Lexicon+TX: Rapid Construction of a Multilingual Lexicon with Under-Resourced Languages

Lian Tze Lim · **Lay-Ki Soon** · **Tek Yong Lim** ·
Enya Kong Tang · **Bali Ranaivo-Malançon**

Abstract Most efforts at automatically creating multilingual lexicons require input lexical resources with rich content (e.g. semantic networks, domain codes, semantic categories) or large corpora. Such material is often unavailable and difficult to construct for under-resourced languages. In some cases, particularly for some ethnic languages, even unannotated corpora are still in the process of collection. We show how multilingual lexicons with under-resourced languages can be constructed using simple bilingual translation lists, which are more readily available. The prototype multilingual lexicon developed comprise six member languages: English, Malay, Chinese, French, Thai and Iban, the last of which is an under-resourced language in Borneo. Quick evaluations showed that 91.2 % of 500 random multilingual entries in the generated lexicon require minimal or no human correction.

L. T. Lim
School of Engineering, Science and Technology, KDU College Penang, 32 Jalan Anson, 10400 Georgetown, Penang, Malaysia
Tel.: +604-226-6368, Fax: +604-228-0362
E-mail: liantze@gmail.com

L. T. Lim · L.-K. Soon · T. Y. Lim
Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
E-mail: liantze@gmail.com, {lksoon,tylim}@mmu.edu.my

E. K. Tang
Linton University College, Persiaran UTL, Bandar Universiti Teknologi Legenda, Batu 12, 71700 Mantin, Negeri Sembilan, Malaysia
E-mail: enyakong1@gmail.com

B. Ranaivo-Malançon
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
E-mail: mbranaivo@fit.unimas.my