NTU Site Introduction
NTT, NICT Site Reports

Francis Bond †,‡

† **Division of Linguistics and Multilingual Studies**

Nanyang Technological University

‡ **NICT Language Infrastructure Group**

National Institute for Information and Communications Technology

<bond@ieee.org>

2009-07-20

Overview

- NTT Report (Sanae Fujita)
  work on WSD and tagging photos (not very open)
- NICT Report (Takyuki Kuribayashi, Francis Bond)
  work on WordNet, some treebanking, SMT with rewriting
- NTU multilingual corpus
  - Chinese/English/Japanese WordNet (with NICT)
  - Jacy Development, JaEn development
- SMT with rewritten data (with NICT)

SMT with rewritten data

- Try to balance the training data for some phenomena
- Starting with negation and questions
  Common errors for SMT

For every English sentence (e_i) in the corpus
  parse, (negate, question) and generate
  For each Japanese sentence
   parse, (negate, question) and generate
   — pull out new aligned pairs

Example of rewriting

(1) Dogs bark. ↔ 犬が吠える。
(2) Dogs don't bark. ↔ 犬が吠えない。 (NEG)
(3) Do dogs bark? ↔ 犬が吠えるか。 (QUE)
(4) Do dogs not bark? ↔ 犬が吠えないか。 (NEG, QUE)

- Jacy generation cover still less then ERG
- generation slower than parsing
- Now testing with SMT

NTU multilingual corpus

- Small, deeply analysed corpus
  - 2,000 sentences x 3 languages (cmn, eng, jpn)
    - Mainichi Newspaper (NICT translations)
    - Sherlock Holmes
    - Cathedral and the Bazaar (plus)
  - Hand alignment, WordNet tagging, treebanking
- Plus lot more Japanese-English (and some Chinese)
- Also Chinese, English, Malay, Tamil.

Cross Lingual Disambiguation

- 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)
- Results to come.

Good News

- New PostDoc: Petter Haugerid (from NTNU)
  working on cross-lingual disambiguation (alignment, transfer)
- New PhD: ???
  working on cross-lingual disambiguation (syntax)
- JSPS-Singapore bilateral funding (NTU-NICT):
  Revealing Meaning Using Multiple Languages

Future Work

- New languages
  Chinese (Mandarin), Malay, Tamil, Singlish
  WordNets and hopefully grammars

- Combining Syntax and Semantics (revisited)

- More visibility

Visibility

- Two PhDs
  - Sanae Fujita. *Constructing, Refining and Exploiting Rich Linguistic Resources*. PhD thesis, Nara Institute of Science and Technology, 2009
  - Eric Nichols. *Applying Deep Grammars to Machine Translation, Paraphrasing and Ontology Construction*. PhD thesis, Nara Institute of Science and Technology, 2010