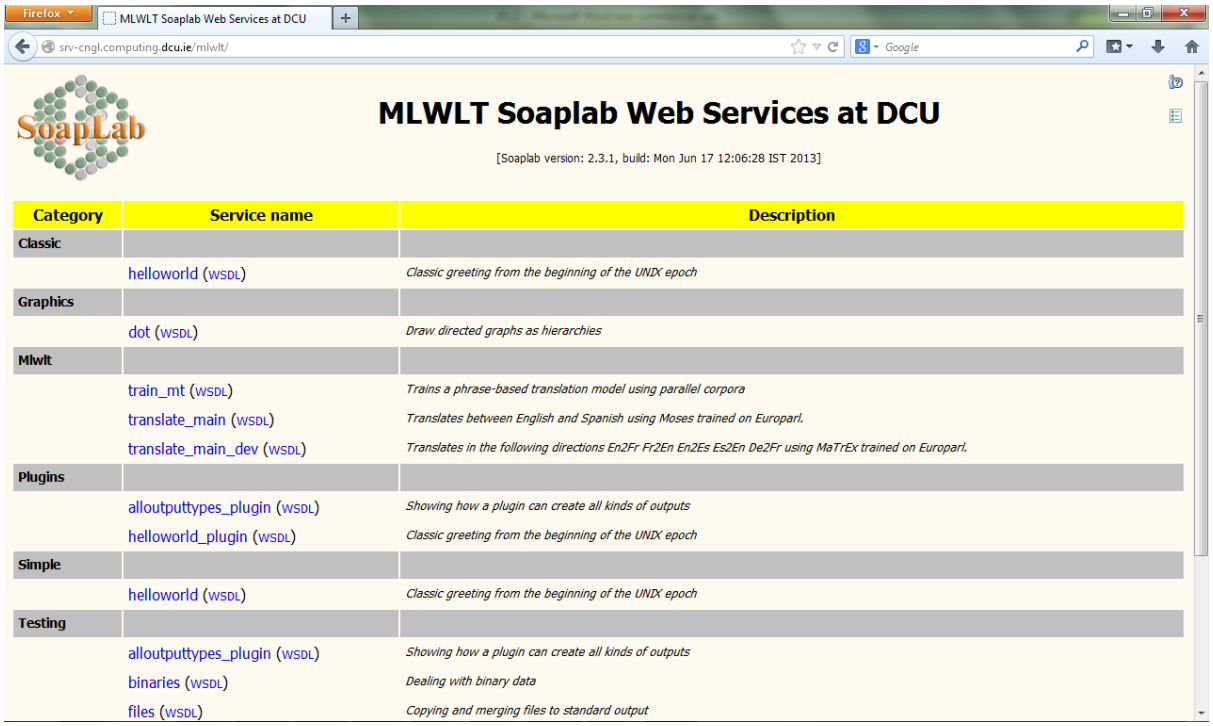
Previously, the MaTrEx engine was configured to receive only plain text, for both training and testing, and to run in batch mode, rather than for on-the-fly instant translation. To make the MT system ITS2.0-aware, a number of modifications were made to the engine (highlighted in Figure 1).

A large proportion of the modifications that are required to enable ITS 2.0 compliant training have also been required as part of the work in WP4 (Deliverable D4.1.2). In order to correctly recognise and process ITS 2.0 metadata, we have created additional ITS 2.0 specific processing modules, implemented in PERL. The main internal functionality has been left unchanged.

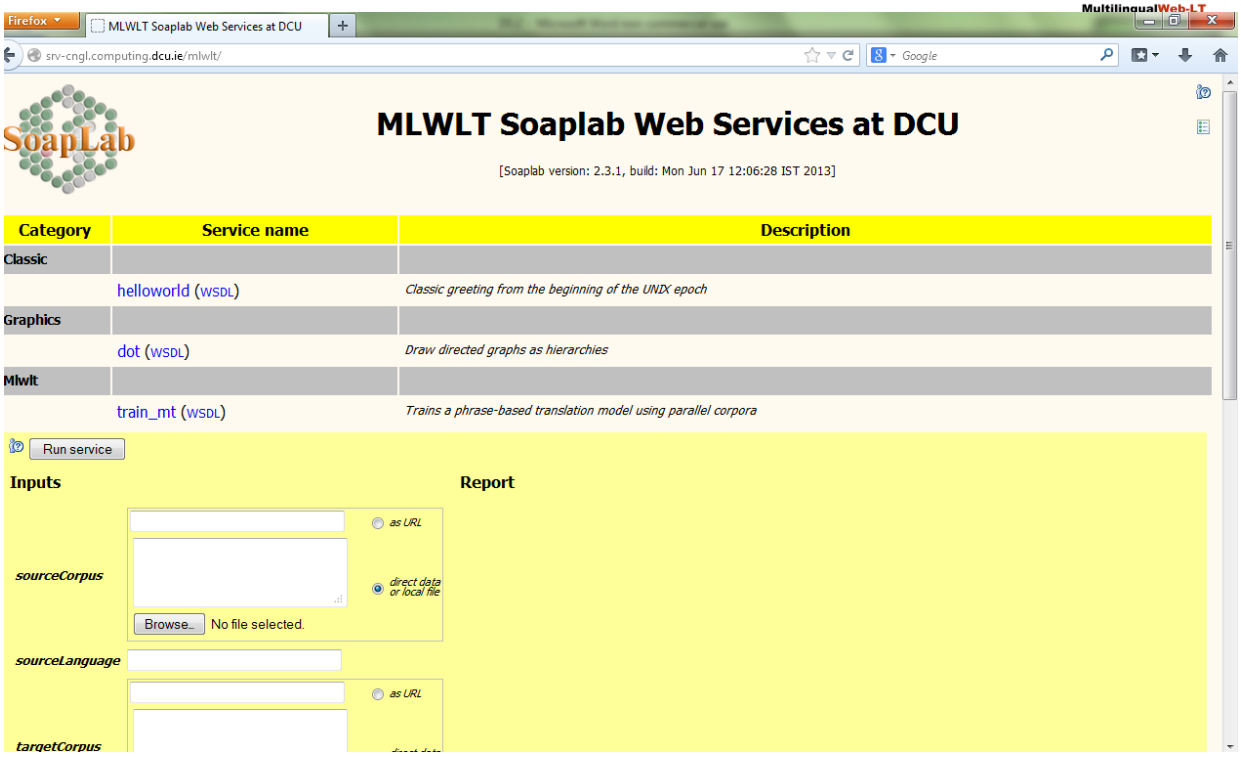
As part of WP5, we offer a new service in addition to the MaTrEx Translation web service: MaTrEx Training / Retraining web service. The system takes as input ITS 2.0-tagged parallel content (corresponding data files in the source and target language) and retrains the MT components with the help of the pre-processor scripts used by MaTrEx Decoder.

4. LINK TO WEB SERVICE

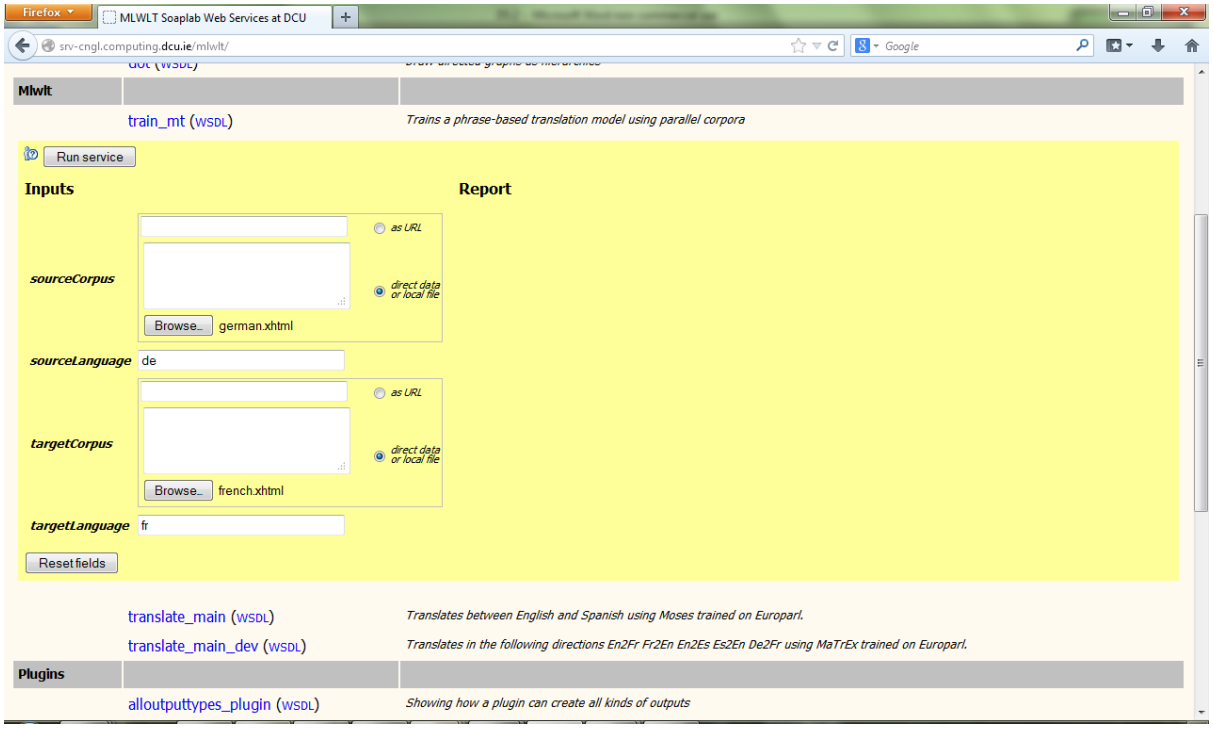
The MaTrEx MT system for LT-Web project can be accessed via a web service located at http://srv-cngl.computing.dcu.ie/mlwlt/



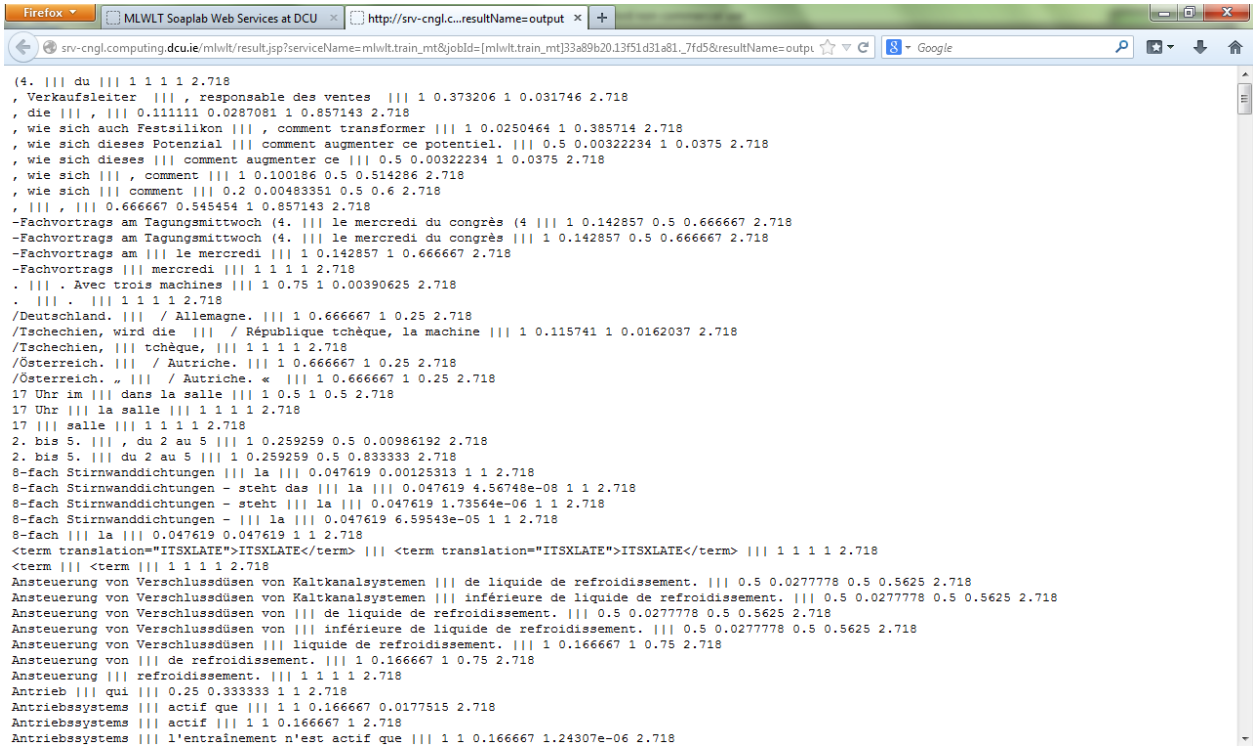
The specific web service for training MT components (translation model) is named **train_mt** and can be accessed by clicking on the link under **Mlwlt** as shown below



Upload the source language file (**source_corpus)** and target language file (**target_corpus)**. Type in the code for the source language and target language name (For example, **de** for German, and **fr** for French). Then, click on **Run Service** to initiate training as shown below



The Segmenter and Retrainer Modules described in Section 3 run in the background and output the retrained translation model (**accessed by clicking on the output link**). A sample output is shown below (The ||| is a field separator for each line consisting of source_language_phrase, target_language phrase, sequence of 5 probabilities encapsulating Translation Model features)



In the next section (Section 5), we describe a set of experiments performed using this training web service.

5. ITS 2.0 DATA CATEGORIES IN MT TRAINING

For Task 5.2, Metadata-aware MT Training, the most relevant ITS 2.0 data categories are:

- **Translate** *Selection of segments which are not to be translated*
- **Terminology** *Selection of segments whose translations are specified (e.g. named entities or linking to a dictionary)*
- **Domain** *Selection of segments belonging to a different domain from the default setting implying a different translation / language model*
- **Language Information** *Selection of segments belonging to a different language from the default setting implying a different translation / language model*

We demonstrate training of translation models on:

- **Spanish-English** *data obtained from Linguaserve generated as part of WP4; 120 documents spanning 100,000 words; domain: economics and tax*
- **German- French** *data obtained from Cocomore, generated as part of the Drupal MT training module developed in Task 5.1; 140 documents spanning 87,000 words; domain: technical documentation*

For WP4, we developed both Spanish-English and French-English baseline MT systems, making use of freely available resources (European Parliamentary Proceedings available at http://www.statmt.org/europarl/) . As per the DoW, these systems were set to be used in WP5 as well as in WP4, as the back-end[1] systems for the online MT showcase. However, we were unable to procure French-English content tagged with ITS 2.0 metadata. Instead, we acquired German-French content generated by Cocomore as part of Task 5.1

Based on Cocomore's real-world usage, we built an additional German-French baseline system. We then used ITS2.0 training material for German-French to iteratively tune this baseline system based on the ITS2.0 metadata. Facilitating this language pair was not originally envisaged in the DoW, but we have made

[1] Owing to the relatively small size of available parallel content which is ITS 2.0 tagged, we build baseline systems on large amounts of non-ITS 2.0 data (approximately 50 million words) and then supplement these back-end systems with ITS 2.0 tagged content

the decision to do so in order to support the real-world use-case for the CMS providers.

We implemented two versions of the metadata-aware MT Training module:

- **Passive Training** (Version 1) *Retrains models by processing out the metadata tags. Demonstrates compatibility of MaTrEx system with parallel content tagged with ITS 2.0 metadata*
- **Active Training** (Version 2) *Retrains models by using metadata information (E.g. terms marked as do-not-translate) to force the translation of terms in the training phase. Demonstrates application of metadata information in training and contrasts with baseline system (version 1).*

The Passive Training module differs from the baseline training functionality in only that the engine can now parse ITS 2.0 tagged parallel content and hence is metadata-aware. This allows retraining of MT systems on ITS 2.0 content (generated by other partners in WP4 and WP5 Task 5.1).

The Active Training module is the novel contribution of this deliverable wherein we experiment on leveraging ITS 2.0 metadata in the training phases of the SMT process. Our experiments focuses on the ITS 2.0 data category **translate.** The MT engine is forced to acknowledge the terms marked as "do not translate" instead of relying on letting the probabilities chose a translation. This results in consistent translation of do-not-translate terms across the entire corpus.

Procedure for Active Training

1. Take as input source and target documents tagged with ITS 2.0
2. Extract translatable segments from both source and target sides
3. Use Preprocess (Segmenter) scripts outlined in Section 3
4. Run Term Substitution Algorithm for <translate> segments, i.e. replace do-not-translate terms with unique IDs across both source and target sides of the corpus
5. Run MT Training Scripts to re-train translation, language models
6. Revert the unique IDs to the original forms

7. Translate documents and compare with baseline system performance

Usually in a SMT system, the translation model consists of multiple translations for each word / phrase with varying probabilities. The purpose of the Term Substitution Algorithm is to force the MT engine to only store one specific translation (with probability 1.0) for any term which is marked as do-not-translate. This results in as much as a **3% absolute improvement** over the baseline system in terms of MT accuracy. An added benefit is the consistency in translation across the entire corpus.

Note that although we only experimented on the **translate** data category, the scripts are generic and can be implemented to leverage information from other data categories like terminology, domain, etc.

6. CONCLUSIONS

We have demonstrated that it is possible and potentially useful to train MT translation and language models on ITS 2.0-tagged documents. The main benefit of this feature is the added consistency in translation of frequently occurring terms (encapsulated by the ITS 2.0 data categories Translate and Terminology) of multiple documents in a localization cycle.