"Assessing the operational usability of languages resources for the purpose of machine translation"

Brussels, 4 December 2015
Workshop Report

Attendance

Thierry Etchegoyhen, VICOMTECH
Andrew Joscelyne, LT-Innovate
Margaretha Mazura, EMF
Sussi Olsen, University of Copenhagen and colleagues (via Skype)
Blanca Rodriguez, Zabala
Gert Van Assche, Datamundi
Joachim Van Den Bogaert, CrossLang
Philippe Wacker, LT-Innovate
Bartholomäus Wloka, University of Vienna
Andrzej Zydrón, XTM International

Purpose

The purpose of this four-hour workshop was three-fold:

- gain vital feedback from a group of experts on the nature and quality of the first set of language resources collected in the LT-Observatory Catalogue, with a focus on usability in an operational business/commercial context.

- work with experts to fine tune the essential criteria that would render the Catalogue as useful as possible to anyone seeking practical information about language resources in future. The LT-O criterion is "operational usability" (the ability for a potential user to inspect the metadata or description and rapidly decide whether it is relevant to his/her task) and requires that the language resources offered by the catalogue should be maximally "mission-ready."

- explore any other ways to open up the LT-O vision of the Catalogue to the best possible opportunities for further developments, e.g. in Open Linked Data, emerging technology enhancements, and changing needs among users of translation automation services.

The group of experts was briefed in advance of the meeting and through an introductory presentation (see Annex).

Key points raised on "operational usability"

In general, the experts present were positive about the first version of the Catalogue and user interface. This made it easier to deepen their engagement with the tasks at hand. Here follows a point by point résumé of their comments on the features of the LT-O Catalogue:

1. **Content variety**: The Catalogue contains a varied mix of resources (terminology, tabular information, aligned corpora, parallel corpora, etc.) This means it is vital to have a user-friendly description of each resource so there is no confusion about which task a given translation resource can serve.
2. **Terminology**: Some terminology resources in fact consisted of word lists extracted from parallel corpora. These are of little use to a translation automation user as they have not necessarily been validated or even checked.
3. **The OPUS collection of corpora** was collectively deemed to be the most valuable collection listed in the Catalogue. This is due to the fact that it has high standards of resource description, and selects only the best curated content. It also contains already 80% of the resources discovered elsewhere by the LT-O project. This suggests that some of the current research-oriented resources in the Catalogue would need to be filtered out.
4. **Versions**: It may be necessary to specify in different cases which version of a given resource is available at the website in question. Some widely-used resources have been through various transformations and there versioning would be necessary to identify the latest version. Other resources such as the Belgian Official Journal (Moniteur/Staatsblad) in 2 languages are dynamic resources that are constantly updated.
5. **Human checking**: It is currently believed that only extensive human checking of a given corpus will ensure the level of "operational usability" that LT-O wishes to provide. There are (and will be more) technology solutions to "validating" a corpus for certain features, but human checking of the entire corpus would be the ideal procedure to ensure "mission-ready" resources.
6. **Monolingual resources**: Statistical MT systems require not only bilingual but monolingual resources which are used to train the engine to produce a language model that provides a fluent output in the target language.
7. **Language data formats:** TMX is the best minimum file standard to recommend for parallel translation resources. XLIFF files are an alternative.
8. **Contacting the resource owner**: It should not be necessary to contact the owner (as is requested in some repositories). This is time consuming and unnecessary.
9. **In-domain MT system requirements**: it is estimated that a system would need 5 million tokens/500,000 segments of language data in the source to minimally build an MT system for a particular task.
10. **Benchmarking**: Europarl, MultiUN are considered "best in class" or model parallel corpora.
11. **Actions to be taken**: LT-O should draft a set of guidelines for best practices based on the benchmarking input from the experts. This will ensure that language resource suppliers gradually standardise their metadata or descriptions of their resources when sharing them in future.

Key criteria to assess "operational usability"

- Data is available for commercial use, ideally for free or at a low/reasonable price
- Data is "low noise", i.e. has undergone human validation or curation
- Data is "in-domain"
- Data sample is available
- Contact person is clearly identified

Key points raised about the catalogue interface

1. **Source / version:** Many repositories are not the primary data source. It would be useful to list all available sources for one and the same LR combined with versioning information wherever possible
2. **Standardised metadata**: Due to the variety of datatypes, it would be advisable to apply a standard type of metadata that covers all such resources. The Dublin Core offers a useful resource definition scheme.
3. **Sampling tools**: Users could benefit from a sampling tool so they could check a given resource to see if it is adequate to needs. Alternatively, potential downloaders could benefit from a preview. However, as the resource will not be hosted on the LT-O Catalogue, downloading does not apply directly to the project. LT-O should encourage LR Repositories to provide preview and/or sampling as a sine qua non functionality.
4. **Related quality check tools**: A number of related tools were mentioned, including the outcomes of the recently finished FALCON project, which aims to provide various techniques to "fingerprint" resources so as to make them easy to evaluate by potential users.
5. **Rating systems**: The proposed 5-star rating system that users of resources could apply to the Catalogue was considered as problematic as ratings very much depend on the context in which a LR is used). It was suggested that it should be replaced by a simple way of "flagging" resources that present technical problems.
6. **Fields**: There was intense discussion about the fields needed on the Catalogue interface to facilitate the user experience. Overall, there appears to be a *maximum* model (including such fields as size, version, licence information, types of segmentation (paragraphs or sentences, etc.), error flagging mechanisms, user corpus classifying) and a *minimum* model, whereby the search engine would automatically filter the resources that matched the search term, and users would not need to add terms to the fields to select the best resource.
7. **In-domain resources**: In-domain resources are critical to successful, high-quality machine translation. A search/filter over resource domains would therefore be very useful. To facilitate resource domain classification, the Catalogue should provide a "standard" list of domains that cover current resources (using that found in the IATE term service for example, or other existing practical categorisations used by the translation community), combined with an open field which would enable the Catalogue to collect additional domain information from users.

Key success factors for the Catalogue

- Data can be filtered by language pairs
- Data can be filtered by domain
- Data well documented (metadata is standard and complete)

Conclusions on language resources and Open Data

A final discussion raised the issue of the role of Open Linked Data, which has been explored more deeply in projects such as BabelNET and FALCON, which propose open source solutions for very large multilingual semantic support in translation and other language-based operations.

More crucially for the future of translation in Europe, Open Data generated by governments and shared using standard file formats opens up a new opportunity for multilingual data – i.e. existing translations in "multilingual" countries or in countries that practice government translation for its minority or immigrant communities – could form a new source of translation data that LT-O could examine and ultimately refer to.

At the same time, some of these data would require pre-processing using a range of LT tools that have become available in recent EC projects. This suggests there is a longer-term need to associate translation data with LT tools of various kinds that can improve or support various kinds of existing but current "hidden" translation data in the EU.

In terms of a general software environment capable of hosting a range of relevant software solutions to help Catalogue users consult and process resources more closely, one option to explore in the medium-term would be to create an open source LR factory or to develop a Language Resources service in the context of the LTI Cloud.