KRG (Korean Resource Grammar) Status Update

Jong-Bok Kim

Kyung Hee U.
jongbok@khu.ac.kr

DELPH-IN Summit
June 25, 2011
Suquamish, WA

Project: Developing an MT System through Deep Language Processing

- Project title: Developing an MT System through Deep Language Processing
- Project period: 2 years (May 2011 – April 2013)
- Funding: about 82,000 US dollars
- Funding Source: National Research Foundation of Korea

Project members

- Korea: Jong-Bok Kim (PM), Jaehyung Yang, and Yonghoon Lee,
- USA: Emily Bender and Sanghoun Song
- UK: Peter Sells
- Singapore: Francis Bond

Main Goals

- Develop a more comprehensive and wide-coverage KRG grammar (KRG 3.0)
- Construct MT baselines for KoJa/JaKo and KoEn/EnKo
- Develop foundations to build a Korean treebank with rich and dynamic syntactic and semantic information (next stage project)

First Year Plan (Present – June 2011)

- Try to refine the KRG and expand its coverage with deep processing
- Target sentences:
  1. SERI test set sentences (about 360 sentences)
  2. 850 sentences from the KPSG textbook (reflecting major linguistic phenomena)
  3. 1000 sentences extracted from the Sejong Electronic Dictionary (total app. 70,000 sentences)
  4. 1000 sentences (less-than 15 word) extracted from the Sejong bilingual corpora (Kor-Jpn (total app. 4,000 sentences), Kor-Eng (total app. 40,000 sentences))
- Build a toy MT system that can cover about 20% of the test set sentences

First Year Plan (cont.)

- build KRG 3.0 optimized for the target MT system
- focused linguistic phenomena (refining and adding):

1. wh-question
2. modification (nominal as well as verbal)
3. subordination
4. serial verb constructions
5. pro-drop
6. more ....

Second Year Plan (July 2011 – June 2012)

- Further develop the first-stage toy MT system into a more decent one
    1. build a small-size treebank
    2. build transfer rules for the MT system
- Refine KRG 3.0 with a wider coverage

Second Year Plan: Target sentences

- 2000 sentences extracted from the Sejong Electronic Dictionary
- 2000 sentences extracted from the Sejong Bilingual corpora (Kor-Jpn, Kor-Eng)
- possible addition: data from the Singapore Tourism Corpus & *Little Prince*

Evaluation and dissemination

- Online Demo: http://krg.khu.ac.kr
- Organizing workshops during 2012 HPSG, 2012 DELPH-IN Summit, or at Kyung Hee Univ in 2012 and 2013.