CrossLang Moses SMT Production System

Joachim Van den Bogaert, CrossLang / joachim@crosslang.com
Kim Scholte, CrossLang / kim.scholte@crosslang.com
http://www.crosslang.com

Description

Overview

CrossLang has developed an industry-grade Machine Translation (MT) and Post-Editing (PE) pipeline for the translation of high volumes of data. At its core, the system consist of a multiplexer/router, which allows source documents to be translated with any MT system connected to it. This ensures vendor independence for both CrossLang and its clients.

To facilitate the deployment in typical translation environments, the system features a SOAP XML interface and a dedicated connector to the SDL WorldServer translation management system. This standard set-up allows for very complex workflows including Translation Memory (TM) leveraging and the combination of offline and online translation.

For use in a crowd-sourcing environment, a lightweight Post-Editing platform has been added. The rationale is to allow domain-specialists, rather than translators, to rapidly review MT output for highly technical documents. This makes the acquisition of expensive CAT (Computer Aided Translation) tools and training unnecessary and speeds up the time-to-market.

The pipeline was developed to provide a cost-effective and rapid way to implement continuous localization, as opposed to project-based translation which typically involves a translation agency and the related project management overhead. Additionally, the MT-neutral implementation reduces client-side development costs and allows for multi-system scenarios.

Features

Hardened Moses SMT (Koehn, et al., 2007)**:** for clients with full customization needs, the hardened Moses SMT implementation is probably the most attractive feature. A service layer on top of Moses SMT provides redundancy, load balancing, asynchronous processing, failover support, industry-standard document format support, alignment-based tag-handling, improved normalization and hardened (de)tokenization and (lower/real)casing.

Separation of concerns: the hardened Moses SMT set-up allows the deployment of third-party translation and language models, while still providing the text engineering capabilities built on top of the translation workflow. Named Entity Recognition (NER) and terminology management services, for example, can be added without disrupting the Moses SMT models. From a technical point of view, tagging and annotation are considered to be engineering issues, which are not allowed to interfere with the linguistic issues, as addressed by the SMT models. From a commercial point of view, clients can have third parties focus on linguistic quality, while the CrossLang system can take care of immediate production use.

Scalability and extensibility: the CrossLang system can be modified for performance or quality by adding extra hardware or processing steps through a unified API. The multiplexing capability makes it a suitable platform for MT systems combination.