Linking Localisation and Language Resources

David Lewis, Alexander O'Connor, Sebastien Molines, Leroy Finn, Dominic Jones, Stephen Curran, and Séamus Lawless

Abstract Industrial localisation is changing from the periodic translation of large bodies of content to a long-tail of small, heterogeneous translations processed in an agile and demand-driven manner. Software localisation and crowd-source translation already practice continuous fine-grained distribution of translation work. This requires close integration and round-trip interoperability between content creation and localisation processes, while at the same time recording the provenance of translated content to maximise it reuse in future translation tasks, and, increasingly, in training Statistical Machine Translation (SMT) engines. This work adopts a Linked Data approach to integrating the content translation round-trip process with the logging of process quality assurance provenance. This integration supports a pull-based interoperability model that supports continuous synchronising of content and process meta-data between the generating organisation and any number of language service providers or translators. We present a platform architecture for sharing, searching and interlinking of Linked Localisation and Language Data (termed L3Data) on the web. This is accomplished using a semantic schema for L3Data that is compatible with existing localisation data exchange standards and can be used to support the round-trip sharing of language resources. The paper describes our approach to development of L3Data schema and data management processes, web-based tools and data sharing infrastructure that use it. An initial proof of concept prototype is presented which implements a web application that segments and machine translates content for crowd-sourced post-editing and rating.

David Lewis · Alexander O'Connor · Sebastien Molines · Leroy Finn · Dominic Jones · Stephen Curran · Séamus Lawless
Centre for Next Generation Localisation, Knowledge and Data Engineering Group, Trinity College Dublin, Ireland e-mail: {Dave.Lewis,Alex.OConnor,moliness,finnle,Dominic.Jones,Stephen.Curran,slawless}@cs.tcd.ie

1 Introduction

Industrial localisation is changing from the periodic translation of large bodies of content to a long-tail of small, heterogeneous translations processed in an agile and demand-driven manner. To enable service providers to compete and innovate in this new market, the capture, reuse and sharing of multilingual data generated in the localisation process becomes essential.

This data is used to train high-quality, domain-specific data-driven Language Technology services . These services include Statistical Machine Translation (SMT) and text classifiers. This training process is key to addressing long-tail localisation efficiently..

This work adopts a Linked Data approach (Bizer et al, 2009) to the massively scalable sharing, searching and interlinking of Localisation and Language Linked Data (termed here L3Data) on the web. This is accomplished using a semantic schema for L3Data that is compatible with existing localisation data exchange standards and can be used to support the round-trip sharing of language resources. Multilingual Terminology Databases, or Term-Bases, and Translation Memories are shared as they have the greatest impact on conventional localisation productivity and on the training of relevant language technologies such as SMT. Commercially sustainable sharing of L3Data must work within constraints of copyright, confidentiality and competitive concerns (Allemang, 2010), but by reflecting these in the L3Data schema, fine-grained access control rules allow enterprises to flexibly balance them against the benefits of sharing data.

The paper describes our approach to development of L3Data schema and data management processes, web-based tools and data sharing infrastructure that use it. An initial proof of concept prototype is presented which implements a simple web service for segmenting and machine translating content for crowd-sourced post-editing and rating, built on a triple store that tracks steps in this workflow using an integration of the Open Provenance Model (Moreau et al, 2008) and the XLIFF localization interchange format from OASIS (XLIFF, OASIS, 2007).

2 Motivation

In the software industry, the trends towards software as a service offering and smartphone apps has greatly levelled the playing field for small and medium enterprises (SMEs) entering a global market. At the same time, technical documentation has shifted from monolithic technical documents to forms such as FAQ, knowledge articles, wikis and Question-Answer forums, which contain high proportion of user-generated content or vendor content produced by customer care staff asynchronously from the product release (Marcus, 2006). For the localisation industry this represents an increasing challenge in terms of growing demand for content translation coupled with reduced job size and increased heterogeneity of content style and domain, which can be characterised as long tail localisation. How-

ever, in recent decades the principle efficiency gains in the localisation industry have resulted in the sharing of Translation Memories (TM, enabling maximal reuse of previous translations) and Term-Bases (TB) between client, localisation service providers (LSPs) and translators. These resources amount to multilingual terminology databases that ensure consistent use of terms in the source content and consistent translation of those tools. However, the characteristics of long tail localisation tend to weaken these efficiency gains, and especially for SMEs in the localisation chain, reduce the attractiveness of investing in the assembly and maintenance of TMs and TBs.

3 Approach

In summary, SMEs now have near-transparent global distribution opportunities, but they are faced with the need to support a bewildering number of locales and languages to gain the benefit of such worlwide availability. To address this localisation barrier for SMEs engaged as provider or consumers in long tail localisation, we propose a solution that combines the carefully targeted and controlled sharing of language resources with open innovation in the use of data-driven language technology services. Specifically this project will develop controlled sharing solutions for multilingual terminology management to support consistency in source language content authoring and its translation and for parallel text management for leverage as TMs and for the training of statistical machine translation (SMT) engines. The premise is that effective use of language resources in long-tail localisation requires low-cost acquisition of domain-specific, quality-assured resources that are best sourced from prior localisation tasks provided these have systematically integrated quality-related provenance annotation into the resources they produce. The sharing solution takes the form of Linked Data on the Web. This builds on the W3C's Semantic Web Standards RDF, RDFa, RDF(S) and OWL, which allows web content and web data (i.e. deep web content) to be interlinked in a decentralised and distributed manner. The approach is inherently multi-lingual as Linked Data supports multi-lingual data and meta-data representation enabled through Unicode, element-level language tagging in RDF and International Web Resource identifiers.

A key requirement is therefore to develop an extensible meta-data schema for localisation data to support fine-grained, low cost interlinking of source language web content and language resources in the form of multilingual term-bases and TMs as used in the localisation process. This schema will also enable terminology and translations generated in the localisation process to be harvested for further sharing, including its use for SMT training. We refer to the subject of this Linked Data schema as Linked Localisation and Language Data or L3Data. This schema will then support the development of terminology management, parallel corpora management and translation management tools that establish and maintain the links between localised content, its localisation-related meta-data and the shared language resources leveraged during the localisation process. Where appropriate these will be devel-

oped as plug-ins to existing localisation support products. Our aim is specifically not therefore to establish an open access repository of Localisation and Language Linked Data, but to provide instead an open data-linking platform that, with suitably fine-grained and auditable access control, will support low-cost data sharing suitable for new commercial value networks that address long tail localisation. This platform will also then support new innovative localisation support services such as: rapid, domain specific SMT training; interlink discovery services to integrate L3Data in support of new domain-specialised value chains and data cleaning and link maintenance services.

3.1 Support for SMT Training

Data-driven language technologies are demonstrating benefits at different stages of the localisation process such as: SMT to support human translation, named entity recognition for terminology management and text classification for translation review. The development of bespoke SMT engines in particular has become increasingly accessible to SMEs thanks to the popularity of the MOSES open source toolkit (Koehn et al, 2007). However, the effectiveness of these technologies is highly sensitive to the relationship of the data upon which they have been trained to the localisation task at hand. Assuring this in long tail localisation requires low cost, but highly agile and responsive acquisition of training data. This can be achieved by continuous and targeted sharing of linguistic resources that are the by-products of localisation – i.e. translation memories, term-bases and question-answer (QA) annotations of these – across the broadest range of LSPs and their clients. Long-tail localisation inherently requires data linking as the fragmentary nature of the content means that the heterogeneous data must be collated and links kept up-to-date, and places even greater value on agility in sharing to be able to assemble sufficient data to support rapidly changing application domains. Until now there has not been an effective model for rapid language data exchange with suitable provenance needed to address commercial quality and access control concerns.

3.2 Relationship to Existing Localisation Standards

However, this requires high levels of interoperability, both in the linguistic resources that are shared and the language technology-based localisation support services, often from third parties, that would operate over them. Language resources, even when made openly available, exist in information silos with their own distinct schema and access portals. Even when standardised exchange formats are available, such as TermBase Exchange (TBX) for terminology (Localization Industry Standards Association, 2008), Translation Memory Exchange (TMX) (Localisation Industry Standards Association, 2005) or XLIFF for localisation job hand-off format (XLIFF,

OASIS, 2007), they are implemented through tool import/export functions that do not support round-trip consistency management as the language resources are updated over time. Though there have been some proposals to export existing language resources in RDF, such as an RDF mapping for ISO TC37 data categories (see, e.g., Windhouwer and Wright, this vol.), these efforts focus more on rich linguistic annotation of semantic web ontologies (Buitelaar et al, 2009) rather than commercial terminology management for which TBX provide a suitable lightweight starting point. Based on a subset of ISO TC37 data categories, the potential for extending with these richer lexical schema remain. The Multi Lingual Information Framework (Cruz-Lara et al, 2005) extends the ISO data categories work to integrate localisation meta-data as exchanged in TMX and XLIFF, but has remained at the specification stage and does not attempt to exploit Linked Data. The OASIS OAXAL initiative (Zydroń, 2011) proposed mechanisms for linking from source web content to termbase entries (using the Termlink XML schema) and segment-level translations (using the XML Text Memory, xml:tm schema), and can therefore be easily extended to reference L3data potentially held by partner organisations.

3.3 Use of Linked Data Infrastructure

The L3Data approach combines Linked Data architectures with fine-grained access control to address the need of data to be distributed but connectable in the rapidly forming value chains. These are the key properties needed to address the heterogeneous but short-lived jobs typical of long tail localisation. The dereferencable property of RDF-based Linked Data allows decentralised development of unambiguous schema terms that can be subsequently interlinked, potentially by third parties. The approach enables rapid specialisation of language resources to support innovation in what data and meta-data can be readily shared. The W3C's standardisation of RDF and RDFS to author schemas and publish data directly on the web and RDFa to link XML content to RDF elements has resulted in a growing range of performant and robust Linked Data stores ('triple stores') with web accessible query interfaces (for which a standard query language SPARQL, is typically used). However, localisation support systems typically place emphasis on data handling performance, especially for interactive translation tasks such as TM searches. However, an initial comparison on TM look up, using a 3 million word TM typical of the size supported by desktop translation tools, showed only a 20% overhead incurred by an unoptimised off-the-shelf triple store compared to the leading tool (SDL's Trados^{TM}.[1]).

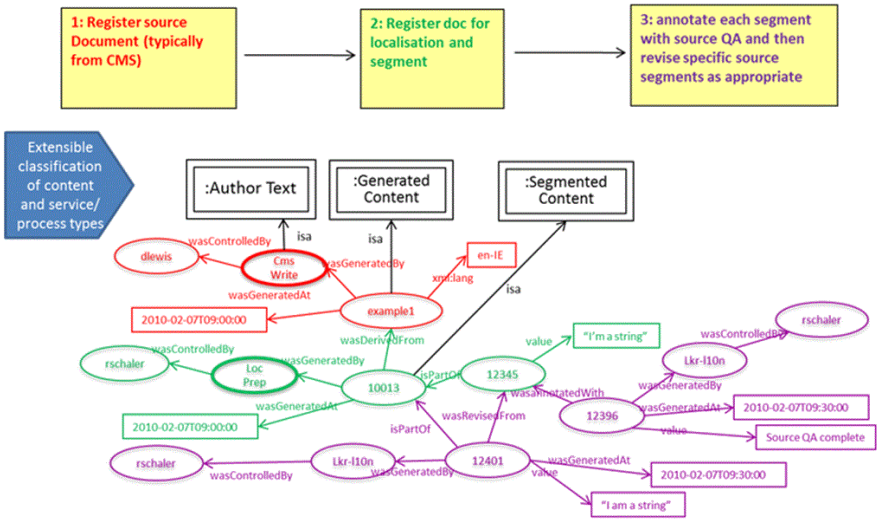

Fig. 1 Example of RDF provenance capture for a three step workflow

4 Implementation

Despite working in a domain that requires controlled access to Linked Data, we follow the decentralised evolution principles of Linked Open Data, in making use of and extending existing RDF vocabulary, specifically the Open Provenance Model (Moreau et al, 2008, OPM). A content processing approach is therefore taken for the L3Data schema, where OPM is used to express the state transformation that operates on content (principally terms and translation units) and its meta-data as the result of content processing by different processes such as authoring/revision, source language question-answering (including terminology usage), TM leverage, SMT usage, human translation, post-editing, target language question-answering etc. Figure 1 shows an example of the RDF provenance data captured for three steps of registering a source document with the system, preparing it for localisation by segmenting and running a controlled language quality check on those segments. This approach can also be used to record value adding operation to shared TM or TB elements, such as adding term translations, definitions or morphologies or identifying terms, or style and domain classification TM entries. OPM serves to capture both the process resulting in the recorded content transform or annotation and the agent responsible, thereby support the management of acknowledgement and credit for shared language resources.

Our initial implementation focused on recording transforms produced by external service in a translation workflow that used XLIFF to exchange job data between different web services implemented as part of the Service Oriented Localisation

[1] http://www.trados.com

Fig. 2 Provenance visualiser screenshot

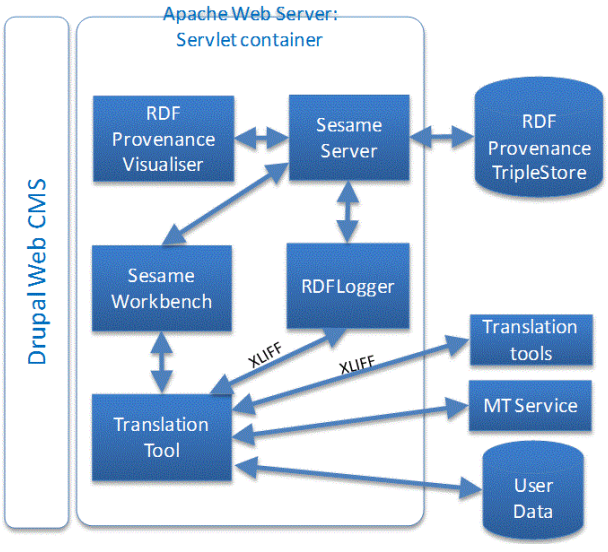

Fig. 3 Content Management System (CMS) Localisation System architecture

Architecture Solution (SOLAS) platform developed by the Localisation Research Centre at the University of Limerick, Ireland. A generic XLIFF to RDF/OPM transformer was implemented, while user defined extensions in the XLIFF recording the workflow activity routing was used to populate the process property of the activity transform. The different forms of translated content and the processes performed on them were classified from taxonomies for localisation content and service categories (Lewis et al, 2010) defined by the Centre for Next Generation Localisation (van Genabith, 2009).

To demonstrate this approach a crowdsourced translation application which has been implemented with a Drupal frontend, via which users can create and contribute to translation jobs in XLIFF. An RDFLogger component is used to change the XLIFF document into RDF provenance statements and then logs these to a triple store. Further, the Sesame Triple Store[2] provides an open source Java framework for storing, querying and reasoning with RDF. A RDF Provenance Visualiser (Fig. 2) has been implemented for exploring outcomes of process steps using the provenance based logging of localisation activities. Subsequently a second application offering a simple web based crowd-sourced translation and rating application was developed and is currently being evaluated. Figure 3 outlines the implementation architecture.

5 Future Directions

Our initial evaluation and proof of concept implementation show that a provenance based Linked Data approach to localisation interoperability is possible, can interoperate with existing standards, namely XLIFF, can provide the basis for web-based translation applications, and operates with acceptable performance for small jobs at least.

Our next steps will be working to establish a platform that will allow L3Data to be shared between the conventional actors in the localisation chain, namely the content developing localisation client, LSPs and translators, but to allow this sharing to occur between value networks of peers (rather than just chains of clients and providers (Allee, 2002)), for instance groups of translators cooperating in addressing a large customer domain. We envisage such shared will be mediated by Language Data Resource (LDR) Curators. The TAUS Data Association[3] is a good existing example of such a curator for TMs, but we could see similar roles for TB and for the interlinking and assembly of LDR for SMT training as depicted in the figure below.

The L3Data platform would need to support common features such as the posting of LDR to a value network (including posting corrections, alternates, new language translation or annotation such as ‘see also’ to existing term or translation segment resources), interlink management, access control (possibly using content annotation using RDF encodings of Creative Commons), annotating LDR with question-

[2] http://www.openrdf.org

[3] http://www.tausdata.org

answering outcomes from translation work (e.g. a poor question-answering rating on reuse of a previously submitted TM segment in a particular context) and auditing of LDR usage in a job conducted within an actor, (in order to help gauge the benefits yielded by participating in a LDR sharing value network)

Acknowledgements This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (http://www.cngl.ie) at Trinity College Dublin.

References

Allee V (2002) The Future of Knowledge: Increasing Prosperity through Value Networks. Butterworth-Heinemann

Allemang D (2010) Semantic web and the linked data enterprise. In: Woods D (ed) Linking enterprise data, Springer, pp 3–23

Bizer C, Heath T, Berners-Lee T (2009) Linked data - the story so far. International Journal on Semantic Web and Information Systems 5:1–22

Buitelaar P, Cimiano P, Haase P, Sintek M (2009) Towards linguistically grounded ontologies. In: Proceedings of the 6th European Semantic Web Conference (ESWC 2009), Heraklion, Greece, LNCS, vol 5554, pp 111–125

Cruz-Lara S, Gupta S, García J, Romary L (2005) Multilingual information framework for handling textual data in digital media. In: Proceedings of the 3rd International Conference on Active Media Technology (AMT 2005), Kagawa, Japan, pp 81–84

van Genabith J (2009) Next generation localisation. Localisation Focus: The International Journal of Localisation 8:4–10

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) MOSES: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007). Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, pp 177–180

Lewis D, Curran S, Jones D, Moran J, Feeney K (2010) An open service framework for next generation localisation. In: LREC 2010 Workshop on Web Services and Processing Pipelines in HLT: Tool Evaluation, LR Production and Validation, Valetta, Malta, pp 52–59

Localisation Industry Standards Association (2005) TMX 1.4b Specification OSCAR Recommendation. http://www.lisa.org/fileadmin/standards/tmx1.4/tmx.htm, retrieved on 25 Feb 2010

Localization Industry Standards Association (2008) Systems to manage terminology, knowledge, and content – TermBase eXchange (TBX). http://www.lisa.org/TBX-Specification.33.0.html, retrieved on 25 Feb 2010

Marcus A (2006) A demand-based view of support: From the funnel to the cloud. Tech. rep., Service Innovation Consortium, San Carlos, CA, retrieved 18/8/11

Moreau L, Freire J, Futrelle J, McGrath R, Myers J, Paulson P (2008) The open provenance model: An overview. In: Freire J, Koop D, Moreau L (eds) Provenance and Annotation of Data and Processes, LNCS, vol 5272, Springer Berlin / Heidelberg, pp 323–326

Windhouwer M, Wright SE (this vol.) Linking to linguistic data categories in ISOcat. P. 99-107

XLIFF, OASIS (2007) Xliff 1.2. a white paper on version 1.2 of the xml localisation interchange file format (xliff). http://xml.coverpages.org/XLIFF-CoreWhitePaper200710-CSv12.pdf, revision: 1.0, 17 Oct, retrieved on on 25 Feb 2010

Zydroń A (2011) Reference model for open architecture for XML authoring and localization 1.0 OASIS committee specification. http://www.oasis-open.org/committees/oaxal/, retrieved 18/8/11

see bib file