

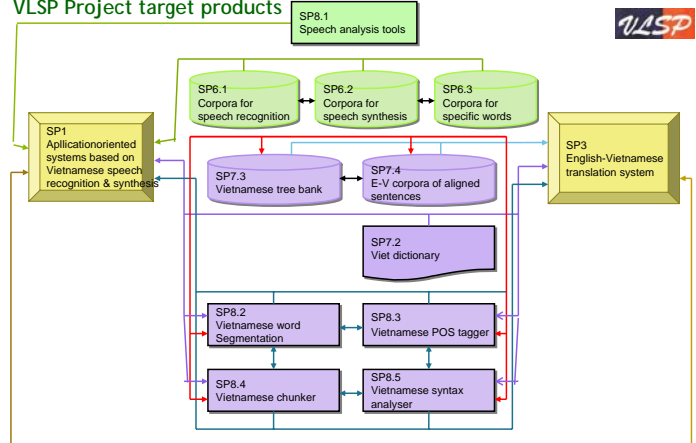


Vietnam Country Report

Activities on O-COCOSDA 2010

Luong Chi Mai
Institute of Information Technology
Vietnam Academy of Science and Technology

VLSP Project target products



Finish National VLSP Project of the years 2007-2009 with the target results



NLP Resources & Tools

- Viet Treebank (parsed corpus - a text corpus in which each sentence has been annotated)
 - 10,000 trees, 1,000,000 morphemes
 - Develop Treebank editor
- Vietnamese Machine Readable Dictionaries (MRD):
 - 35,000 Vietnamese common used words in modern Vietnamese
 - Develop a tool for building VCL with XML representation
- NLP tools (based on the same view of words, label assignment, sentences, Viet dictionary and Viet Treebank, Using statistical and machine learning methods in building such tools)
 - word segmentation (n-gram, dictionary, 98%)
 - POS tagging (MEMs, CRFs, 90%)
 - Chunking (CRF, online learning, 94%)
 - syntax analysis (LPCFG, Bikel's implementation, >70%)

3

Website: sharing NLP tools for a community

<http://vlsp.vietlp.org:8080/demo/>



New approved National Project for Speech-to-Speech Translation, 2011-2012



- Develop, improvement of the engines
 - ASR
 - Noise removal
 - Tonal integration
 - MDSL for HMM based
 - Hybrid SMT:
 - incorporate linguistic features to a translation model
 - Investigation how to combine a rule based to SMT model
 - TTS:
 - HMM based TTS for the Northern Dialect
 - Concatenation based TTS for Southern Dialect
- Develop speech and text corpora for a translation in daily communication domain for tourism
- Develop S2S system for a smartphone under Windows Mobile operating system
- Develop a Portal to share speech, NLP tools and resources to a community

5

U-STAR Speech-to-Speech Translation



- Vietnam is a member of the new U-STAR consortium
 - Signed MoU in August 2010
- Completion
 - Translation 160K BTEC text corpus from English to Vietnamese for SMT purpose
 - Economical Law Parallel corpus
 - Law Corpus: 100K sentences.
 - Collect bilingual law documents from official sources.
 - Sentence-alignment over the collected.
 - Refine manually the alignment result.

6