Machine Translation

Overview

This result is about the development of four Machine Translation (MT) systems (for translation directions German-to-English, Dutch-to-English, English-to-Italian and German-to-Italian) targeting the generic news domain. Automatic performance measures and field trials showed that the quality of the generated translations is far away to be optimal, however it allows a rough understanding of television news program in a foreign language.

In depth description

Each Machine Translation system is build upon the open-source MT toolkit Moses licensed under the LGPL (http://www.gnu.org/licenses/lgpl.html). The decoder implements a statistical log-linear model including a phrase-based translation model, a language model, a distortion model and word and phrase penalties.
The phrase-based translation model provides direct and inverted frequency-based and lexical-based probabilities for each phrase pair included in a given phrase table. Phrase pairs are extracted from symmetrized word alignments generated by GIZA++. The n-gram language models (typically of order 5) smoothed through the improved Kneser-Ney technique are estimated on monolingual texts via the IRSTLM toolkit (http://sourceforge.net/projects/irstlm). The 14 weights of the log-linear interpolation models are optimized on the development sets by means of the standard MERT (minimum error training) procedure provided within the Moses toolkit.
As in all statistical system, performance strictly depends on the amount, quality and domain-pertinence of the training data. During the project translation quality of the developed systems – measured with the standard BLEU and TER metrics -- has been improved basically by increasing the size and pertinence of both the training and development sets and by adopting language-depending pre-processing techniques.

Potential fields of application

Machine translation can be successfully utilized in all the application areas that require the management of multi-lingual content. Perfect quality of the translated texts is not necessary for tasks such as content retrieval. Even contexts with linguistic limitations such as video captioning can benefit from the use of MT systems to increase the understanding of the video content when the original language is not known to users. Machine translation technology has a great potential of application when targeting specific domains for which sufficient amount of training data is available: in this case an appropriate tuning of the MT models is feasible, example of such domains are computer assisted translation of manuals of use and of weather and avalanche bulletins.

Possibilities for exploitation

Exploitation of the result is on a bilateral licensing agreement basis.

Further Information

Further technical information is available in TOSCA-MP Deliverables D2.1, D2.2 and D2.3 “Automatic Metadata Extraction and Enrichment”.

Contact Person

Diego Giuliani
Fondazione Bruno Kessler
Human Language Technology research unit
Via Sommarive 18, I-38123
Trento – ITALY
giuliani@fbk.eu