Integration of Machine Translation Paradigms

Results in Brief
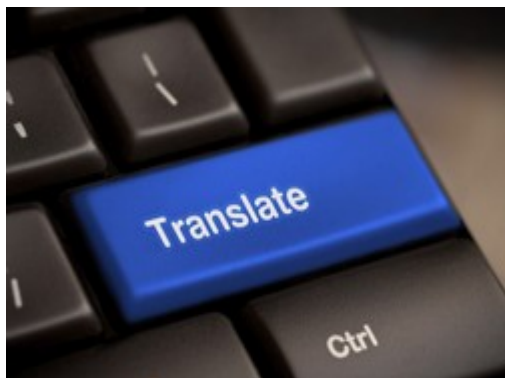
Open-source advanced linguistic system

An advanced tool that can efficiently translate language in real time would have an enormous payoff for society. EU-funded scientists proposed an advanced machine translation (MT) paradigm to further enhance the quality of translated texts.

DIGITAL ECONOMY

SOCIETY

© scyther5, Thinkstock

MT is a highly interdisciplinary and multidisciplinary field requiring input from professionals ranging from translators to engineers to computer scientists to mathematicians to linguists. The EU-funded IMTRAP ⤢ (Integration of machine translation paradigms) project worked on developing and validating an open-source hybrid MT system.

Researchers focused on multiple aspects of linguistics such as morphology, syntax and semantics. The resulting cutting-edge hybrid system prototype combines different MT paradigms, including statistical and rule-based MT (RBMT), and can be trainable in any pair of languages.

Researchers successfully introduced baseline statistical MT (SMT) systems for Chinese-to-Spanish and English-to-Spanish through a collection of corpora for these pairs of languages.

Another important IMTRAP achievement was development of the first Chinese-to-Spanish open-source hybrid system. The input of this system was pre-processed with an RBMT system and its output was passed to an SMT system. SMT uses models whose parameters stem from the analysis of monolingual and bilingual corpora. RBMT was used to define the structural transfer rules for phrases, and SMT was considered as the only source for the lexical transfer of words. By using SMT techniques, notable enhancements were observed in the final output of translation.

Furthermore, the output of this new hybrid system was contrasted with a state-of-the-art SMT system in the out-of-domain test set. Results showed that the new RBMT system outperforms the SMT system in all linguistic levels except the syntax level. Specifically, the new hybrid system far outperformed the state of the art in terms of lexical coverage.

In addition, IMTRAP achieved a higher level of hybridisation in statistical and RBMT. Work also focused on extracting transfer rules, assigning a probability to a sequence of n words, and introducing a language model to the generation step.

The results of the research, which succeeded in achieving its aim, were published in journal papers and books, as well as through international conferences. Commercialisation of a cost-effective hybrid MT system will have wide-ranging applications in information access systems and document translations. Society at large stands to benefit enormously, as will the European civil service and international relationships, not least with Asian parties since the project focused initially on the Chinese language.
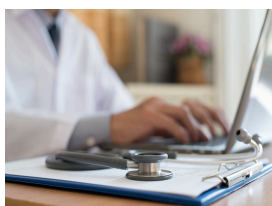
Keywords

Linguistic   machine translation   IMTRAP   languages   statistical MT

Discover other articles in the same domain of application

Real-time cloud-based monitoring at home will make hearts sing

31 October 2019

Closing the digital gap between poetry and technology

23 April 2021

News monitoring platform makes the job of media professionals easier and more efficient than ever

21 June 2019

Using data to understand – and predict – fashion trends

2 June 2020

Project Information

IMTRAP
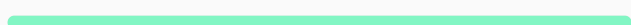
Grant agreement ID: 299251

Project website ↗

Project closed

Start date
7 December 2012

End date
28 October 2016

Funded under
Specific programme "People" implementing the Seventh Framework Programme of the European Community for research, technological development and demonstration activities (2007 to 2013)

Total cost
€ 173 212,80

EU contribution
€ 173 212,80

Last update: 8 May 2017

Permalink: https://cordis.europa.eu/article/id/147448-opensource-advanced-linguistic-system