Position Paper: Interoperability Challenges for Linguistic Linked Data

David Lewis (dave.lewis@cs.tcd.ie)
Centre for Next General Localisation
Trinity College Dublin

Abstract:
This position paper reviews the growing need for seamless interoperability and interlinking between multilingual and multimedia web content and linked data that captures linguistic knowledge. It outlines some of the state of the art in this area and highlights some interoperability issues that may be fruitfully addressed in the short to medium term.

1 Introduction

The explosive growth of content on the Web in terms of volume, variety and velocity demands new approaches to content processing and analytics. Such approaches must address issues in large scale analysis, interpretation and reuse of heterogeneous data sets that capture human linguistic judgements that originate from different media, in different human languages and from across different jurisdictions, business settings and social/community contexts. Language diversity in particular has become an increasingly important aspect of the Web as it reflects increasing globalization trends.

The market for outsourced language services and technologies was estimated at €33.5 billion in 2012. This market is growing at 12% annually and is crucial to the global trade in goods and services. The process of localisation involves the transformation of content from the language and formats used in a source language to that used in a target locale. Commercial localisation workflows involve content generating enterprises and the Language Service Providers (LSPs) that they contract to translate source content, often with a Multilanguage provider subcontracting smaller single language providers operating in the target locales.

In recent decades, the main technological innovations to yield productivity improvements in these workflows have involved the collection and reuse of language data resources. Specifically these resources take the form of: term-bases, which are multilingual glossaries that improve consistency in both authoring and translation of terms, and translation memories, which are databases of previously translated sentences that assist translators in translating identical or similar sentences, phrases or terms. More recently, translation memories (TM) and term-bases are being reused by LSPs as good quality training corpora for Statistical Machine Translation (SMT) engines.

Language service provision is therefore a strongly data-driven industry, where the sharing and reuse of data along translation value chains is key. The sharing of translation memory and terminology databases within value chains for use by human translators is enabled by well-established transaction discounts, while similar translation data reuse for statistical machine translation is growing rapidly. The industry therefore benefits from a well established model that allows linguistic data reuse to be financially quantified. The key here is that it is the content itself that forms the basis of this commercially valuable data, and it is

the human linguistic judgements that localisation workflows perform on this content, i.e. term identification, translation and linguistic QA, that provides this value.

2 Problems

However localisation value chain workflows can be varied and complex and overheads due to poor data and meta-data interoperability are estimated as being upto 20%. Moreover, the distribution of providers by size exhibits an extremely long tail, with 99% being SME. The industry as a whole therefore faces challenges to both handle the overhead of poor interoperability and to reap the benefits of large scale language data reuse arising from large volumes of translation traffic.

Their low throughput of content for translation means handled by smaller LSPs and small clients means they have little opportunity to amass significant term-bases and TMs as assets to reuse between jobs or to train SMT engines tailored to their domain specialisms and language pairs, legal text translated from French to German. This is compounded by the lack of language resource curation skills needed to maintain these resources. The potential for the localisation industry to benefit from pooling and sharing language resources for training SMT engines has already been recognised, but only realised to date through centralised data sharing models, e.g. the repositories run by TAUS Data Association or LetsMT! In parallel, large online dictionaries, e.g. WordNet, and term data bases, e.g. IATE, EuroTermBank, Terminology-as-a-Service, Eur-Voc, have emerged. However, these systems are closed in the range of resource meta-data that they support. Therefore third parties are prevented from innovating with novel corpora assembly methods that can be supported by new meta-data via data queries, e.g. by translation quality rating, range of target language, term usage or resource similarity to a reference corpora. Further, the sustainability of these language resource curation approaches is limited either through the high cost of access, difficulty in predicting return-on-investment (ROI) or reliance on episodic public funding resulting in unpredictable levels of freshness, linguistic quality and data integrity. Such language resource reuse could therefore benefit from publish parallel sentences and terms as linked data in an open format, to enable third party linking to external meta-data, e.g. generated by text analytics components, without the need to maintain a copy of the original resource.

The high workflow cost overheads is related to poor interoperability between the high variety of content formats in use and the range of workflow systems and translation tools deployed in the industry. For example, language resource reuse in the localisation industry has relied on segment or term matching, which is tightly constrained to domains, i.e. largely useful for similar content from the same client. Consequently, reuse formats, e.g. Translation Memory eXchange (TMX) and Term Base eXchange (TBX), support the delivery of pre-packed resources from clients to LSPs and onto translators, rather than offering queryable repositories from which resources can be easily searched, filtered and retrieved.

3 Current Interoperability Appraoches and Issues

The W3C Internationalization Tag Set (ITS) Recommendation [1] aims to reduce elements of the interoperability overhead cost by defining standard meta-data attribute that can be used to annotate HTML and XML content both in its source content form and in the data flow

[1] http://www.w3.org/TR/its/

formats used within localisation workflows, e.g. the XML Localization Interchange File Format (XLIFF) standardised by OASIS[2]. It is defined as a number of data categories each of which support different individual use cases requiring well defined annotations. The key data here though is the textual content of documents.

Annotation schemes oriented toward the semantic web and linked open data, i.e. RDFa and microdata, are not well suited to this task as text is treated only as literal objects of data triples and not the subject of meta-data annotations. Textual content can be the subject of standoff linked data annotations through the construction of URLs using character offset or hashing of surround text to identify specific text fragments, such as proposed in the NLP Interchange Format (NIF)[3]. However, to support close integration with content processing and localisation tool chains, ITS associated meta-data with with textual content either through well-defined attribute added to enclosing elements (e.g. HTML span) or through rule element that associate attributes with enclosing elements (or attributes) using XPath selectors. Well defined inheritance, override and default rules enable dedicated ITS processor functions to be implemented and conformance tests for such processors to be formulated. Attributes are defined as abstract data categories, with mappings defined for implementing these definitions as attributes and elements in HTML or XML documents or as linked data in the NIF ontology. Data categories target independent use cases, but can also be use productively in combination. Ease of adoption is supported by conformance being attainable through implementation of a single data category.

The original ITS recommendation (ITS1.0) provides data categories that support the preparation of text for translation, e.g. indicating what text should be translated and what should not, identifying terminology use, marking text subflows and providing free form instructions for translators. An extension of the standard, ITS2.0[4], is currently being finalised by the Multilingual Web – Language Technology working group at the W3C. ITS adds further data categories designed to ease the integration of language technologies and linked open data into the localisation process. Machine translation integration is supported by annotation of the content's application domain and of automated translation confidence scores. Text analysis is supported with annotation to associate words or phrases with external resources, e.g. DBpedia for classification and definitions or WordNet or BabelNet for lexical definitions. Such annotation may be generated by text analysis components such as Named Entity Recognition (NER) engines. ITS2.0 therefore offers a flexible palette of well-defined data categories to support the generation and consumption of content annotations by multiple processes and the translation workflow, spanning from content creation to its translation, consumption and reuse. In this sense ITS2.0 fulfils a role for the multilingual Web similar to that which the Dublin Core has played for interoperability of monolingual content publishing.

Open published language resources, such as parallel translated text sentences, e.g. Europarl, and term bases, e.g. EurVoc, provide a means of bootstrapping the development and use of machine translation for SMEs. They have often been subject to curation activities to provide some consistency of quality, but the degree to which this has been conducted is often not clearly reported, and typically not part of the resulting language resource meta-data. Commercial parallel corpora typically reflect the static output of translation projects and lack any indication of the variation in quality that typically occurs across such a resource.

[2] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff
[3] http://nlp2rdf.org/
[4] http://www.w3.org/TR/2012/WD-its20-20121206/

Similarly, terminological or lexical resources assume a flat quality model, with their scope often limited by the quality bar set by the curators of that particular corpus. This quality-agnostic assumption is largely reflected in linked data vocabularies and ontologies that have been used to support publication of linguistic linked data. These include: General Ontology of Linguistics Descriptions (GOLD)[5], RDF binding for ISO TC37/SC4 Data Category Registry (ISOcat)[6], OntoTag[7], the Ontology for Linguistic Annotations (OliA)[8] as well as the NLP Interchange Format[9]. However, localisation workflows produce a wealth of provenance meta-data that can help differentiate quality in different contexts. This includes: details of MT engines used; translation confidence scores; experience and domain knowledge of post-editors; use of terms and coverage of available term-bases. As previously demonstrated via the CMS-LION platform[10], we can log localisation provenance using the W3C's standard RDF provenance vocabulary (PROV)[11], from both language technology components and commercial translation and terminology tools. Predicates and classes from the above linguistic ontologies can then be incorporated as appropriate to the type of component or tool involved. We therefore advocate a more active approach to the curation of language resources. Here passively curated language resources from public sources are re-published as linked data and then annotated, adapted and extended with resources collected from commercial translation projects, providing a progressively richer and more relevant set of assets for reuse. We refer to the on-going harvesting of human linguistic judgements (e.g. translation post-editing and term extraction) and its assembly into corpora for re-training language technology components as **active curation**. Initial evaluation of this approach[12] for dynamic SMT retraining, showed for an SMT engine trained on available public corpora (Europarl) and used to translate domain specific Wikipedia content from Spanish to English, resulted in an automated translation quality score (using the BLEU score) increase from a 0.3395 baseline to 0.4268 after retraining over 4 iterations with a filtered output of 6000 sentence post-edits by non-professional translators. This is an improvement of 25.71% - a considerable margin within the context SMT optimisation research.

We advocate a decentralised approach to curating, locating and reusing language resources for active curation in the language services industry, due to its complex value chains featuring large numbers of small providers. The use of linked-data, by offering fine grained, inter-linked data elements accessible via individual URLs, therefore provides an ideal technical platform for the active curation of language data. This approach allows resource consumers to search and filter over distributed sources at different levels of granularity by using standard meta-data vocabularies and SPARQL queries.

[5] An OWL-DL implementation of GOLD: An ontology for the Semantic Web, Farrar & Langendoen, in Linguistic Modeling of Information and Markup Languages, Springer 2010

[6] Linking to linguistic data categories in ISOcat, Windhouwer & Wright, in proc Linked Data in Lingusitics 2012

[7] OntoTag: A semantic web page linguistic annotation model, Aguando de Cea et al, in proc ECAI'02

[8] An ontology of linguistic annotations, Chiarcos, LDV Forum 23(1): 1-16, 2008

[9] http://nlp2rdf.org/nif-1-0

[10] On Using Linked Data for Language Resource Sharing in the Long Tail of the Localisation Market, Lewis et al, LREC 2012

[11] http://www.w3.org/2011/prov/wiki/Main_Page

[12] Retraining MT with Post-edits to Increase Post-editing Productivity in Content Management Systems, A. Toral, L. Finn, D. Jones, P. Pecina, D. Groves, D. Lewis, International Workshop on Expertise in Translation and Post-editing Research and Application. Copenhagen, Aug 2012

4 Summary

The ITS2.0 standard can be seen as an initial set to provide an equivalent to the Dublin Core schema for annotating content and interlinking it with meta-data and thereby form the foundation of a linguistic linked data cloud. Other initiatives such as the NLF Interchange Format and ISOcat for lexical information also fill important parts of the ecosystem needed to enable linguistic knowledge to be extracted, exchanged and shared as linked data. However there are still important missing components in the representation of language- and media-specific information that are required for the correct interpretation of this data - across different media and across the increasing variety of human languages used nowadays on the Web. In this paper, we highlight the issues developing such an ecosystem of interlinked, and semantically interoperable language resources (corpora, dictionaries, lexical and syntactic metadata, etc.) and media resources (image, video, etc. metadata) that will allow for low-cost and open exploitation of such resources in multilingual, cross-media content processing and analytics.

Several groups are investigating this topic and are exploring broader international collaboration: W3C Internationalization Activity and the W3C Multilingual Web-Language Technology WG (working on ITS2.0); the Ontology-Lexica W3C community group and the Open Linguistic Working Group at the Open Knowledge Foundation (working on NIF). Collaboration is currently aimed at the development of the linguistic linked open data cloud and its relationship with multilingual web. We seek to broaden links and collaborate on interoperability topics related to:

- Internationalization of annotated content to support consistent extraction of Linguistic Linked Data resources across languages (involving best practice in applying ITS2.0 with HTML5 and specific XML vocabularies)
- Provenance of linguistic linked open data and how that can be tied to language service industry best practice in language quality assessment (integrating provenance ontology[13] into linguistic linked data, such as the NIF vocabulary);
- Leveling the current anglo centric bias of LOD - making cross lingual resources available as linked data seems a good place to start (see work of the MONNET project[14]);
- Taking techniques established by ITS for linking multilingual textual content with stand-off provenance and quality meta-data and applying it to multimedia and multimodal web content, aligning with W3C work on Multimodal Architecture[15], Media fragment URIs[16] and media ontology[17].

[13] http://www.w3.org/TR/prov-o/
[14] www.monnet-project.eu
[15] http://www.w3.org/TR/mmi-arch/
[16] http://www.w3.org/TR/media-frags/
[17] http://www.w3.org/TR/mediaont-10/