Work package 3: Text Analytics

Description of work

Description of work

Task 3.1: Classification of forum content (acrolinx, SYMANTEC, month 10-16)
Initially, it will be important to automatically classify the type of content present in forum posts. For instance, specific utterances such as questions will have been differentiated from answers. We will also investigate specific types of questions, matching simple categories such as “compatibility”, "troubleshooting", "licensing", etc. The classification rules will be implemented in the acrolinx IQ rule formalism. The acrolinx IQ language technology server will be used for annotating the data.

Task 3.2: Classification of NGO content (acrolinx, month 16-22)
In the NGO domain, a similar simple content taxonomy will be developed with a view to classify the data. The rules will be implemented in the acrolinx IQ rule formalism. The acrolinx IQ language technology server will be used for annotating the data.

Task 3.3: Sentiment analysis (acrolinx, month 22-30)
We will develop rules to automatically determine that content is heavily charged with positive or negative sentiment. Rules will be developed for all of the project’s languages (English, German, French and Japanese).

Task 3.4: Analyze relationship between classification output and PE data (acrolinx, SYMANTEC, month 30-36)
We will analyze the relationship between the classification data obtained in Tasks 3.1, 3.2 and 3.3 and the PE data obtained in WPs 7 and 8 (productivity data, quality data, usage data). We wish to establish whether certain content types lend themselves better to a specific type of post-editing (monolingual, bilingual).
We will also investigate whether Sentiment Analysis data can be used to check that translations (after editing) correctly conveys source sentiments. We will investigate whether the system can be set up to check that sentiment polarity (whether an utterance was positive or negative) is maintained from source to target.

Deliverables	Delivery date	Description
D 3.1	PM 16	Taxonomy of forum content and rules for automatic classification.
D 3.2	PM 22	Taxonomy of NGO content and rules for automatic classification.
D 3.3	PM 30	Rules for automatic sentiment detection.
D.3.4	PM 36	Report of evaluation results.