Expanding Arabic Treebank to Speech: Results from Broadcast News
Mohamed Maamouri, Ann Bies, Seth Kulick
{maamouri,bies,skulick}@ldc.upenn.edu

Treebank of large corpus of relatively structured speech transcribed from various Arabic Broadcast News (BN) sources
- 432,976 source tokens/517,080 tree tokens, in 120 manually transcribed news broadcasts
- Annotation challenges & guidelines additions
- Challenges of parsing an Arabic speech corpus & initial parsing evaluation for BN data
- Technical challenge of maintaining correlation with SAMA
- Will be available to the community as an LDC publication (LDC2012T07)

Challenges of Spoken Language in Arabic Broadcast News Data

1. Cross-linguistic speech effects
- BN = mostly scripted → mostly Modern Standard Arabic (MSA)
- Lexical and syntactic structures are very similar to the MSA in written newswire data
- BUT, spoken = cross-linguistic speech effects are common
- Restarts, fillers, hesitations, repetitions, etc.
- ATB BN annotation guidelines developed based on English Penn Treebank Switchboard
- Annotate similar speech effects across languages in a similar way, while focusing on the transcribed BN corpus at hand

2. Impact of Arabic dialect issues
- Constructions specific to spoken language, to broadcast style (as opposed to written style), certain novel MSA usages
- Some dialect data
  - Variety of Arabic dialects in BN
  - From on-the-street interviews and similar informal contexts
- 4,760 (1%) of the 432,976 source tokens include a DIALECT part-of-speech (POS) tag.

3. Transcription issues
- Manual transcription, potential transcription errors
- Affect downstream annotation in Arabic-specific ways
- Partial/incomplete words given POS “PARTIAL”
- Errors in transcription given POS “TRANSERR”
- Similar to TYPo tag used for text corpora

2. Impact of Arabic dialect issues
- Constructions specific to spoken language, to broadcast style (as opposed to written style), certain novel MSA usages
- Some dialect data
  - Variety of Arabic dialects in BN
  - From on-the-street interviews and similar informal contexts
- 4,760 (1%) of the 432,976 source tokens include a DIALECT part-of-speech (POS) tag.

Dialect examples:
- Discourse filler: يتعطّب (he/it means) frequent in spoken Arabic
- Much like “you know” in English
- Annotated as PRN

(PSL (VP يتعطّب (NP-PROPR (NP-MSG (NP "You know") ]

He wants to become a lawyer

3. Transcription issues
- Manual transcription, potential transcription errors
- Affect downstream annotation in Arabic-specific ways
- Partial/incomplete words given POS “PARTIAL”
- Errors in transcription given POS “TRANSERR”
- Similar to TYPo tag used for text corpora

Transcription examples:
- Partial word:
  - Trailing hyphen indicating transcribed incomplete word → “PARTIAL” POS tag
  - Typically inside a tree node marked → UNF for unfinished, and included as SAMA Status #4

Transcription error:
- Token is left as transcribed
  - For consistency with other annotation work on the same transcribed corpus
  - POS = TRANSERR, but syntactic annotation = as if written correctly

Status of BN Corpus Integration with SAMA
- Status flag for each source token to make explicit the connection between morphological analysis from Standard Arabic Morphological Analyzer (SAMA) and ATB POS annotation

Status #4 EXCLUDED from check with SAMA
- Source tokens that are not expected to have a solution in SAMA
- For NW ATB3 corpus, almost entirely
  - Punctuation
  - Numbers written as digits (not part of transcription specs for BN)
- For BN, 12,843 Status #4 tokens include
  - 4,760 with a DIALECT tag (almost none in NW)
  - 3,001 with a TRANSERR tag (very few analogous TYPo in NW)
  - 4,765 with a PARTIAL tag (no partial words in NW)
- In addition, numbers in BN are transcribed as written out words rather than as digits, and so are not included as Status #4 for BN

3. Transcription issues
- Manual transcription, potential transcription errors
- Affect downstream annotation in Arabic-specific ways
- Partial/incomplete words given POS “PARTIAL”
- Errors in transcription given POS “TRANSERR”
- Similar to TYPo tag used for text corpora

Transcription examples:
- Partial word:
  - Trailing hyphen indicating transcribed incomplete word → “PARTIAL” POS tag
  - Typically inside a tree node marked → UNF for unfinished, and included as SAMA Status #4

Transcription error:
- Token is left as transcribed
  - For consistency with other annotation work on the same transcribed corpus
  - POS = TRANSERR, but syntactic annotation = as if written correctly

SAMA status # BN tokens % BN tokens in tokens status % ATB3 tokens in status
#1 INCLUDED in SAMA 415,924 96.1% 84.6%
#2 LIMITED solution (no vocalization) 735 0.1% 0.3%
#3 PENDING SAMA solution 3,474 0.8% 1.3%
#4 EXCLUDED from check with SAMA 12,843 3.0% 13.9%
TOTAL 432,976

Status of BN Corpus Integration with SAMA
- Status flag for each source token to make explicit the connection between morphological analysis from Standard Arabic Morphological Analyzer (SAMA) and ATB POS annotation

Status #4 EXCLUDED from check with SAMA
- Source tokens that are not expected to have a solution in SAMA
- For NW ATB3 corpus, almost entirely
  - Punctuation
  - Numbers written as digits (not part of transcription specs for BN)
- For BN, 12,843 Status #4 tokens include
  - 4,760 with a DIALECT tag (almost none in NW)
  - 3,001 with a TRANSERR tag (very few analogous TYPo in NW)
  - 4,765 with a PARTIAL tag (no partial words in NW)
- In addition, numbers in BN are transcribed as written out words rather than as digits, and so are not included as Status #4 for BN

Parsing Evaluation: ATB vs. BN
- Parser modes
  - Free to choose a POS tag for each word
  - Forced to use the gold tags

- Two sets of BN data
  - As released (row 2)
  - EDITED nodes removed (row 3)
- Results improve, since EDITED nodes are inherently difficult for parser
- Different nature of BN data remains:
  - UNF, DIALECT, TRANSERR

BN results still close to ATB3 NW, despite the difficult nature of the corpus

Conclusion
- Lessons learned from this first large corpus of treebanked Arabic speech, both annotation and technical
- Perhaps somewhat surprisingly, in some respects the BN data is actually more consistent than NW data
- These lessons will inform our methodologies as we continue to expand the Arabic Treebank into less formal speech and web text domains, where a greater impact from dialects and vernacular usage is expected