Linked Data for Language Technology, 21st March 2014, Athens

The Need for Clean Language Data and the Rules for Cleaning it

Costas Nadalis

CostasNadalis@TMServe.gr

Language Data

- Language data is usually monolingual, bilingual or multilingual data in the form of text, translation memories or glossaries.
- It has been generated through the use of text editors, word processors, web editors, DTP software as well as Computer Aided Translation (CAT) tools.
- Each tool stores language data in proprietary formats together with language irrelevant information (e.g. formatting, structure elements etc.) aimed to serve the specific tool functionalities and the user needs.
- Standards and open formats such as HTML, RTF, XML, TMX, TBX, XLIFF etc. made language data available and allowed the interexchange and utilization by different tools, adding , at times, further language irrelevant information.
- During the recent years, the drive to share and consume data, the changes in the translation environment and the increase in the usage of Machine Translation (MT) systems, have amplified the need for more language data of good quality.
- Removing non language related information and improving the quality of existing, private, purchased, shared and publicly available language data is a frequently repeated task on lack of readily available suitable data sets (for use in MT training or in CAT tools for example).

Cleaning Language Data

While the extent and the form of language data cleaning needed is argued among experts and vary from tool to tool and the intended use, some of the basic steps involve:

Encoding

- ➢ Convert all data to the appropriate encoding (e.g. UTF-8).

Characters

- ➢ Convert character elements to true characters (e.g. &#60; , &#x003C;, &lt; to <).
- ➢ Fix (remove/replace) any invalid or corrupt characters (e.g. control codes etc.).
- ➢ Simplify and unify unnecessary character variations (e.g. –, –, —, — to -).
- ➢ Remove extra spaces, tabs etc.

Markup (formatting, placeholders)

- ➢ Remove all known markup tags (e.g. <b>, <\b> etc.).
- ➢ Remove all RTF formatting elements (e.g. {\f0\fswiss Helvetica;}\f0\pard … \par}).
- ➢ Find any remaining unknown tags and remove or replace as needed (e.g. {>1<}, <1>).

Cleaning Language Data

Units (Us)/ Translation Units (TUs)

- Remove Us/TUs with Source text in a different language.
- Remove TUs with no translation (Source = Target).
- Remove TUs with some/much untranslated text in the Target.
- Remove too short Us/TUs (too many numbers/too few letters).
- Remove TUs with Source too different in length from Target.
- Remove multiple copies of the same U/TU.
- Etc.

Quality

- Correct misspellings, and harmonize spelling variations.
- Fix punctuation, inconsistent numbers and other easily spotted errors.
- Improve language and terminology consistency.
- Etc.

What is Needed

- Definitions for clean language data “types”, based on the needs, tools and intended use.
- Specifications on what is clean language data for each “type”.
- Guidelines for the proper cleaning of language resources.
- Data and/or recommendations for specific language and language pair values relating to data cleaning e.g. min., max., average sentence length etc.
- Tools to support big language data handling and help in the language data cleaning and quality enhancements.
- Clean language data readily available instead of, or in parallel to, the original unprocessed data in linked data or traditional repositories.
- Clean language data made available through automated cleanup of the original data from linked data or other sources.

Linked Data for Language Technology, 21st March 2014, Athens

The Need for Clean Language Data and Rules for Cleaning it

Costas Nadalis

CostasNadalis@TMServe.gr