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

http://projects.ldc.upenn.edu/ArabicTreebank/

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: اليدよい (he/it means) much like “you know” in English
  - Annotated as PRN
  - The two forms require different POS and tree syntactic analyses

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 “TRANERR”
  - Similar to TBY 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

4. 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 # | # BN source tokens | % BN tokens in 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 #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 TRANERR tag (very few analogous TYPs 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

- 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, TRANERR

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

| BN | ATB3 | ATB-BN | ATB-NB |
|----|------|--------|--------|
| Words | 17,854 | 78.2 | 79.6 |
| Parser chooses tags | 28,058 | 76.1 | 77.8 |
| Parser uses gold tags | 28,378 | 77.2 | 76.9 |

Initial hamza:

- BN: All initial hamzas (glottal stops) are heard and transcribed with either <i> or <a>, such as /<i>an-/ (<i>is indeed) or /<a>an-/ (<a>that)
- Newswire (NW): Neutralized /An form very common (1.5% of tokens in ATBS)
- Annotators forced to distinguish between <i> or <a>-forms based on context
- The two forms require different POS and tree annotations, different guidelines

For initial hamza, transcribed speech data actually presents fewer issues for downstream annotation than written NW data

He wants to become a lawyer

http://projects.ldc.upenn.edu/ArabicTreebank/