Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines

Mohamed Maamouri, Ann Bies, Seth Kulick
Linguistic Data Consortium
University of Pennsylvania, USA
{maamouri,bies,skulick}@ldc.upenn.edu

Enhanced and revised Arabic Treebank Guidelines

- Arabic Treebank (ATB) at Linguistic Data Consortium
- Revised and enhanced annotation guidelines and procedure over the past year
- Now being applied in annotation production
- More complete and detailed annotation guidelines overall
- Period of intensive annotator training
- Inter-annotator agreement f-measure scores improved to 94.3%.
- Parsing results improved to 79.7 f-measure

Motivation

- Examination of inconsistencies in annotation
  - Relationship between Part-of-Speech (POS) tags and the syntactic (TB) annotation
  - Priority of semantic vs. syntactic structures in some syntactic constructions
- Lower than expected initial parsing scores

Additional goals

- Representing more finely-grained distinctions
- Aligning more closely with traditional grammar concepts already familiar to annotators

Morphological/Part-of-Speech level: More fine-grained distinctions

- NOUN (common noun)
- NOUN_NUM (number)
- NOUN_QUANT (quantifier)
- NOUN_PROP (proper noun)
- ADJ (common adjective)
- ADJ_NUM (ordinal number)
- ADJ_COMP (comparative adjective)

POS: Follow Arabic traditional grammar

- List of prepositions strictly limited to traditional grammar list (most lexical items previously PREP now categorized as NOUN, or "prepositional nouns")
- Particles now given several POS alternatives: *fA'* فاء
  - CONJ for *fA' Al-EaTf/* فاء العطف for coordination: '*and*'
  - CONNEC_PART for *fA' Al-rabT/* فاء الربط comment after focus particle *>am~A /* أمّا : '*well (then)*'
  - RC_PART for *fA' Al-jazA'/* فاء الجزاء as *a* Response Conditional to introduce result of preceding conditional clause: '*then*' or '*so*'
  - SUB_CONJ for *fA' Al-sababiy~ap/* فاء السببية to introduce subordinate result clause: 'so that'

Syntactic/Treebank level: Quantifiers

- Subordinate semantic needs to syntactic needs in certain constructions (for example, idafa with quantifiers)
- High frequency of idafa construction → change has significant effect on corpus

(NP كتاب kitaAbu book
(NP نحو naHowK grammar))
كتاب نحو *'(a) grammar book'*

(NP every/all/each_one |-kul~u | - كُلُّ
(NP collection/group | majomuwEapK | مَجْمُوعَةٍ))
كُلُّ مَجْمُوعَةٍ *'every collection'*

POS + TB → Quantifier improvement

- POS + TB changes → conceptual and practical improvement

New NOUN_QUANT more fine-grained POS tag
every/all/each_one |-kul~u | - كُلُّ /NOUN_QUANT
+
Priority of syntax over semantics for quantifiers
syntactic head = every/all/each_one |-kul~u | - كُلُّ
semantic head = collection/group | majomuwEapK | مَجْمُوعَةٍ
=
Semantic head information easily recoverable by the end user, since "every" would be clearly marked as a quantifier
+
Inter-annotator agreement is higher

TB: Follow Arabic traditional grammar

- Constructions including
  - Comparatives
  - Numbers and numerical expressions
  - Several pronominal constructions such as
    - Separating pronouns/Damiyr Al-faSl/ضمير الفصل
    - Anticipatory pronouns/Damiyr Al$a>n/ضمير الشأن
  - More careful and complete classification of verbs and their argument structure
  - Thorough treatment of gerunds, participles and verbal nouns

TB: Gerunds and participles

- Simultaneous dual verbal and nominal role possible
  - Behaving as verbs with respect to their objects (assigning accusative case, for example) and complements
  - Behaving as nominals with respect to the rest of the sentence (occupying canonically noun phrase positions, for example, such as the subject of the sentence, or the complement of a preposition)

Representation in Treebank

(S (VP rafaDat رَفَضَتْ
(NP-SBJ Al+suluTAtu السُّلطاتُ)
(S-NOM-OBJ
(VP manoHa مَنْحَ
(NP-SBJ *)
(NP-DTV Al>amiyri الأمير
AlhAribi الهارب)
(NP-OBJ (NP jawAza جَوازَ
(NP safarK سَفَر))
(ADJP dyblwmAsy~AF دِيبلوماسيّاً)))))

رفضت السلطات منحَ الأمير الهارب جوازَ سفر دبلوماسياً
The authorities refused to give the escaping prince a diplomatic passport

Enhanced guidelines: Gerunds and participles

- More comprehensive about contexts for verbal reading
  - Greater grammatical precision → increased inter-annotator agreement
  - For example, a maSdar/gerund, active participle or passive participle followed by a PP complement to the regular verb form (PP-CLR) is now shown with a verbal reading

Verbal participle example

- **Active participle with PP-CLR complement:**

(NP (NP Almawoqifu المَوْقِفُ)
(SBAR (WHNP-1 *0*)
(S (VP AlmuEab~iru المُعَبِّرُ
(NP-SBJ-1 *T*)
(PP-CLR Ean عَنْ
(NP ra}iyi رأي
(NP Al>aglabiy~api الأغلبيَّةِ)))))))
الموقف **المعبر عن** رأي الأغلبية
The attitude which shows the opinion of the majority

Correcting previous annotation level

- Treebank annotators can correct POS tags from the earlier annotation level
- Crucial to reducing the number of mismatches between POS tags and syntactic structures
- Available corrections in the tool include
  - Correcting the core POS tag (changing an active verb tag to a passive verb tag, for example)
  - Correcting tokenization errors
  - Correcting case endings
  - Correcting case endings and specific POS tags such as CONJ → ADV or PREP → NOUN
- Experiments in automatic corrections

Improvements in inter-annotator agreement and training

- Intensive annotator training focused on agreement and consistency
- Improvement of inter-annotator agreement scores (evalb) to f-measure 94.3%
  - Compare to reported score of 93.8% for Chinese Treebank
- Automatic and manual error correction
- 10% dual blind annotation

Error analysis and parsing improvement

- Two main problems parsing ATB
  - Inconsistent or unexpected Part-of-Speech tags in particular syntactic configurations (mismatches)
  - Inconsistent annotation of certain syntactic structures
- Analysis of these problems informed the early direction of the Treebank revision work described here, and continues to be at the core of the ongoing work

Parsing improvement

- Using revised and enhanced annotation as training data → preliminary improvement in statistical parsing results
  - **75.1** f-measure = Initial parsing results for 100K words of the pre-revision Arabic Treebank Part 3 v. 2.0
  - **76.2** = For first 100K F-measure
  - **79.7** = Improves with additional data: 200K F-measure
  - Expect on-going QC and other work will reach the intermediate goal: matching parsing of Chinese 230K 82.7

Fruitful collaboration

- Improved ATB guidelines + improving inter-annotator agreement scores + expected continuing improvement in parsing scores
- Result of fruitful collaboration between data producers and end users, along with the support of sponsors and time to effect the change
- It is hoped that such collaboration will continue to benefit both annotation production and NLP applications in the future
- For a more complete description of the new annotation policies, see the *Arabic Treebank Morphological and Syntactic Annotation Guidelines* (2008) **http://projects.ldc.upenn.edu/ArabicTreebank/**