LT-OBSERVATORY

ASSESSING THE OPERATIONAL USABILITY OF LANGUAGE RESOURCES FOR THE PURPOSE OF MACHINE TRANSLATION

Venue:
The International Auditorium
International Trade Union House
Boulevard du Roi Albert II, No. 5 / 2
B-1210 Brussels
Date: 4 December 2015
12.00-16.00 Hrs.

BRIEFING

The following notes are intended to inform and facilitate the discussions with experts:

1. The **LT-Observatory (LT-O)** is organising the present **workshop with experts** to examine a number of critical issues impacting the construction of a user-centric Language Data Catalogue.

2. This Catalogue aims to facilitate access to / promote language resources (LRs) that provide maximum "**operational usability**" for users of statistical machine translation (MT) applications. The overall aim is to unlock the value of existing resources (often paid for by the public purse) by making this value more effectively available to the European MT community, and eventually other LT constituencies.

3. This MT community covers a broad range of stakeholders, including:

   - today's Language Service Providers that deliver MT and post-editing services to their commercial and public sector clients
   - translation technology suppliers that build or support systems used by public and private services
   - in-house managers of translation projects in very large global companies that prefer to keep their translation business in-house
   - government officials that wish to experiment with the development of MT services for their own applications
   - players of any kind in tomorrow's Digital Single Market who want to build multilinguality into their apps and services.

4. LT-O has now identified a first set of publicly available LRs from databases of language data around Europe, and built a small **central repository prototype, with the aim of testing the critical steps** in developing a broad-based heavy-duty LR service.

   The next step is to **evaluate this first packet of resources on the basis of expert input**, and where necessary re-engineer it into a more practical and powerful asset.

This workshop therefore seeks **input from experts** in the field of MT engineering, by asking them to:

- Examine and evaluate the "**operational usability**" of the catalogue's LRs
- Select collegially the **"top ten" LRs** (i.e. those maximally usable in one MT application or another). This set will act as a benchmark for further evaluation activities.
- Decide on the **minimum critical selection criteria** ("the resource fingerprint" to be expressed as metadata) that underlie this most "operationally usable" set of LRs (e.g. size, speed/ease of access to the LR, domain relevance, language ranges, processing costs, speed of implementation, or other decisive criteria)
- Produce a **practical set of metrics** on the basis of these criteria that could be easily applied to any other LRs collected from project, university, public or private repositories around Europe
- Use these criteria to identify **a series of "strata" representing the descending operational usability levels** (the best 10 LRs, the next best 20 LRs, then a third tier that could be optimised, etc.)
- Provide possible pathways to estimating the potential **size of the entire EU LR population** – where are these LRs, which resources could easily be overlooked when surveying the field, unexpected sources of LRs, LRs that need more technical work to become usable, together with any information on who might know more about them?
- Evaluate the work needed to **cyclically select the best language resources for each "usability level"** to reach a target number of "best resources" in the EU. (The aim is to build serious LR critical mass that can then prepare for the next stage of the LT-O agenda outlined below).

5. Once operational usability has been satisfactorily defined and agreed on, and a process is in place to apply it on a practical, recurrent basis, this LR evaluation and estimation exercise should (in due course) be able to provide expert opinion on the following **longer-term goals**:

   - Establish practical evaluation criteria based on operational usability that could be easily applied to resource identification for the whole community (i.e. not requiring highly specialist skills but leveraging the common-sense experience of any MT stakeholder). The ultimate goal here is to **open-source LRs in the same way as technologies can be open-sourced.** This could liberate the management & supply of LRs from public (and perhaps even private) owner/user siloes and automatically make them available via an easy-to-access API model or similar.
   - Identify collective ways in which **LRs identified from the currently "lower" usability strata could be optimised/improved/upgraded**, using existing or emerging technologies (often developed by EU projects).
   - Following on from (2), foresee the kind of new **projects that might best kick-start a "language resource optimisation factory".** This could include, for example:
     - how existing parallel corpora can be easily (or automatically) cleaned,

- how textual alignments etc. can be extracted from one LR and integrated into another to build a new multilingual resource,
- how LRs can be generated massively/automatically to build new language pairs,
- how LR evaluation and optimisation can be carried out *independently* of a specific use case, or alternatively *in close relation to* a growing range of typical use cases.
- The motivation for such a factory is the need to rapidly open up new markets and use cases for MT, especially in the context of the Digital Single Market.

➢ Propose a simple roadmap of the opportunities (and threats) for moving from where we are today to where we want to be tomorrow. This means **envisaging something like an "LR infrastructure" for a very large range of optimised, open LRs.** This kind of enabling infrastructure would not only help respond to new translation needs, but could also provide easy access to a much broader range of LRs to fuel a new generation of innovative MT, text analytics, and other LT-related solutions by SMEs.

6. The **expected takeaways** from this charrette are as follows:

➢ Consensus on the benchmark of the best 10 and then best 30 LRs

➢ Identification of the third tier of LRs that can serve a basis for further optimisation work

➢ Clearer understanding of the relevance of the resource "fingerprint" approach to LR usability

➢ Clearer understanding of how existing LRs can be improved/optimised in some way

➢ Identification of gaps/problem areas in LR curation/creation/usability that the market cannot yet solve.

The workshop will feed into a **Dialogue Day**, to be organised in May or June 2016, at which the LT-O platform, LR Catalogue and benchmarking methodology will be presented to a broad cross-selection of stakeholders.