Open Space session at the W3C MultilingualWeb Workshop 16 March 2012: MultilingualWeb, Linked Open Data & EC "Connecting Europe Facility"

By Christian Lieske & Felix Sasaki

Overview

As one of the Open Space breakouts, Kimmo Rossi (European Commission; EC) and Christian Lieske (SAP AG) proposed to look into the status quo, and possible actions concerning the intersection of the MultilingualWeb Thematic Network, Linked Open Data & the EC's "Connecting Europe Facility". The breakout attracted attendees from constituencies such as users/service requesters (e.g. users and implementers of Machine Translation Systems), facilitators (e.g. the EC), Service providers (e.g. for translation-related services), and enablers (World Wide Web consortium, W3C). Outcomes of the breakout included a first step towards a mutual understanding of the topic and subtopics, information on actions that already have been started, and suggestions for follow-ups.

Connecting Europe Facility (CEF)

To kick off the discussion, Kimmo Rossi, the European Commission's officer for the Multilingual Web Thematic Network (and the Network's successor, the Multilingual Web – Language Technology project), explained the most likely new context for language technologies in view of the ongoing reorganisation of DG INFSO. As in the 7th Framework Programme, language technology is most likely to remain in close relation with the Data challenge, but the link is likely to become even stronger, necessitating a strategic positioning of language technologies as a **contributor** to the Data Challenge (big data, linked data, open data, data value chain, semantic web).

One of the main areas of LT-related activity at the EC is likely to be the "Connecting Europe Facility (CEF)". This funding program is currently in a status of legislative proposal, under discussion at the Council and the European Parliament. The CEF proposal foresees commoditized multilingual access to online services "from the tap in the wall". CEF is a deployment program, not research: aim is to build infrastructure and services. The CEF proposal also includes an "open data infrastructure" – a European linked open data portal providing standardised one-stop access to EU public sector data in a standardised format.

The future LT infrastructure may include some or all of the following as its initial constituents:

1. MT@EC project – Machine Translation based on the Moses Open Source system; rolled out to European institutions, to serve European eGov services.

2. Multilingual access to pan-European eGov services such as business registers, patient records, etc.
3. Support infrastructure for multilingual virtual meetings and conferences.

In the coming months, it is important to progressively define and specify the CEF infrastructures for multilingual services and data, as the success of the legislative proposal will require us all to bring forward credible, understandable arguments and solid implementation roadmap for the planned actions. The continued input from the MultilingualWeb community will be instrumental to make this happen.

Linked Open Data (LOD)

Christian Lieske from SAP AG provided additional baseline input for the discussion: Since EC is targeting a large scale deployment of Language Technology/Natural Language Processing (NLP), its relationship to Linked Open Data (LOD) needs to be considered:

1. LOD can help to build NLP – LOD can for example generate raw data to build statistical NLP engines/models
2. NLP can help to LOD – NLP can for example help building "clean" LOD data sets, or linking data items automatically; Iván Herman's presentation emphasized that links are the bottleneck in Semantic Web …

For the sake of discussion, a pragmatic definition for LOD was given:

a. machine readable
b. not proprietary formats
c. following Semantic Web standards
d. freely accessible and usable

Structure of the Discussion

To facilitate progress, the breakout was broken down into two areas:

a. Review of status quo with view towards challenges
b. Identification of immediate actions

Hypothetical example of these steps:

a. Review:
    - Wiktionary needs links across languages
    - Currently, we use NLP to find concepts behind the Wiktionary entries, for interlinking the different language Wiktionaries
b. Action:
    - Discuss with Wikimedia foundation what needs to be adapted in their infrastructure to support a concept-based approach

Status Quo

Existing work already builds bridges between NLP and LOD:

- Lemon: a lexical data model compliant with the Semantic Web (in RDF), and the Lexical Markup Framework (LMF)
- EC funded Monnet project
- Dbpedia work in Leipzig
- Wikidata project to put data at the core of Wikipedia (Wikipedia pages are generated out of that data - could help to clear mess of wikipedia data across languages)
- META-SHARE may already provide open data for rule-based NLP

There are various open questions related to the NLP and LOD infrastructures:

a. Do there need to be different infrastructure (one for private sector, one for public sector, one for supplementary sectors)
b. How can as much data as possible be made public without violating regulations and laws?
c. So far there has been a strict separation between terminology and lexical resources, translation memories etc. Do these silos need to be entertained? Integration may help to move NLP to next level.
d. Very often, NLP focussed on prose text. However, industry is about e.g. data in product catalogues, master indices (e.g. for healthcare infrastructures), ... How will NLP perform in this realm?
e. Annotations and their Provenance are important. How can trustful annotations be brought to life? How can responsibilities for annotations be distributed (e.g. between humans and machines)?
f. Which setup (in terms of organizational entities and technical infrastructure is needed for collaboration and self-organization?
g. Health Care, eGov or other application areas are not yet really involved in the discussion. How can these constituencies be reached?

Possible Immediate Actions

- Create base definitions for both NLP and data people. One page as a start is enough.
- Gather use cases via in W3C Ontolex group and W3C MLW-LT group (questionnaire for MLW-LT project https://www.w3.org/2002/09/wbs/1/mlw-lt-requirements/)
- Talk to eGov people at W3C (develop usage examples like “eCommerce portal translation”)
- Take discussion to Dagstuhl event in September (not open to the public)
- Take discussion to Boston Semantic Web event early November
- Investigate possibility of session at a 2013 event hosted by the Globalization and Localization Association (Gala)
- Discussion opportunity: 11-12 June MLW-LT workshop, Dublin

Summary to the Workshop Participants

Participants: Creators of content, providers of infrastructure, facilitators e.g. from EC or W3C.

General question: How to continue work started in Multilingual Web (MLW) Network with a view towards Linked Open Data (LOD) and Connecting Europe Facility (CEF).

- MLW fostered (net)working across communities and discussed Natural Language Processing in open real world contexts. Companies normally work with NLP behind closed doors; MLW brought some of this to the open.
- The Linked (Open) Data community puts lots of machine-readable language data on the web according to the principles of the Semantic Web.

The CEF will centre upcoming funding opportunity and challenge around open data and language related services.

Discussion in the breakout session:

- Still hard to find good freely (re)usable NLP resources (MT systems, rules for them, dictionaries etc.) as LOD on the Web.
- Change of paradigm for data creation is needed; currently hard to have a unified "picture" across languages (example "population of Amsterdam" differs across Wikipedia language versions). New paradigm: Create resources in one language or even language agnostic/neutral and use (automatic) approaches to generate resources across languages. Example: Monnet project.
- Need for more annotation, language neutral representations, easier collaboration e.g. to enable crowdsourcing.

Immediate action items to which breakout participants committed:

- Make relationship between MLW, linked open data and CEF facility easy to understand for everyone: What is LOD, how does it relate to MLW? Iván Herman, Paul Buitelaar and Felix Sasaki have already started working on this.
- Gather more use cases for language resources and language related services. Paul Buitelaar – on behalf of the W3C Ontolex group – will send out call for proposals. See also the questionnaire for MLW-LT project; https://www.w3.org/2002/09/wbs/1/mlw-lt-requirements/ (deadline: end of March)
- Pedro Diez will try to get LOD + LT into a Gala event in 2013.