SUMAT: An online service for SUbtitling by MAchine Translation

European Commission
Information and Communication Technologies Policy Support Programme
CIP-ICT-PSP.2010.6.2 - Multilingual online services
Pilot Type B
270919
http://www.sumat-project.eu/

List of partners
Vicomtech, Spain (coordinator)
Titelbild Subtitling and Translation, Germany
Athens Technology Center, Greece
Univerza V Mariboru, Slovenia
Invision Ondertiteling, The Netherlands
Voice & Script International Limited, United Kingdom
Deluxe Digital Studios, United Kingdom
Applied Language Solutions, United Kingdom
Subcontracted: TextShuttle, Switzerland

Project duration: April 2011 — March 2013

Summary

SUMAT aims to increase the efficiency and productivity of the European subtitling industry while enhancing the quality of its results via the effective introduction of SMT technologies into subtitling processes. In order to achieve this, we will develop an online subtitle translation service addressing nine different European languages divided into the following 14 language pairs: English-German; English-French; English-Spanish; English-Dutch; English-Swedish; English-Portuguese; Slovenian-Serbian. During the first year of the project the consortium's subtitling companies have provided large amounts of professionally produced parallel and monolingual subtitle data, which have been processed into a form suitable for training SMT systems. Baseline SMT systems are being created using the Moses SMT training scripts and decoder and the IRSTLM toolkit. In the near future, subtitles will be enriched with linguistic information and the baseline SMT systems for subtitling will be built upon by: augmenting language models with extra monolingual target data and improved use of linguistic information; enhancing translation models through the use of POS tagged data and factored models; using compound splitters, named entity recognizers and additional lexica to deal with unknown words; and investigating hierarchical decoding to make use of syntactic dependencies.