TaaS - Terminology as a Service

Andrejs Vasiļjevs, Tatiana Gornostay {andrejs, tatiana.gornostay}@tilde.lv
Tilde
http://taas-project.eu

Description

The main objective of the FP7 project Terminology as a Service (TaaS) is to establish a cloud-based platform for acquiring, processing, and reusing multilingual terminological data to benefit both human translators and statistical machine translation systems.

TaaS platform provides a variety of online terminology services for human users – translators, technical writers and terminologists:

- Processing of documents uploaded by users in different formats: DOC, DOCX, PDF, XML-based formats like XLIFF, and others;

- Automated extraction and retrieval of monolingual term candidates (from documents uploaded by users) using state-of-the-art linguistically and statistically motivated terminology extraction techniques;

- Automatic lookup for term translation equivalents (for monolingual term candidates automatically extracted from documents uploaded by users) by querying public terminology databases, such as IATE, EuroTermBank and others, as well as automatically extracting from parallel and comparable Web data;

- Creation of monolingual and bilingual terminology collections in user-defined languages;

- Collaborative terminology data clean-up, e.g., deletion of irrelevant or unreliable term candidates, validation of term candidates in context, etc.;

- Sharing of resulting terminological data with major terminology databases and banks;

- Reuse of terminology collections in various applications within different human and machine usage scenarios via the TaaS application user interface (API) and export of files in different formats widely exploited by users, e.g., TSV, CSV, and TBX.

TaaS demonstrates the efficacy of its terminology services within the following usage scenarios:

- For language workers, to simplify the processing, storage, sharing, and reuse of task-specific multilingual terminology;

- For computer-assisted translation (CAT) tools, to provide instant access to term candidates and translation equivalent candidates via the TaaS API;

- For statistical machine translation (SMT) systems, to facilitate the domain adaptation by a dynamic integration with TaaS-provided terminological data.

TaaS terminology services improve quality of domain specific SMT by integrating relevant terminology in the SMT training and translation processes. We elaborate several techniques to adapt SMT for under-resourced domains, such as:

- Addition of bilingual term pairs that are extracted from domain specific parallel and comparable data;

- Addition of domain-specific language model built on monolingual in-domain comparable corpora;

- Introduction of an additional feature in translation model indicating phrases containing in-domain terms.