

NTU, NTT Site Reports

Francis Bond, Petter Haugereid

Division of Linguistics and Multilingual Studies

Nanyang Technological University

<bond@ieee.org>

2009-07-20



- NTU
 - Machine Translation (Petter)
 - WordNets (Japanese, English, Chinese, Malay)
 - NTU Multilingual corpus
 - Classifiers
 - Teaching HPSG, Grammar Engineering

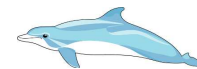
- NTT Report (Sanae Fujita)
 - Release of GoiTaikei — A Japanese Lexicon (NC)
 - Joint work with NTU on corpus annotation and WSD



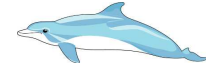
- Japanese WordNet
 - Added definitions and examples
 - Extended POS tags (subclasses of V, A)
 - Improved variant handling

- Chinese WordNet
 - Extending the SouthEastern University WordNet

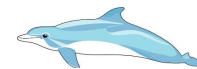
- English WordNet
 - Adding new entries, merging old



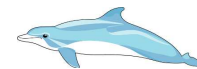
- Biggish EU project <www.kyoto-project.eu/>
 - Parse environmental documents (WWF) extract facts
 - PDF → text → dependencies, NE, WordNet+WSD, . . .
 - Extract new terms, add to domain ontology/WordNet
 - Full logical ontology lookup (with inferencing)
 - Wildly ambitious, kind of fun
- WordNet annotation tools, NER, disambiguation
 - Built Japanese, English, Chinese WSD systems (Dutch, Spanish, Italian) — the devil is in the details of the matching



- Combined Malay/Indonesian WordNet
 - default is both, can tag words as Malay or Indonesian
- 30,000 or so entries by the end of summer
- Collaboration with NTT (using the KAMI lexicon)
- Collaboration with Hammam Riza. BPPT, Indonesia



- Small, deeply analysed corpus
 - 6,000 sentences x 3 languages (cmn, eng, jpn)
 - * Mainichi Newspaper (NICT translations)
 - * Sherlock Holmes
 - * Cathedral and the Bazaar (plus +++)
 - * Singapore Tourist data (plus Korean, Viet, Indo)
 - Hand alignment, WordNet tagging, Treebanking
- Plus lot more Japanese-English (and some Chinese)



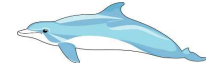
- Built yet another sense annotation system
 - allows discontinuous MWEs
 - annotate word by word (type-based)
- Annotate NT Corpus
 - Students annotating the multilingual corpus (En, Zh)
 - NTT-AT annotating the Japanese
 - Feedback to the WordNets (En, Ja, Zh)
 - Finish $3 \times 6,000$ in July
- Annotate Kyoto Corpus (20,000 sentences) with NTT
 - Semi-automatically (WN constrained by GT)



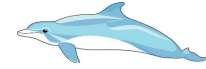
- Use one language to disambiguate the other
 - Lexical Semantics (WordNet)
 - Structural Semantics (DELPH-IN grammars)
 - On the same corpora (NTU multi and Tanaka, NICT Corpora)

- Large amounts of Ja-En, some Ja-En-Zh, a bit of Ja-++++

- We have Japanese translations of the WSJ



- Tagging Chinese and Japanese for classifiers and their targets
 - Induce classifier semantic restrictions
 - New gold standard for parsing/generation of classifiers
- Pyscho-linguistic experiments on count/mass to come



- Erasmus Exchanges
 - Lea, Dominikus (from Saarlandes)
 - Liling (to Saarlandes)
 - Mathieu Morey (from Aix)

- **2011 AAMT Nagao Award** for individuals or groups who have contributed to R&D leading to or facilitating commercialization of MT systems.

- **AAMT Standardization Working Group** for work on a Common User Dictionary Format (UTX) (山本ゆうじ, 伊藤肇, 村田稔樹, Francis Bond, 島津美和子, 大倉清司, 加藤マイケル 孝仁)