PRESEMT project: Publishable Summary

Project objectives

The objective of the PRESEMT project is to develop a flexible and adaptable MT system, based on a language-independent method, which is easily portable to new language pairs. This method attempts to overcome well-known problems of other MT approaches, e.g. bilingual corpora compilation or creation of new rules per language pair. PRESEMT will address the issue of effectively managing multilingual content and is expected to suggest a language-independent machine-learning-based methodology.

The key aspects of PRESEMT involve syntactic phrase-based modelling, pattern recognition techniques towards the development of a language-independent analysis and evolutionary algorithms for system optimisation. PRESEMT is intended to be of a hybrid nature, combining linguistic processing with the positive aspects of corpus-based approaches, such as SMT and EBMT. In order for PRESEMT to be easily amenable to new language pairs, relatively inexpensive, readily available language resources as well as bilingual lexica will be used. The translation context will be modelled on phrases, as they have been proven to improve the translation quality. Phrases will be produced via an automatic and language-independent process of morphological and syntactic analysis, removing the need for compatible NLP tools per language pair.

Parallelisation of the main translation processes will be investigated in order to reach a fast, high-quality translation system. Furthermore, the optimisation and personalisation of the system parameters via automated processes (such as genetic algorithms or swarm intelligence) will be studied. To allow for user adaptability, all the corpora used in PRESEMT will be retrieved from web-based sources. User feedback will be integrated through the use of appropriate interactive interfaces. PRESEMT is expected to be easily customisable to both new language pairs and specific sublanguages.

Work performed since the project start

The work performed during the third semester (M13-18) of PRESEMT relates mainly to the further development of the individual system modules (involving workpackages WP 3-6), mainly encompassing the pre-processing modules and the main translation modules, as well as the post-editing modules. Another important activity within the third semester has been the integration of the different modules into a unified platform (WP7), in anticipation of the release of the first PRESEMT prototype early in the fourth semester. Moreover, the consortium has been continuing to perform various dissemination activities (WP8), as detailed herewith. Furthermore, initial work on the evaluation infrastructure (related to WP9) has started, even though the relevant workpackage has been scheduled to start after the completion of the third semester of PRESEMT.

Main results achieved within the reporting period

The main project results that have been achieved within the first semester of the 2nd year can be summarised as follows:

- A substantial part of the monolingual corpora for all languages has been compiled and annotated. These corpora have been made available to the whole consortium.
- The Phrase aligner module (PAM) and the Phrasing model generator (PMG) have been integrated with the PRESEMT platform. Also, comparative experiments have been carried out to established methods in order to evaluate objectively the effectiveness of the newly-developed modules.
- Different corpus modelling methods have been investigated, with the aim of efficiently performing the word disambiguation task.
- The first translation phase (Structure selection) has been enriched with more features, and alternative approaches have been investigated. In addition, an initial version of the second translation phase (Translation Equivalent Selection) has been completed.
- A token generation component for English and German (i.e. the initial SL sentences) has been developed.
- The basic functionalities of the Post-processing module have been implemented.
- A substantial part of peripheral tools has been integrated into the main system platform.
- Several system modules have been integrated into the main system platform.
- The 1st annual review (corresponding to work carried out up to and including January 2011) has taken place with a successful outcome. The official results of the review were received and comments and suggestions have been adopted by the consortium during the 2nd year of the project.

Expected project results and potential impact

The project impact and results are expected to cover the research/scientific community as well as the general public. The MT approach being developed within the project has a number of innovative aspects, ranging from the method for implementing the division of sentences into phrases to the way of optimally utilising the linguistic information in the monolingual and bilingual resources. These aspects are project-specific, and thus can be expected to contribute as a whole to the state-of-the-art of the MT area. In this respect, the design of the prototype so that it allows the rapid development of new translation systems for different language pairs is of prime importance. The project results are envisaged to follow two main directions, (a) the dissemination of the MT system design as a whole and (b) the release of a prototype via the web for use by the academic community and the general public. The latter involves the release of specific modules individually, via dissemination activities, most likely following strategies defined in the exploitation plan. Hence, it is expected that a number of advances may be achieved, relating to the specific field of machine translation (entire system) as well as to other scientific areas, where mainly isolated PRESEMT modules are studied, the respective areas including general computational linguistics, pattern recognition and parallel processing, to name but a few.

Regarding the MT field, the impact is expected to be substantial. As noted by one of the PRESEMT peer reviewers, “the project is ambitious and its impact on future MT, both research and practice, could well be highly significant”. Also, as noted by another peer reviewer, the aspect of meta-evaluation of the system is important as “very useful information could come out”. The prime impact is of course expected to be to the machine translation community, which could filter through to the public both in terms of the final PRESEMT prototype as well as future MT systems, which may well be influenced by PRESEMT. It is hoped that advances in both the translation quality, the ease of development of systems for new language pairs and the speed of MT processing will all benefit from the project. Similar areas of impact are expected in the other research and application areas mentioned earlier, though these impacts are less evident at this point and will become more concrete as the project progresses further.

PRESEMT website

For further information and for keeping up-to-date regarding the PRESEMT project, please visit our website at **www.presemt.eu, which is being regularly updated.**

PRESEMT consortium & contact persons

Institute for Language and Speech Processing/R.C. "Athena"

Coordinator

http://www.ilsp.gr/

Contact person: **Dr. George Tambouratzis,** giorg_t@ilsp.gr

Gesellschaft zur Förderung der Angewandten Informationsforshung e.V.

http://www.iai-sb.de/iai/index.php/en/Die-GFAI.html

Contact person: **Dr. Paul Schmidt,** paul@iai.uni-sb.de

Norges Teknisk-Naturvitenskapelige Universitet

http://www.ntnu.no/

Contact person: **Prof. Björn Gambäck,** gamback@idi.ntnu.no

Institute of Communication and Computer Systems

http://www.iccs.gr/eng

Contact person: **Dr. Georgios Goumas,** goumas@cslab.ece.ntua.gr

Masaryk University

http://www.muni.cz/

Contact person: **Prof. Karel Pala,** pala@fi.muni.cz

Lexical Computing Ltd.

http://www.sketchengine.co.uk/

Contact person: **Dr. Adam Kilgarriff,** adam.kilgarriff@gmail.com