D5.2 - ANNEX
METADATA-AWARE MT EXPERIMENTS

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 Annex
Deliverable title:	Metadata-Aware Machine Translation Experiments
Dissemination level:	PU
Contractual date of delivery:	30th September 2013
Actual date of delivery:	30th September 2013
Author(s):	Ankit K. Srivastava, Declan Groves, John Judge
Participants:	Dublin City University (DCU)
Internal Reviewer:	Dublin City University (DCU)
Workpackage:	WP5
Task Responsible:	Dublin City University (DCU)
Workpackage Leader:	Dublin City University (DCU)

Revision History

Revision	Date	Author	Organization	Description
1	30/09/2013	Ankit K. Srivastava	DCU	Draft Version

CONTENTS

1. Executive Summary

The Annex is an accompaniment to the main document (D5.2 Report) detailing the specific experimental setup and results on retraining MT systems with ITS 2.0 metadata.

The purpose of this annex is to provide empirical evidence and implementation details of leveraging ITS 2.0 metadata to retrain MT systems. It is structured as follows:

- **Data & Tools** Summary of parallel seed data, ITS 2.0-tagged data, and MT tools used in all experiments
- **Retrain on translate** Translation performance of MT systems retrained on ITS 2.0 data category translate
- **Retrain on terminology** Proof-of-concept experiment on MT system retrained on ITS 2.0 data category terminology
- **Retrain on domain** Proof-of-concept experiment on MT system retrained on ITS 2.0 data category domain

2. Data and Tools

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 (http://www.statmt.org/moses/).

Pre-process and Post-process wrapper scripts (*in PERL)* to parse ITS 2.0 tagged documents, developed as part of WP4 and WP5.

Seed data refers to large amounts of non ITS 2.0 parallel data used to train a baseline system owing relatively small size of readily available parallel content which is ITS 2.0 tagged.

- **Spanish-English** *data obtained from freely available European Parliamentary Proceedings available at http://www.statmt.org/europarl 1.9 million sentences spanning 50 million words; domain: parliament*
- **German- French** *data obtained from freely available European Parliamentary Proceedings available at http://www.statmt.org/europarl 1.3 million sentences spanning 30 million words; domain: parliament*

ITS 2.0-tagged data refers to small amount of parallel text tagged with ITS 2.0 data categories used to retrain baseline systems trained on seed data

- **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*

In our experiments we compare performance of 3 MT systems

- **Baseline** MT system trained on seed data used to translate test set of 100 ITS 2.0 tagged documents
- **Passive** MT system retrained on ITS 2.0 tagged data with processed out tags used to translate test set of 100 ITS 2.0 tagged documents
- **Active** MT system retrained on ITS 2.0 tagged translation model used to translate test set of 100 ITS 2.0 tagged documents

3. Translate Data Category

The translate data category is used to select words and phrases which retain their translation across source and target language. We experimented on German-French system. The evaluation is done using the BLEU metric, which compares the similarity of the translated document to a reference document and gives a score between 0 and 1 rounded

to a percentage such that a score of 100 signifies perfect translation. Both Passive and Active systems perform over the Baseline system as shown in Figure 1. The Passive system simply discards all ITS 2.0 information but gives a better output on account of being retrained on the smaller ITS data. The Active system force feeds the translations in the translation model and is shown to give a significantly better performance over both the baseline and passive systems.

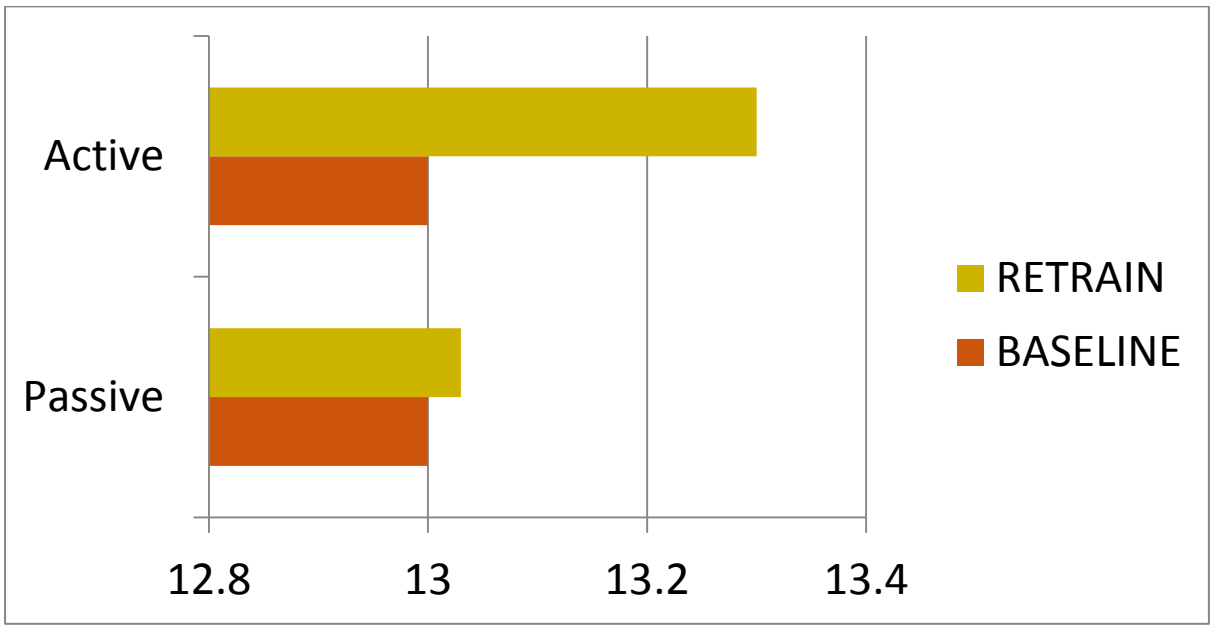

Figure 1 Comparative Translation Performance of MT systems retrained on “translate”

4. TERMINOLOGY DATA CATEGORY

This is a proof-of-concept experiment demonstrating use of translations provided by external sources (like dbpedia) and encapsulated in the ITS 2.0 terminology data category. Select words and their translations were marked with ITS 2.0 data tag terminology and used to retrain the translation model (Spanish-English). The test set was an artificially created set from the same domain as seed data (parliament). Therefore, the Passive system which simply discards all ITS 2.0 information does not differ from the baseline system. The Active system force feeds the term translations in the translation model and is shown (Figure 2) to give a significantly better performance over both the baseline and passive systems.

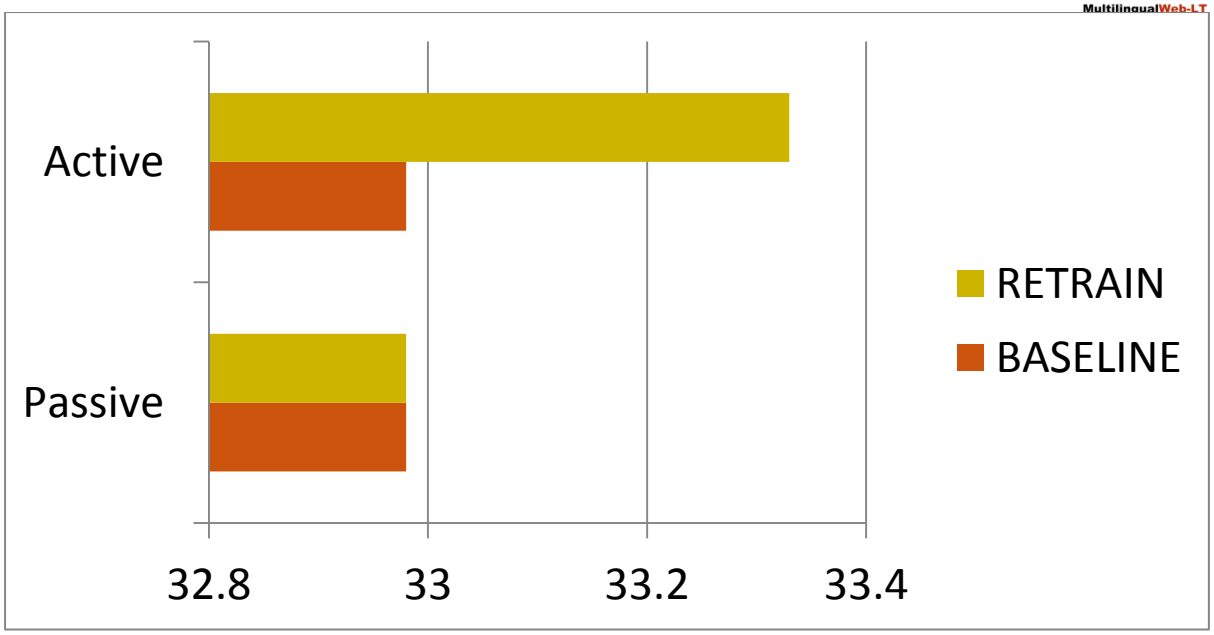

Figure 2 Comparative Translation Performance of MT systems retrained on "terminology"

5. Domain Data Category

This was another proof-of concept experiment wherein we trained two different translation and language models for Spanish-English for two different domains: parliament and tax. The test set was a mixture of documents from both domains (specified using the ITS 2.0 data category domain). The purpose was simply to demonstrate the functionality of selecting a domain-appropriate model during translation using information stored in the ITS 2.0 metadata. The baseline system did not differentiate between the domains. The domain-tuned system performed significantly better, as expected (Figure 3).

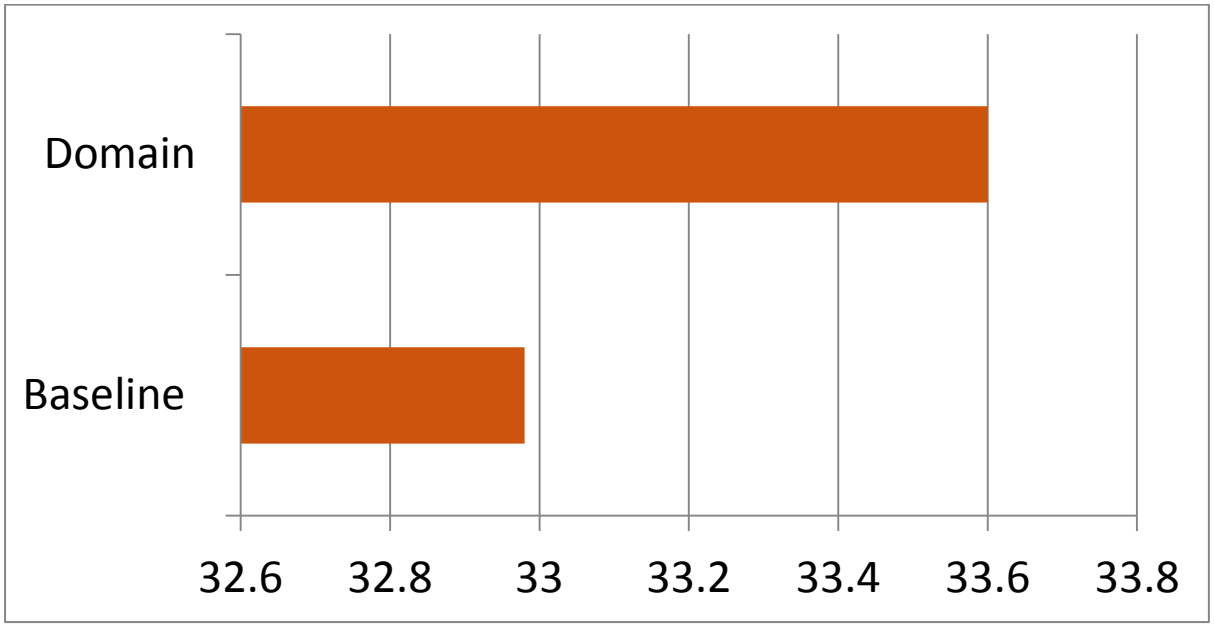

Figure 3 Translation Performance of MT systems retrained on "domain"

6. CONCLUSIONS

We have demonstrated that it is possible and potentially useful to train MT translation and language models on metadata leveraged from ITS 2.0-tagged documents. Incorporation of the information stored in metadata resulted in significantly better performing MT systems.