FALCON: Building the Localization Web

Andrzej Zydroń MBCS
CTO XTM International Ltd.
azydron@xtm-intl.com

Abstract

This document describes the EU FP7 funded FALCON (Federated Active Linguistic data CuratiON) project.

1 Introduction

FALCON (http://falcon-project.eu/) is a European Union funded FP7 project comprising Trinity College Dublin (TCD), Dublin City University (DCU), Easyling/SKAWA, Interverbum/TermWeb and XTM International. FALCON stands for Federated Active Linguistic data CuratiON and is largely the brainchild of David Lewis, Research Fellow at Trinity College Dublin. FALCON initially had the following important goals:

1. To establish a formal standard model for Linked Language and Localisation Data (L3Data) as a federated platform for data sharing based on a RDF metadata schema.
2. To integrate the Skawa/Easyling proxy based web site translation solution, Interverbum/TermWeb web based advanced terminology management and XTM web based translation management and computer assisted translation products in one seamless platform.

To integrate and improve SMT performance benefitting from the L3Data federated model as an integral part of the project as well as integration of the DCU SMT engine with XTM

2 General Description

Manuscripts must be in single column format. Type single-spaced. Start all pages directly under the top margin. See the guidelines later regarding formatting the first page. The paper should not exceed the maximum page limit described in Section 4.

2.1 Background

The FALCON project started in October 2013 and is scheduled to run for two years ending in September 2015.

FALCON will provide a mechanism for the controlled sharing and reuse of language resources, combining open corpora from public bodies with richly annotated output from commercial translation projects. Federated access control will enable sharing and reuse of commercial resources while respecting business partnerships, client relationships and competitive and licensing concerns.

2.2 Detailed Description

You can think of the L3Data aspect of FALCON as a distributed, federated database that points to the domain specific training and terminology data that is available, given certain commercial restrictions as regards private data, that can be used to build custom SMT engines

Proceedings of the 37th Conference Translating and the Computer, pages 33–36,
London, UK, November 26-27, 2015.

on the fly. In the world of the Internet only a distributed federated linked data database can achieve this. FALCON will use the highly flexible Resource Descriptor Framework (RDF) and a Simple Protocol and RDF Query Language (SPARQL) database. Using the Semantic Web concept, FALCON will provide a fast and efficient mechanism for sharing translation memory and terminology data for specific domains.

As Don DePalma of Common Sense Advisory describes very eloquently in his article entitled 'Building the Localization Web': http://goo.gl/jE6zuz, this will potentially allow smaller LSPs to have access to a much broader range of linguistic assets than would otherwise be the case. A federated, distributed L3Data store will allow for a very flexible and very scalable model, without the limitations and restrictions associated with centralized repositories.

3 Innovation

The improvements to SMT foreseen at the start of the FALCON project were to cover the following aspects:

1. Continuous dynamic retraining of the SMT engine with real-time feedback of post-edited output.
2. Named Entity Recognition (NER) to protect personal and product names etc. from being processed accidentally by the SMT engine: e.g. 'President Bush' from being transliterated as 'President Small Shrub'.
3. Providing an optimal segment post-editing sequence which will provide maximum benefit for the continuous retraining of the SMT engine.
4. Integration of terminology into the SMT chain by forcing the SMT engine to use terminology, where it exists and is identified, (so called 'forced decoding') rather than relying on the statistical probabilities for the translation.
5. Active translation memory (TM) and terminology resource curation through the L3Data RDF database built as part of FALCON.

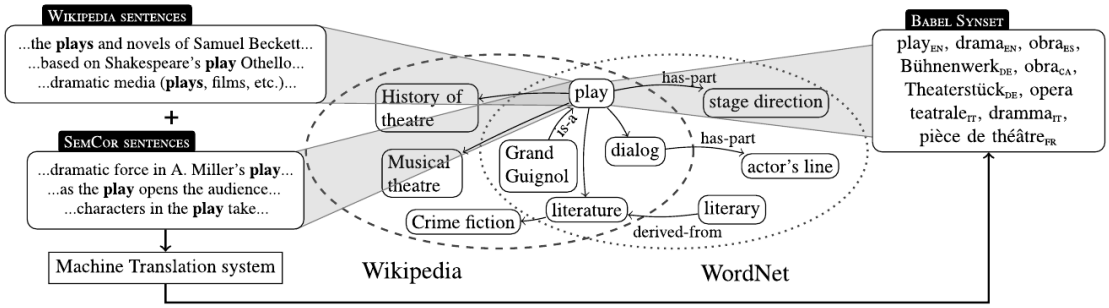
Apart from the L3Data store, which in its own right is a very important step forward in terms of establishing a federated way of holding relevant data, and curation and optimal translation sequence, the improvements build on existing advances in SMT. Nevertheless their integration into a production workflow based around XTM represents an important incremental step forward in terms of automation and consolidation of techniques. More importantly the investigation process around SMT improvements has yielded another 'golden egg'.

The initial SMT engine for FALCON was going to be OpenMaTrEx (http://www.openmatrex.org/) from DCU. OpenMaTrEx was an adaptation of the Moses SMT engine, but with an added twist: it introduced the concept of 'marker' or function words to assist in phrase alignment. All languages use around 230 function words such as prepositions, conjunctions, pronouns, ordinals such as 'if', 'but' 'above', 'over', 'under', 'first' etc. to delineate phrases and sub-segments in sentences. This was an interesting avenue of experimentation that in the end did not provide the hoped for improvement in alignment, but the concept was nevertheless very sound from the linguistic point of view. More on the OpenMaTrEx concept later.

An additional important aspect of the FALCON project has become BebelNet (http://www.babelnet.org). The implications of BabelNet were not immediately apparent during the initial design phase. It was only while investigating ways of improving SMT performance in terms of word and phrase alignment that its significance became truly apparent. An initial review of the BabelNet dataset and API provided a revelation.

4 BabelNet

BabelNet is a truly marvellous project funded by the European Research Council (it is part of the MultiJEDI (Multilingual Joint word sensE DIsambiguation) project). BabelNet is a multilingual lexicalized semantic network and ontology. So far so good. What is truly impressive about BabelNet is its sheer size, quality and scope: BabelNet 2.5 contains 9.5 million entries across 50 languages. This is truly Big Lexical Data. Roberto Navigli and his team at the Sapienza Università di Roma have created something quite remarkable, The plan for BabelNet 3.0 is 13+ million entries across 263 languages. What is truly astounding about BabelNet is the sheer size, breadth and depth of the semantic data:



By trawling through Princeton's remarkable WordNet lexical resource for the English language and then through Wiktionary, Wikipedia and following through additional resources on the Internet BabelNet has produced a veritable multilingual parallel treasure trove. Its richness also allows for word sense disambiguation (WSD) for homographs, one of the big 'bug bears' of MT and SMT.

Using the BabelNet API it is very easy to produce bilingual dictionaries. It does not take a great deal of imagination to work out what the addition of truly large-scale dictionaries can have on the accuracy of SMT engines. Even just adding the dictionary data to the training data for a Moses based SMT engine has a significant effect on the accuracy and quality scores.

Big Lexical data has the potential to remove the 'blindfolds' that have shackled SMT to date, significantly improving both accuracy and performance through bilingual dictionaries and word sense disambiguation.

BabelNet will continue to grow in size and scope over the next few years adding further online dictionary data such as IATE (http://iate.europa.eu/) and other multilingual open data resources.

5 The future

There is still much work to be done. The Moses GIZA++ word aligner is not optimized for dictionary input and has no direct notion of mechanism for WSD. The Berkeley Aligner can take dictionary input as it is designed for both supervised and unsupervised operation but is primarily designed for word and not phrase alignment. Much research word remains to be done, but the fundamentals of SMT have now been significantly shifted. BabelNet in its current form does not tackle function words, but it is relatively simple using existing Internet resources to 'harvest' the bilingual equivalents between various languages. The use of function words can then be used to assist with sub-segment and phrase alignment in the manner foreseen by OpenMaTrEx.

The SMT team at DCU, Trinity College and the rest of the FALCON team will be working on adapting existing Open Source software such as Moses and the Berkeley, Apache and Stamford tools to take maximum advantage of BabelNet.

Many other features of SMT regarding morphology and differences in word sequences between languages remain to be fully resolved in the Open Source domain, but the basic building blocks for truly effective machine translation are now in place. Just as search engines revolutionised the way we access data on the Internet in ways unforeseen in the early 1990's, SMT is well on the way of becoming the primary way that we translate (if not the way we are already doing so to get the 'gist' of what is on a given web page or email in a language that we do not understand).

Human endeavour is always based on incremental improvements. Just as OCR reached a tipping point in the mid 1990's so SMT is going to be the predominant tool for translation within the next 5 years. Just as translation memory, terminology tools and integrated translation management systems (TMS) have helped to automate and reduce translation and more significantly project management costs, integrated and automated quality SMT will further automate the actual translation process itself. Translation will become in the main a SMT post-editing process.

The quality and data resource issues have been largely addressed in theoretical terms: implementation of these ideas is well on the way. The translation workflow will be mainly around post-editing for most commercial translation projects. This can only be a good thing for all concerned: the demand for translation is growing at around 8% pa. and further automation of the process is the only way to meet this growing need which contributes so much to the increase in global trade helping lift billions of people from levels of poverty

Acknowledgments

The Falcon project was funded by the European Union under the auspices of the FP7 program.

References

Roberto Navigli, Simone Paolo Ponzetto, Artificial Intelligence. 2012. BeblNet: *The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*, Elsevier Artificial Intelligence 193 (2012) 217-250

Sandipan Dandapat, Mikel L. Forcada, Declan Groves, Yanjun Ma, Sergio Penkale, John Tinsley, and Andy Way. Pavel Pecina 2011. OpenMaTrEx. *OpenMaTrEx, a free/open-source marker-driven example-based machine translation system,* http://www.openmatrex.org/, 2011

Philipp Koehn Hieu Hoang Alexandra Birch Chris Callison-Burch. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. http://homepages.inf.ed.ac.uk/pkoehn/publications/acl2007-moses.pdf