SUMAT

CIP-ICT-PSP-270919

An Online Service for SUbtitling by MAchine Translation

Annual Public Report 2012

Editor(s):	Arantza del Pozo
Contributor(s):	Mirjam Sepesy Maucec, Lindsay Bywood, Yota Georgakopoulou
Reviewer(s):	Consortium
Status-Version:	Final
Date:	15th November 2012

Table of Contents

1. Introduction

Subtitling and subtitle translation face some important problems that are preventing the expansion of the market and are hindering new business opportunities: cost, time and quality. There is a clear need to increase the productivity of current subtitle translation procedures, reducing costs and turnaround times while enhancing the quality of the translation results.

Research in such line has shown that subtitling and audiovisual translation could greatly benefit from the introduction of Statistical Machine Translation (SMT) followed by post-editing, in order to increase productivity and enhance the quality of results.

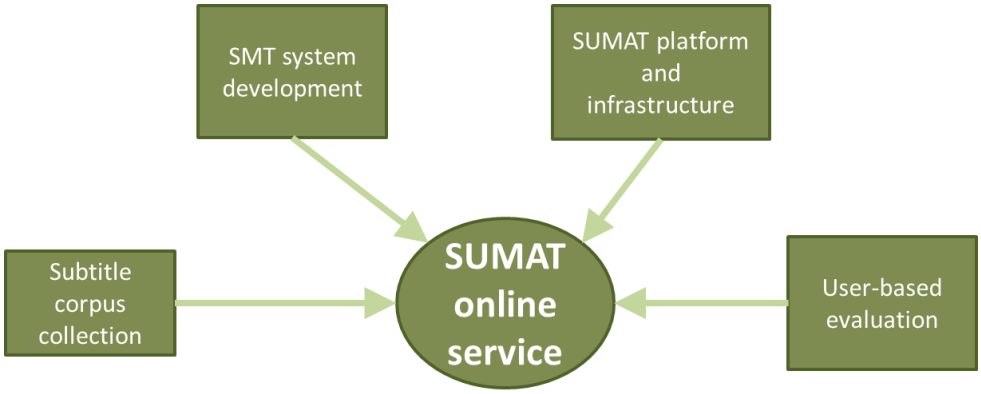
That is the SUMAT goal, to increase the efficiency and productivity of the European subtitle industry, while enhancing the quality of its results, thanks to the effective introduction of SMT technologies in the subtitle translation processes.

In order to achieve its objective, the SUMAT consortium is developing an online subtitle translation service addressing 9 different European languages combined into 14 different language pairs. The targeted language pairs are: English-Dutch; English-French; English-German; English-Portuguese; English-Spanish; English-Swedish and Serbian-Slovenian. The translation service will be working in both directions.

The rest of this document describes the progress of the SUMAT project so far in more detail, together with the corresponding results and future plans.

2. Summary of activities

The SUMAT project is divided into the following four main activities and their supporting subtasks:



Subtitle corpus compilation. A key task within the project is to collect high-quality subtitle data from the professional subtitle translation companies of the consortium and to convert and pre-process it into a format suitable to train SMT engines.

Although experiments in the literature have reported that 700.000 parallel subtitles are enough to obtain good results for SMT of subtitles, better results are expected with higher amounts. For this reason, one of our goals within this activity has been to collect as much high-quality subtitle data as possible for each language pair targeted in the project.

The subtitle collection task has been extended in order for subtitling companies to provide additional data from unexplored archives and newly generated content. Both, parallel and monolingual subtitles are being collected. The first as the basis for SMT training and the second, to build larger target language models – an approach that has been shown to be beneficial in most instances and, in particular, for language pairs with smaller training sets. The subtitle corpus collection task will finish by the end of this year. The overall amount of high-quality SUMAT subtitle data estimated to be collected is shown in the following tables:

PARALLEL CORPORA	
English-Dutch	1.452.963
English-French	1.566.431
English-German	2.282.329
English-Portuguese	762.716
English-Spanish	987.818
English-Swedish	959.304
Serbian-Slovenian	215.097
Total	**8.226.658**

MONOLINGUAL CORPORA	
Dutch	2.609.869
English	1.891.677
French	1.060.885
German	1.958.171
Portuguese	1.547.372
Swedish	3.147.588
Total	**12.215.562**

The collected raw subtitles then need to go through a conversion step in which: (a) they are all converted into plain text; (b) their language is automatically identified; (c) and the parallel subtitle files are also automatically document aligned. Once converted, the subtitle files can be pre-processed for SMT purposes: tokenized, sentence split, normalized and, finally, aligned at both sentence and subtitle level.

Unfortunately, the conversion and pre-processing steps above can reveal some problems present in the raw subtitle data. The subtitle files delivered may be corrupt, there may be errors in the raw language tags assigned or subtitle file pairs tagged as parallel may not actually be translations of each other. Moreover, not all subtitles present in two parallel subtitle files may be aligned. As a result, we expect to lose around 12% and 15% of parallel subtitle data in the conversion and pre-processing steps respectively and, thus, a 27% overall. However, unaligned subtitles will still be exploited by adding them into the monolingual dataset and using them for target language model training.

In addition to the SUMAT corpora, one extra source of high-quality professional subtitle data will be exploited: EuroparlTV[1]. Experiments will also be done with publically available extra parallel subtitle corpora, such as TED and OpenSubtitles. Despite these two last corpora contain mainly subtitles translated by amateur subtitle translators, the amounts available per language pair are considerably large and, thus, their impact on translation quality is worth exploring.

SMT system development. This activity involves developing the best possible SMT systems for each of the 14 language combinations of the project. The better the systems we develop, the bigger productivity and efficiency gains we expect to achieve with their integration into the current subtitle translation processes.

The development of the SMT systems will be incremental and linked to SUMAT online service prototyping and end-user evaluation cycles. We have started by developing baseline SMT systems with the available amounts of parallel subtitle data per language pair. Experiments with linguistic annotations and features are ongoing, towards the development of advanced systems that will exploit linguistic information and extra data. The translation quality of the advanced systems will then be evaluated by subtitle translators of the consortium and their feedback will be used to build the final SUMAT SMT systems. More details on each development step are provided in the next subsections.

Baseline SMT systems

Baseline SMT systems have been built for the targeted 14 language pairs. The related tasks included (1) selecting and preparing training, development and test material; (2) training and building translation models; (3) building baseline language models and (4) producing and evaluating baseline SMT system translations.

A number of training/development/test sets from the assembled parallel data were selected for each language direction. The test sets will henceforth be used for the remainder of the project to evaluate future iterations of the MT systems against the baselines.

Our baseline SMT systems made use of the state-of-the-art open-source Moses [Koehn et al., 2007] SMT training scripts for translation and reordering model building. To build the language models (LMs) we used the state-of-the-art open-source IRSTLM toolkit [Federico & Cettolo, 2007]. Taking the trained translation, reordering and language models, the Moses phrase-based decoder was then used to calculate baseline scores, for the following four baseline models:

1. 500K subtitle pairs of data (except for Serbian-Slovenian, English-Portuguese and English-Swedish, where only just over 500K subtitle-pairs were available, so we just built one model using all the data here);

[1] SUMAT has signed an agreement with EuroparlTV to employ their subtitles within the project.

2. All subtitle-pairs of data;

3. 500K sentence-pairs of data (except for Serbian-Slovenian, English-Portuguese and English-Swedish, where only just over 500K sentence-pairs were available, so we just built one model using all the data here);

4. All sentence-pairs of data.

We trained systems on subtitles and sentences, and for the systems trained on sentences, we also performed a cross-evaluation where we tested the engines on subtitles. The models were optimised using the designated development sets (2000 subtitles/sentences) and we evaluated the performance of each of the baseline systems on the appropriate test sets (4000 subtitles/sentences).

Figure 1 provides a visual representation of the obtained evaluation results with respect to the BLEU score for all 14 language pairs.

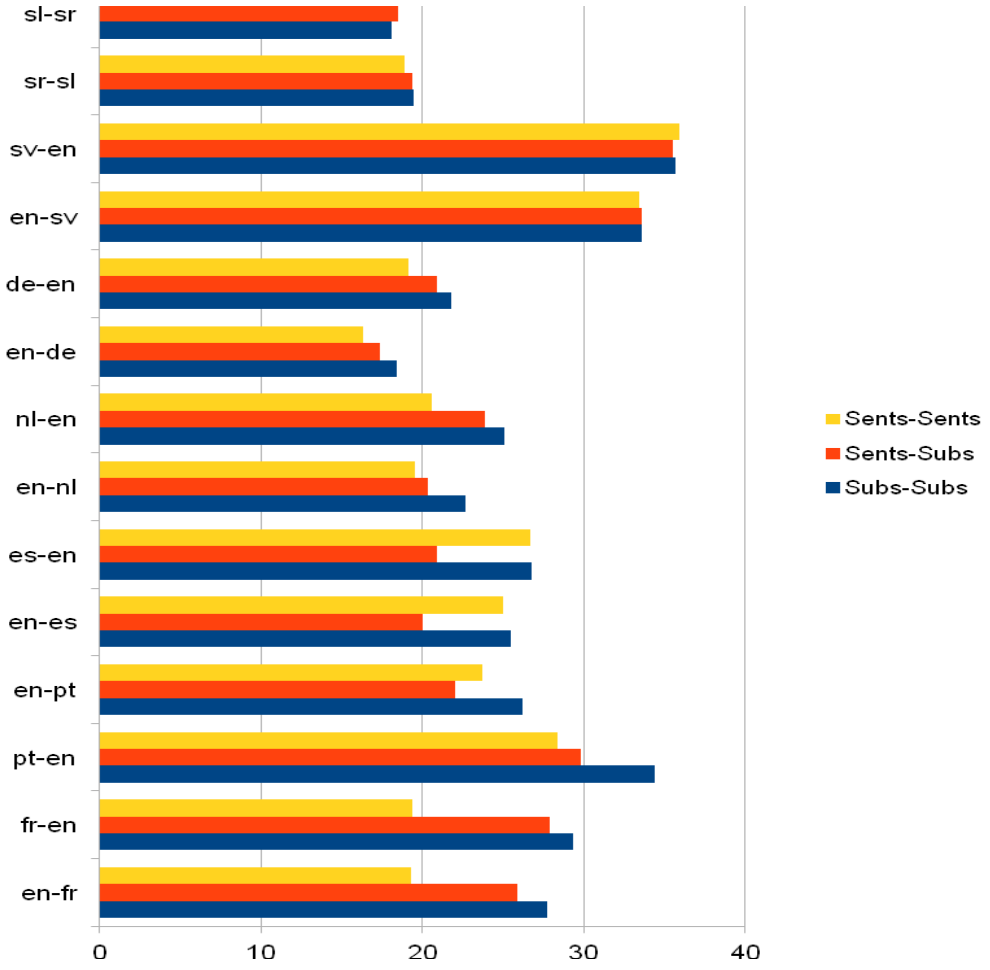


Figure 1: Overview of BLEU scores obtained on all language pairs

The scores obtained on the subtitles training and test sets are quite promising having obtained BLEU scores above 20 (except for Slovenian-Serbian, Serbian-Slovenian and English-German), that can be translated into reasonable quality for the majority of the SUMAT systems.

The main recommendation regarding future experiments is that as we aim to always test on subtitles, all future MT engines built in the project should also be trained on subtitles.

Experiments with linguistic annotations and features

This task is concerned with the exploration of the impact that linguistic annotations and features of several types, such as POS-tagging, lemmatization, dependency parsing, compound splitting, named entity recognition and phrase tables filling may have in the quality of subtitle translation.

Experiments have been distributed among partners, who are currently running them in parallel on selected language pairs. In each experiment, one Germanic and one Romance language is being tested. In addition, experiments with Slavic languages will also be run. Linguistic models for all language pairs will then be trained for those features that appear to be valuable in the distributed experiments.

Advanced SMT systems

The outcome of this task will be the set of translation models enriched with: (1) the linguistic information shown to be valuable in the distributed experiments described above; (2) all the collected, converted and pre-processed SUMAT parallel subtitle data; (3) all the monolingual SUMAT subtitle data used to build larger target language models; and (4) additional publically available in-domain (subtitle) and out-of-domain (non-subtitle) data.

We expect these advanced features to increase translation performance compared to the baselines and to positively impact the productivity gain of post-editing machine translations of subtitles.

Final SMT systems

The development of the final SMT systems for the targeted 14 language combinations will be done here. The final systems will implement the improvements suggested by the second end-user evaluation cycle in order to further optimize translation performance and the productivity of subtitle translation through machine translation.

SUMAT platform and infrastructure. The first versions of the envisaged SUMAT Online Service prototype and Demo have been deployed, following the architecture and functionalities defined at the beginning of the project[2] and integrating the developed baseline SMT systems. A preliminary version of the Demo (see Figure 2) was showcased at the SUMAT booth of the LREC EU Projects village.

[2] For more details, see the 2011 Annual Public Report

Within the project, two more versions of the SUMAT prototypes will be deployed integrating the advanced and final SMT systems respectively, and incorporating the improvements derived from the feedback provided by the user-based evaluations after each iteration.

Once ready, the final SUMAT Demo will be made publically available through the project website.

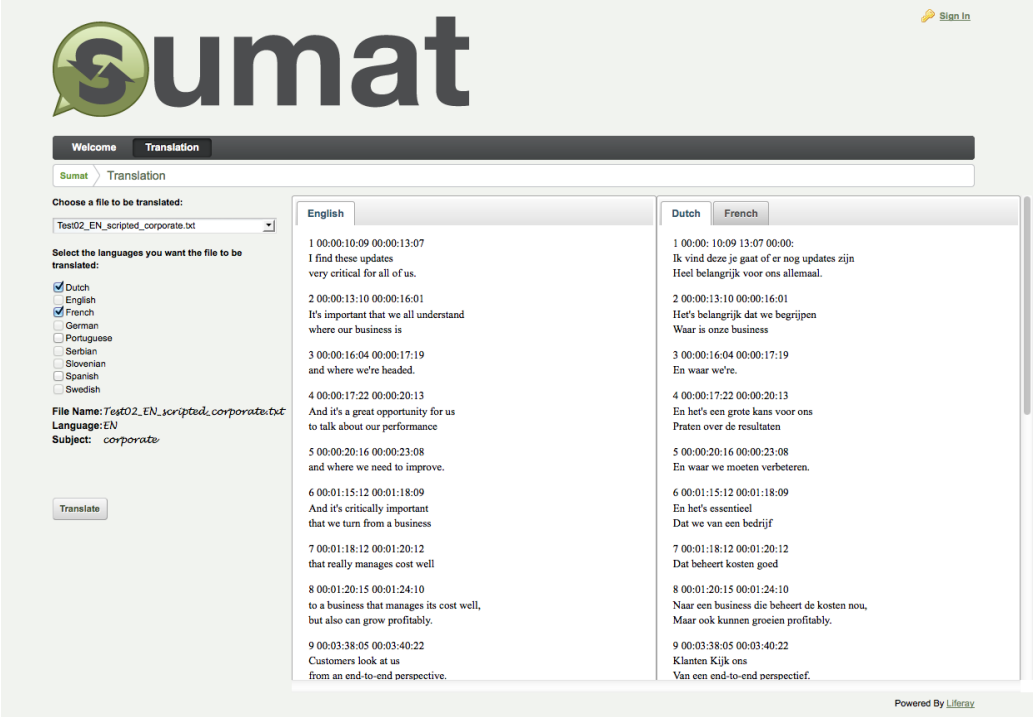


Figure 2. SUMAT Demo v1

User-based evaluation. It is important for the final users of the envisaged SUMAT Online Service to be involved in its evaluation throughout deployment, in order to provide feedback regarding its usability and translation performance. In this activity, three user-based evaluation cycles are planned:

(1) A small-scale translation performance evaluation case study with one language pair has already been carried out. In addition, end-users will evaluate the usability of the first prototype in terms of functionalities, persistence and expandability. Once finalized, the results of this evaluation cycle will include recommendations for the deployment of the second pilot service prototype, the development of the advanced SMT systems and the design of the next end-user evaluation cycle.

(2) Once the advanced SMT systems are ready and integrated in the second SUMAT pilot service prototype, end-users will evaluate its usability performance and translation quality and again provide feedback for their improvement.

(3) Finally, end-users will carry out the evaluation of the third pilot service prototype and final SMT systems deployed after the second evaluation cycle.

So far, a small-scale evaluation case study with the English-Swedish language pair has already been carried out. The translations produced by the current system for this language pair have been evaluated by subtitle translators from the consortium, following the strategy employed in previous similar evaluation experiments [Volk, 2008]. Results will be presented at the Languages and the Media conference to be held in Berlin, on 21-23 November 2012.

The first round of usability evaluation is ongoing.

3. Dissemination

The dissemination strategy of the project has started to take a more product oriented focus during the second year. A new plan is underway to coordinate the dissemination transitions from the initial simple project and objectives description towards the SUMAT product exploitation.

Since the final SUMAT Online Service will not be available until the last quarter of 2013, the current dissemination objective is to emphasize the results achieved so far and build up expectations within the potentially interested user community. The already available SUMAT LinkedIn Group and Twitter accounts will be used for this purpose, together with revamped versions of the existing website and leaflets. In addition, mock-up images and videos of the foreseen SUMAT product will also be produced and disseminated with the same objective.

Website and dissemination material

Updated versions of the website and existing leaflets will be produced for the Languages and The Media conference, to be held on 21-23 November in Berlin, where SUMAT will be:

- presenting the results of the small-scale end-user evaluation carried out for the current English-Swedish SMT system in a talk entitled *"What is the Productivity Gain in Machine Translation of Subtitles?"*
- participating in the closing panel of the conference, where the focus is likely to be on MT for subtitles
- showing the first version of the Demo in a booth at the conference Exhibition

Dissemination events

In parallel, the project partners have participated in the following dissemination events:

- **Multilingual Web, Luxembourg** (March 2012)
  SUMAT exhibited a poster at the event.
- **Language Resources Evaluation Conference, Istanbul** (May 2012)
  SUMAT exhibited at the event and a paper was presented at the conference.
- **European Association for Machine Translation, 16th annual conference** (May 2012)
  SUMAT representatives attended and presented two papers.
- **META-FORUM 2012, Brussels** (June 2012)
  SUMAT had an exhibition stand at the forum.
- **Formal Approaches to South Slavic and Balkan Languages, The Eighth International Conference (FASSBL-8)** (September 2012)
  SUMAT presented the work carried out for Slovenian and Serbian within the project so far in a talk entitled *"Towards Slovenian-Serbian Machine Translation of Subtitles"*
- **MIPCOM 2012, Cannes** (October 2012)
  SUMAT also had a stand at the show.
- **Eighth Language Technologies Conference, Slovenian Language Technologies Society, Ljubljana** (October 2012)
  SUMAT presented a paper with the work carried out for Serbian-Slovenian within the project.

Publications

V. Petukhova, R. Agerri, M. Fishel, S. Penkale, A. del Pozo, M. Sepesy Maucec, A. Way, P. Georgakopoulou and M. Volk, "*SUMAT: Data Collection and Parallel Corpus Compilation for Machine Translation of Subtitles*", Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), May 2012

M. Fishel, Y. Georgakopoulou, S. Penkale, V. Petukhova, M. Rojc, M. Volk, A. Way 2012, "*From Subtitles to Parallel Corpora*", proceedings of EAMT 2012, Trento, Italy

"*SUMAT: An online service for SUbtitling by MAchine Translation*", Proceedings of the 16th EAMT Conference, 28-30 May 2012, Trento, Italy

SEPESY MAUČEC, Mirjam, PRESKER, Marko, ZIMŠEK, Danilo, ROJC, Matej, VLAJ, Damjan, VERDONIK, Darinka, KAČIČ, Zdravko. Izdelava slovensko-srbskega vzporednega korpusa podnapisov za razvoj strojnega prevajanja v projektu SUMAT = "Building the parallel Slovene-Serbian corpus of subtitles for machine translation in the SUMAT project". V: ERJAVEC, Tomaž (ur.), ŽGANEC GROS, Jerneja (ur.). Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012, [Ljubljana, Slovenia] : zbornik 15. mednarodne multikonference Informacijska družba - IS 2012, zvezek C : proceedings of the 15th International Multiconference Information Society - IS 2012, volume C, (Informacijska družba). Ljubljana: Institut Jožef Stefan, 2012, str. 167-172, ilustr.

L. Bywood, M. Volk, M. Fishel, & Y. Georgakopoulou. *"Parallel Subtitle Corpora and their Applications in Machine Translation and Translatology",* Perspectives: Studies in Translatology. Special Issue: Corpus linguistics and AVT: in search of an integrated approach. (forthcoming)

4. Future work

The SUMAT future work will involve:

- finalizing the collection and pre-processing of the subtitle corpora;
- training advanced and final SMT systems;
- developing versions two and three of the online pilot service;
- and finally, running two more end-user usability and translation performance evaluation cycles of the SUMAT service.

These tasks will follow the more specific time plan shown in the following diagram:

1.	Compilation of final parallel corpora	December 2012
2.	Advanced SMT systems are developed	February 2013
3.	V2 of the online pilot service is deployed	March 2013
4.	End-users evaluate V2	June 2013
5.	Final SMT systems are developed	August 2013
6.	V3 of the online pilot service is deployed	September 2013
7.	End-users evaluate V3	December 2013
8.	The SUMAT Online Pilot Service is ready for use and testing	January 2014

5. Further Information

For further information please visit the SUMAT web site at www.sumat-project.eu for information on the project and its progress.

References

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst: Moses: open source toolkit for statistical machine translation. *ACL 2007: proceedings of demo and poster sessions,* Prague, Czech Republic, June 2007; pp. 177-180.

Marcello Federico & Mauro Cettolo: Efficient handling of n-gram language models for statistical machine translation. *ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation,* June 23, 2007, Prague, Czech Republic; pp. 88-95.

Volk, M. (2008) *The Automatic Translation of Film Subtitles. A Machine Translation Success Story?* In: Joakim Nivre, Mats Dahllöf and Beáta Megyesi (eds.): Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein. Uppsala