ARABIC TREEBANKING AT LDC: Update on Current GALE Efforts

Mohamed MAAMOURI

Linguistic Data Consortium

mamouri@ldc.upenn.edu

Arabic Treebank (ATB) Accomplishments

- Newly developed, comprehensive, and stable POS guidelines/documentation (now about 215 pages)
- Expanded documentation for TB guidelines (300 pages) (http://projects.ldc.upenn.edu/ArabicTreebank/)
- Completed training of three annotators (for POS & TB)
- Completed enhancement, automatic and partial manual revision of ATB1, ATB2 and ATB3 (total of 738,845 words combined)
- Completed POS annotation of ATB5 (over 100K words of Arabic Broadcast News, parallel to existing EATB corpus)

Improved ATB Training and IAA

- Planned (as of this quarter) on-the-job training sessions focusing on BN corpus specificities
- Expanded QC searches for ATB for better error detection
  - 76 total CorpusSeach scripts targeting annotation errors (16 only in 2006)
  - Results hand-corrected by annotators in two passes
- IAA improved to 94.3% (output of production workflow based on ATB3 – all newswire)
  - Up from a baseline of 85% established in Fall 2006 – IAA of current BN annotation will be released with the completed treebanked data
  - Compare to Chinese Treebank IAA 93.8%
  - Emphasis on improving inter-annotator agreement (IAA) based on 10% blind annotation during production

ATB Technical Improvements

- Improved software for ATB corpus analysis and QC
  - Automatic checking of all "function word" tokens for consistency with morphological guidelines
  - Automatic modification of words based also on consistency with tree annotation
- Retrained parser for better output in annotation pipeline
- Latest parsing results show important improved consistency & core syntactic relations recovery (cf. Seth Kulick's presentation later on in this meeting)
- Update and revision work on the LDC-BAMA Arabic Morphological Analyzer – new version soon to be released

ATB GALE Year 4 Planning

- Continue syntactic annotation of ATB5 (over 100K words of MSA Broadcast News (BN), parallel to existing EATB corpus)
- Release POS annotation of ATB5 by March 2009
- Continue POS and TB annotation of another 200-250K words of MSA Broadcast News
- Continue work on MSA tagger with significant results expected during this period
- Continue experimentation and improvement of parsing results and error analysis detection and correction