PANACEA: Platform for Automatic, Normalised Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies

Núria Bel[*], Marc Poch[*] and Antonio Toral[+]
nuria.bel@upf.edu, marc.pochriera@upf.edu, atoral@computing.dcu.ie
***Universitat Pompeu Fabra and [+]Dublin City University**

http://www.panacea-lr.eu

Description

A strategic challenge for Europe in today's globalised economy is to overcome language barriers through technological means. In particular, Machine Translation (MT) systems are expected to have a significant impact on the management of multilingualism in Europe, making it possible to translate the huge quantity of textual data produced, and thus, covering the needs of hundreds of millions of citizens. PANACEA addressed a critical thread to this vision: the so-called, language-resource bottleneck. Although MT technologies may consist of language independent engines, they highly depend on the availability of language-dependent knowledge for their real-life implementation, i.e., they require Language Resources (LRs). In order to equip MT for every pair of European languages, for every domain, and for every text genre, appropriate LRs covering every language, domain and genre must be produced. Moreover, a Language Resource for a given language can never be considered complete or final. Languages change and new knowledge domains emerge at rapid pace. Traditionally, LRs production is done by hand, and its high cost (highly skilled human work and development time) hinders full coverage. A company willing to cover the enlarged Union market needs to produce and maintain 500 bilingual glossaries, for instance.

The PANACEA project has focused on the development of a factory of LRs that automates the stages involved in the acquisition, production, updating and maintenance of LRs required by MT systems, and by other based on Language Technologies (LT) applications. This automation is meant to cut down costs significantly, in terms of time and human effort. Such reductions are the only way to guarantee a continuous supply of LRs that MT and other Language Technologies demand in a multilingual Europe. In order to address this objective, PANACEA has worked in (i) the development of a platform, designed as a dedicated factory for the composition of a number of LR production lines based on combinations of different web services and (ii) the integration of advanced components for the acquisition and normalization of corpora, monolingual and parallel corpora, their alignment; the derivation of bilingual dictionaries out of aligned corpora; and the production of monolingual rich information lexica using corpus based automatic methods.

The PANACEA factory has been thoroughly evaluated within R&D and industrial settings. The platform and the LRs production lines based on advanced technological components have proved the feasibility of the concept. PANACEA's contribution and potential impact has been demonstrated in an industrial evaluation carried out with the adaptation of Machine Translation products to a specific/specialized domain. In terms of effort, to produce a domain-adapted bilingual glossary of 1000 entries with PANACEA reduces costs from 30 person/hours to 0.5 person/hours. In terms of quality, there were no significant negative effects in the translation quality of the systems using automatically produced resources. A human evaluation showed that PANACEA domain-tuned SMT gained in quality up to a 6% with respect to the not tuned baseline, and that quality was not significantly worse than the achieved by other state-of-the-art systems as Google Translate.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 435.