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

Plataforma para la adquisición automática y la anotación normalizada eficiente de Recursos Lingüísticos para las tecnologías del Lenguaje Humano. PANACEA

Núria Bel
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona, España
Telf. 0034 935422307
nuria.bel@upf.edu

Resumen: El objetivo de panacea es engranar diferentes herramientas avanzadas para construir una fábrica de Recursos Lingüísticos (RL), una línea de producción que automatice los pasos implicados en la adquisición, producción, actualización y mantenimiento de los RL que la Traducción Automática, y otras tecnologías lingüísticas, necesitan.
Palabras clave: Tecnologías y Recursos Lingüísticos, Adquisición Automática.

Abstract: PANACEA's objective is to join a number of advanced interoperable tools to build a factory of Language Resources (LR). A production line that automates the stages involved in the acquisition, production, updating and maintenance of the LR required by Machine Translation and other Language Technologies.
Keywords: Language Resources and Technologies, Automatic Acquisition.

1 Motivation

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 (written or oral) data produced, and thus, covering the needs of hundreds of millions of citizens. PANACEA is addressing the most critical aspect for MT: the so-called language-resource bottleneck. Although MT technologies may consist of language independent engines, they depend on the availability of language-dependent knowledge for their real-life implementation, i.e., they require Language Resources. In order to supply MT for every pair of European languages, for every domain, and for every text genre, appropriate language resources covering all these aspects must be found, processed and supplied to MT developers. These should be provided in the format and with the information demanded by their systems. At present, this is mostly done by hand. Moreover, a Language Resource for a given language can never be considered complete nor final because of the characteristics of natural language: language changes and the emergence of new knowledge domains and new language varieties. What is needed is an automatic system for compiling, producing and validating language resources, a system conceived as integrated machinery for the production of LRs.

2 Objectives

The objective of PANACEA is to build a factory of Language Resources that automates the stages involved in the acquisition, production, updating and maintenance of language resources required by MT systems, and by other applications based on Language Technologies, and in the time required. This automation will cut down the cost, time and human effort sig-

nificantly. These reductions of costs and time are the only way to guarantee the continuous supply of the Language Resources that Machine Translation and other Language Technologies will be demanding in the multilingual Europe.

In order to address this objective, PANACEA will work in the following areas: 1) the creation of a platform, which will be designed as a dedicated workflow manager, for the composition of a number of processes for LR production based on combinations of different web services. 2) the automatic production of massive amounts of LRs for MT and other Language Technologies by the use of advanced components for the acquisition and normalization of corpus, monolingual and parallel corpora, the alignment of parallel corpora; the derivation of bilingual dictionaries out of subsentential aligned corpora; and the production of monolingual rich information lexica using corpus based automatic methods. 3) The evaluation of the platform and the LR production chain within the framework of both R&D and industrial settings.

3 Technologies and Components

The PANACEA platform will incorporate different technology components that will make possible a step-by-step automation of the whole process of producing LRs. Research to be undertaken during the project will address the optimization of these components for increasing accuracy and promoting precision results. The ultimate goal is to offer results that can be used in industrial processes with a high confidence.

Components to be integrated are well-known GNU and GNL licensed programs, such as FreeLing (Atserias et al. 2006) and Bitextor (Esplà 2009,) and other developed by the PANACEA partners: Subcategorization frame (SCF) acquisition system, which can be used to acquire comprehensive lexicons for verbs, nouns and adjectives from un-annotated corpus data (Preiss et al., 2007). Selectional Preference acquisition inferring semantic classes directly from corpus data (Korhonen et al., 2008); lexico-semantic class classification of nouns and adjectives (Bel et al. 2007). Subtree aligner (Zhechev & Way, 2008). PANACEA's contribution & impact will be demonstrated with a significant time and cost reduction in producing LR's. A real life use case will be used to measure the achievements

4 Project details

The PANACEA project is funded by the DG INFSO of the European Commission through the Seventh Framework Programme, Grant agreement no.: 7FP-ITC-248064.

The consortium PANACEA consists of seven partners leaded by the following researchers and engineers: Núria Bel, Universitat Pompeu Fabra, Barcelona, who coordinates the project; Nicoletta Calzolari, CNR-ILC, Pisa; Stelios Piperidis, ILSP - Athena Research, Athens; Gregor Thurmair, Linguatec, Munich, Anna Korhonen, University of Cambridge, Andy Way, Dublin City University and Khalid Choukri, EDA.The project has started 1st January 2010 and will finish in December 2012.

5 References

Jordi Atserias and Bernardino Casas and Elisabet Comelles and Meritxell González and Lluís Padró and Muntsa Padró. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library *Proceedings of LREC 2006*, ELRA. Genoa, Italy. May, 2006.

Núria Bel; Espeja, S.; Marimon, M. (2007). Automatic Acquisition of Grammatical Types for Nouns. In HLT 2007: The Conference of the North American Chapter of the ACL. Rochester, New York.

Miquel Esplà. (2009). "Bitextor, un cosechador automático de memorias de traducción a partir de sitios web multilingües". Procesamiento del lenguaje natural. N. 43.

Anna Korhonen, Yuval Krymolowski and Nigel Collier. 2008. The Choice of Features for Classification of Verbs in Biomedical Texts. To Appear in Proceedings of Coling 2008. Manchester, UK.

Judita Preiss, Ted Briscoe and Anna Korhonen. 2007. A System for Large-scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic.

Ventsislav Zhechev and Andy Way. 2008. Automatic Generation of Parallel Treebanks. In Proceedings of the CoLing '08, pp. 1105–1112. Manchester, UK.