Arabic Treebank Annotation of Broadcast News Speech

- Coordination necessary for Broadcast News (BN) data at LDC across Arabic Treebanks (ATB)
- English translation of Arabic BN transcripts
- word-level alignment of Arabic and English data
- Corresponding English Treebank

- Robust enough to account for the new genre of data?
  - Newly improved ATB annotation pipeline
  - Revised ATB annotation guidelines

- Yes! (with some adaptation)

Issues of Broadcast News Data

- Metadata ➔ Not Annotate in Treebank
  - Metadata to convey several kinds of information in addition to the text of what each speaker is saying
  - Coughs, laughter, background noise or music, etc.
  - Speech in a language other than Arabic, or a colloquial dialect of Arabic rather than MSA

- Speech Effects ➔ Annotate in Treebank
  - Discourse markers, hesitation sounds, word fragments, mispronunciations and other disfluencies

- Indistinct Audio Signal ➔ Annotate if possible in Treebank
  -Difficult to guess at from context rather than from the audio signal
  - Cascading impact on higher-level annotations

Tool Development for ATB BN Data

- Tools adapted to filter out the metadata that ATB would ignore, while preserving the ability to align the annotation results to the initial transcripts
- Give POS annotators access to the original audio files when necessary to disambiguate doubtful words in the transcript
  - Transcribed “ضرب” /sˤrˤ/ to “prbrace” in place of “بتر” /bˤrˤ/ to “pierce”
  - Transcribed “أين” /ʔe’en/ to “there” in place of “أين” /ʔe’en/ to “there”

Guidelines Development for BN Data

- Syntax: adapted the Penn English Treebank Switchboard annotation guidelines for use with Arabic BN data
  - Treatment of speech effects, disfluencies and metadata is not language-specific
  - Arabic-specific dialect-related structures were addressed
  - Revised and enhanced Arabic Treebank Syntactic Guidelines for general syntax

- Morphology/POS: Dialect words given DIALECT tag
  - Dialect is low frequency in highly monitored BN speech

- ATB Annotation Pipeline

  - ATB annotation and processing pipeline improved overall

  - Adapted to support the production of treebanked broadcast news corpora

  - Several components devoted to handling word forms outside the vocabulary and grammatical repertoires of SAMA
  - Feedback to upgrade SAMA’s lexicon and morphotactic tables
  - Careful setting of POS labels and glosses assigned to novel/terms

- Pipeline stages, including consistency checking at every stage

  1. Speech Transcription and SU Annotation

  2. Morphological Analyzer and Morphological/Part-of-Speech Annotation

  3. Clitic Separation, Parsing, and Syntactic Annotation

  4. Quality Control Searches and Corrections

Arabic Treebank Annotation Pipeline Stages

Issues of Broadcast News Data

- Metadata ➔ Not Annotate in Treebank
- Speech Effects ➔ Annotate in Treebank
- Indistinct Audio Signal ➔ Annotate if possible in Treebank

Tool Development for ATB BN Data

- Tools adapted to filter out the metadata that ATB would ignore, while preserving the ability to align the annotation results to the initial transcripts
- Give POS annotators access to the original audio files when necessary to disambiguate doubtful words in the transcript

Guidelines Development for BN Data

- Syntax: adapted the Penn English Treebank Switchboard annotation guidelines for use with Arabic BN data

- Morphology/POS: Dialect words given DIALECT tag

- ATB Annotation Pipeline

- ATB annotation and processing pipeline improved overall

- Adapted to support the production of treebanked broadcast news corpora

- Several components devoted to handling word forms outside the vocabulary and grammatical repertoires of SAMA
  - Feedback to upgrade SAMA’s lexicon and morphotactic tables
  - Careful setting of POS labels and glosses assigned to novel/terms

3. Clitic Separation, Parsing, and Syntactic Annotation

- Clitic Separation
  - Clitics separated automatically according to the POS annotation
  - Segmentation necessary for treebanking phase
  - “Wrong” clitics in BN English can distinguish between a clitic and a noun
  - “Wrong” clitics in BN English can distinguish between a clitic and a noun

- Parsing
  - Bit parsing used to automatically create syntactic trees for treebanking
  - Input: “good” POS annotation, as split for the treebank tokens

- Syntactic Annotation
  - Treebank Annotation Tool
  - Treebank annotation
  - Correct the parse output when necessary
  - Add function tags not included by the parser (most adverbial tags)
  - Add empty categories with appropriate co-indexing

- Quality Control Searches and Corrections

- Quality Control Searches
  - Corpus Search tool
  - 35 error-search queries to locate known problems involving improper patterns of tree structures and node labels

- Corrections
  - Treebank annotation tool: displays search results and allows annotators to click-through directly to affected portion of each tree

Conclusions

- New challenges posed by Broadcast News data

- ATB’s improved pipeline and revised annotation guidelines robust enough to require few changes for BN data

- Similar adaptations planned in the future to account for additional new data genres (webtext and dialectal speech, etc.)

- Expect current pipeline will continue to prove flexible and robust enough to accommodate the morphological and syntactic annotation of the necessary data