NICT Site Report

Francis Bond †,‡

<www3.ntu.edu.sg/home/fcbond/>

† **NICT Language Infrastructure Group**

National Institute for Information and Communications Technology

‡ **Division of Linguistics and Multilingual Studies**

Nanyang Technological University

<bond@ieee.org>

2009-07-20

Overview

- Jacy Development
- KRG Development
- JaEn development
- SMT with paraphrased data
- Japanese WordNet
- NICT multilingual corpus
- Infrastructure (Lextype DB; Egad)

Jacy Maintenance

- zero-pronouns changed from rules to types
  20% speed up
- Verb clusters changed from right headed to left headed.
- N-N compound (was NP-NP)
- Still meeds more linguist love
- Documentation (LexType DB: CH, TB)
- Book in progress

KRG

- Increasing robustness and cover
- Matrixification: pet version and generation
- All new morphology
- Increased lexicon
- Some phenomena temporarily unavailable
  classifiers, serial verbs, relative clauses, . . .
- Some new features: STYLE

Song Sanghoun (SSH)

JaEn (Noja, EnJa, EsJa) development

- Dictionary Construction from JMDict and Moses phrase table (EN)
- Documentation and various prototypes
- Rule acquisition from Corpora (EN)
- Still worse than MOSES (BLEU, Meteor, Human)
- Now working on feedback cleaning
  - Remove harmful rules

SMT with paraphrased data

For every English sentence (e_i) in the corpus

parse and generate For each Ja-En pair

create n translations consisting of

the original translation

the best $m < n$ distinct paraphrases (up to n)

$n - m$ copies of the original translation

- Still issues with unknown words
- Now doing EnEn MT ($\rightarrow$ Paraphrase)
- Several requests for standalone paraphraser

Example of paraphrasing

(1) この ことから、会社には事故の責任が無いことになる。

It follows from this that the company is not responsible for the accident.

It follows that the company isn't responsible for the accident from this.

It follows that the company is not responsible for the accident from this.

That the company isn't responsible for the accident follows from this.

(2) 打合せの記録

The minutes of the meeting/the meeting's minutes/the meeting minutes/the minutes of the meeting's

Infrastructure

- LexType DB: Lexical Type Database
  - Now in the LKB
  - Can be any type
  - Combined with simple corpus lookup
- Egad: Erroneous Generation Analysis and Detection

Japanese WordNet v0.9 released

- Japanese translation of En WordNet 3.0
  - v0.9 freely available: `nlpwww.nict.go.jp/wn-ja`
    - 49,190 Synsets
    - 85,966 Words
    - 156,684 Senses
    - Illustrations for 541 Synsets
- Still being extended
  - Revised Structure, External links
  - Sense Tagged Corpora
  - Japanese Definitions and Examples

Free is good

- $> 1,000$ downloads
- Interface in the following languages
  - Perl (FCB)
  - Python
  - Java
  - Ruby
  - Gau
- Cool graphical lookup
- Linking to EDICT

Kyoto Project

- 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
- Would like to make compatible with DELPH-IN

NICT multilingual corpus

- Aiming to have 10 million sentences of parallel text
  - 2-3 million Ja-Zh
  - Remainder Ja-En
  - Small amount of other languages
- Make as free as copyright allows us
  - Used for SMT, EBMT
  - MASTAR - tourism, manga, JC - scientific
- Cathedral and Bazaar test corpus (Language Grid)
  - En, Zh, Ja, Ko, Fr, Es, De, It, Pt

Wishlist

- Robust Parsing
  - Unknown words we can generate from
  - Restrict parses with suppertagger
  - Unified Pet/LKB preprocessor
- Shared Multilingual Corpus
- Training treebank from GOLD (reforesting)

- Harmonization
  - Universal Predicate names/features
  - Directory Structures
  - PET/LKB configuration
  - Lexicon tools
- Shared Multilingual Corpus
- WSD

New DELPH-IN site?

- I have moved to NTU
  - Division of Linguistics and Multilingual Studies
  - School of Humanities and Social Sciences
  - Nanyang Technological University
- NICT still active
  - Treebanking
  - Paraphrasing
  - WordNet and WSD