CLARA - Common Language Resources and their Applications — a Marie Curie ITN
Early Stage Researcher (ESR) at SIA TILDE

Septina Dian Larasati
Duration: August 1st, 2011 - April 30th, 2013

I was a CLARA Early Stage Researcher (ESR) appointed for Project 6c: "Translation tools and resources for under-resourced languages". The research was conducted at SIA Tilde in Riga, Latvia. This also includes a secondment at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic. At the same time, I was also enrolled in a doctoral program at Charles University in Prague. I was starting my second year of my PhD when I started becoming an ESR.

The project was a research in translation tools and resources for under-resourced languages. I explored and investigated the possibilities to facilitate translation tools (mainly SMT systems) and resources development for languages that currently have limited translation technologies and resources. The language study case for this project was Indonesian, which was chosen because it was aligned together with my current PhD topic. And the application in Statistical Machine Translation (SMT) was by creating a translation system using MOSES[1].

The research includes creating language resources for the input of the SMT system. I composed a bilingual corpus (Indonesian-English) which then known as IDENTIC[2](1). One of the SMT experiments done was to improve translation quality by exploiting Indonesian morphological feature and analysis(2,5) produced by an Indonesian morphological tool, MorphInd[3]. I also did an SMT experiment to fix word alignment by taking account word similarity(3). During my stay as CLARA, I also collaborated with another CLARA ESR working in the work package 7: "Parsing Technologies and Grammar Models" by providing him a manually annotated Indonesian Dependency Treebank(4). Regarding research activities with people outside the CLARA network, I had the chance to be one of the Google Summer of Code 2012 (GSoC 2012)[4] mentor on a development of translation between two under-resourced languages, Indonesian and Malay, with a Rule-based Machine Translation approach(6).

I documented all the work as publications at international conferences and all the language resources made during the CLARA ESR stay are available for research purposes.

Riga, April 30th 2013
Septina Dian Larasati

References:

[1] www.statmt.org/moses/
[2] ufal.mff.cuni.cz/~larasati/identic
[3] ufal.mff.cuni.cz/~larasati/morphind
[4] www.google-melange.com/

Publications:

(1) Larasati, S.D.
IDENTIC Corpus: Morphologically Enriched Indonesian – English Parallel Corpus (Poster Session)
Proceedings of the 8th international conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 2012

(2) Larasati, S.D.
Towards an Indonesian-English SMT System: A Case Study of an Under-Studied and Under-Resourced Language, Indonesian
Proceedings of the 20th Week of Doctoral Studies (WDS), Prague, Czech Republic, 2012

(3) Larasati, S.D.
Improving Word Alignment by Exploiting Adapted Word Similarity (Poster Session)
Proceedings of the Workshop on Monolingual Machine Translation (MONOMT) at AMTA 2012, San Diego, USA, 2012

(4) Green, N., Larasati, S.D., & Žabokrtský, Z.
Indonesian Dependency Treebank: Annotation and Parsing
Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC), Bali, Indonesia, 2012

(5) Larasati, S.D.
Handling Indonesian Clitics: A Dataset Comparison for an Indonesian-English Statistical Machine Translation System (Poster Session)
Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC), Bali, Indonesia, 2012

(6) Susanto, R.H., Larasati, S.D., & Tyers, F.M.
Rule-based Machine Translation between Indonesian and Malaysian (Poster Session)
Proceedings of the Workshop on South and Southeast Asian Natural Language Processing (WSSANLP) at Coling 2012, Mumbai, India, 2012