Creating a Methodology for Large-Scale Correction of Treebank Annotation: The Case of the Arabic Treebank

Mohamed Maamouri, Ann Bies, Seth Kulick
Linguistic Data Consortium
University of Pennsylvania
{maamouri,bies,skulick}@ldc.upenn.edu
## Arabic Treebank Newswire Corpora Sizes

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Source Tokens</th>
<th>Tokens after Clitic Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATB1: AFP</td>
<td>145,386</td>
<td>167,280</td>
</tr>
<tr>
<td>ATB2: Umaah</td>
<td>144,199</td>
<td>169,319</td>
</tr>
<tr>
<td>ATB3: Annahar</td>
<td>339,722</td>
<td>402,246</td>
</tr>
<tr>
<td>ATB123 Total</td>
<td>629,307</td>
<td>738,845</td>
</tr>
</tbody>
</table>
Enhanced and revised Arabic Treebank (ATB)
Preview of key features & results

- Revised and enhanced annotation guidelines and procedure over the past 2 years. More complete and detailed annotation guidelines overall.
- Combination of manual and automatic revisions of existing data to conform to new annotation specifications as closely as possible (ATB123)
- Now being applied in annotation production
- Period of intensive annotator training
- Inter-annotator agreement f-measure scores improved to 94.3%.
- Parsing results improved to 84.1 f-measure
What is a Penn-Style Treebank

Penn-Style Treebanks are annotated CORPORA, which include linguistic information such as:

- Constituent boundaries (Clause, VP, NP, PP, ...)
- Grammatical functions of words or constituents
- Dependencies between words or constituents
- Empty categories as place holders in the tree for pro-drop subjects and traces
The authorities refused to give the escaping prince a diplomatic passport
Choice of Morphological Annotation Style

◆ BAMA: Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002)
◆ Input string → Analyzer provides
  ● fully vocalized solution (Buckwalter Transliteration)
  ● unique identifier or lemma ID
  ● breakdown of the constituent morphemes (prefixes, stem, and suffixes)
  ● their POS values
  ● corresponding English glosses
◆ Guidelines available at http://projects.ldc.upenn.edu/ArabicTreebank/
Morphological Annotation Tool Screenshot
Choice of Syntactic Annotation Style

◆ Similar to Penn Treebank II
◆ Accessible to research community
◆ Based on a firm understanding and appreciation of traditional Arabic grammar principles
◆ Guidelines available at http://projects.ldc.upenn.edu/ArabicTreebank/
Syntactic Annotation Tool Screenshot
Revision Process

◆ Motivation
  ● Examination of inconsistencies in annotation
  ● Lower than expected initial parsing scores

◆ Complete revision of annotation guidelines, both morphological and syntactic

◆ Combined automatic and manual revision of annotation in existing corpora: ATB₁ (AFP), ATB₂ (Umaah), ATB₃ (Annahar)
## Stages of Correction

<table>
<thead>
<tr>
<th>Stage</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Complete manual revision of trees according to new guidelines</td>
<td><strong>Human</strong> only</td>
</tr>
<tr>
<td>2. Limited manual correction of targeted POS tags</td>
<td><strong>Human</strong>, based on <strong>automatic</strong> identification</td>
</tr>
<tr>
<td>3. Revision of targeted tokenization and POS tags according to new guidelines, based on purely lexical information</td>
<td><strong>Automatic</strong> only</td>
</tr>
<tr>
<td>4. Revision of targeted tokenization and POS tags according to new guidelines, based on tree structure information</td>
<td><strong>Automatic</strong>, based on <strong>human</strong> trees</td>
</tr>
<tr>
<td>5. Corrections based on targeted error searches</td>
<td><strong>Human</strong>, based on <strong>automatic</strong> identification</td>
</tr>
</tbody>
</table>
Manual and Automatic Revision

- **Stage 1** focused on a human revision of all of the trees.
- **Stages 2, 3 & 4** focused on revising lexical information, based in part on the new tree structures, using a combination of automatic and manual changes.
- **Stage 5** focused on error searches targeting both lexical information and tree structures.
Stage 1: Manual Revision of Trees

- Introduction of iDAfa structure, e.g. (formerly flat NPs)

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

(NP كُلُّ -kul-u every -kul-
  (NP collection majomuwEapK ((مَجمُوعَةٌ مَجمُوعَةٌ 
كُلَّ مَجمُوعَةٌ
every collection

MEDAR 2009
Stage 2: Manual correction of targeted POS tags

- Specific tokens ambiguous with respect either to multiple POS tags or to tokenization were revised by hand (about 13 passes deemed important include such tokens as wa-, fa-, laysa, <il~A, Hat~aY etc.)

- **Example: mA values in SAMA**
  
  1. mA/REL_PRON what/which
  2. mA/NEG_PART not
  3. mA/INTERROG_PRON what/which
  4. mA/SUB_CONJ that/if/unless/whether
  5. mA/EXCLAM_PRON what/how
  6. mA/NOUN some
  7. mA/VERB not be
  8. mA/PART [discourse particle]
mA: Relative Pronoun vs. Negative Particle

mA=REL_PRON

لِيَحْصُلَُ عَلَىُ ماُ يَسُدُُّ رَمَقَهُ

li+yaHoSula ElaY mA yasud~u ramaqa+hu

for+gets (he) what fill breath of life+his

in order for him to get what he really craves

mA=NEG_PART

ماُزالَُحَيّا ُإِلَىُالآنَ

mA zAla Hay~AF <ilaY Al|na

not finished (he) alive until the+now

He doesn’t cease to be alive now
Ma SUB_CONJ vs. mA REL_PRON

- بعدَما أظهرته لَه
  after she showed (it) (to) him
- بعدَ ما أظهرته لَه
  After what she showed (it) (to) him
- بعدَما أظهرته لَه مِن حُبٍّ
  after she showed (it) (to) him of love
- بعدَ ما أظهرته لَه مِن حُبٍّ
  after what she showed (it) (to) him of love
Stage 3: Automatic revision of targeted tokenization and POS tags based on lexical information only

- Use lexical information in revised guidelines and new SAMA for “function words” as in PREP → NOUN
- Create a version of the corpus associating each original token from the source text file with the one or more Treebank tokens that together make up that original token
- Use this characterization of all original tokens to modify the tokenizations to match the new guidelines
  - Example: “limA*A” لمَاذَا → single token in new guidelines, from both single token and two token forms (“li” and “mA*A”) in pre-revision corpus
### Stage 4: Automatic revision of targeted tokenization and POS tags based on lexical and tree information

<table>
<thead>
<tr>
<th>Original unvocalized token</th>
<th>Possible vocalization/POS alternatives</th>
<th>Count in ATB123</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;nmA or AnmA</td>
<td>&lt;in~amA/RESTRIC_PART</td>
<td>138</td>
</tr>
<tr>
<td>إنما انما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>يِفِما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;in A</td>
<td>&lt;in A/RESTRIC_PART</td>
<td>14</td>
</tr>
<tr>
<td>فيما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;nm A</td>
<td>&lt;nm A/RESTRIC_PART</td>
<td>225</td>
</tr>
<tr>
<td>إنما انما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>يِفِما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;in A</td>
<td>&lt;in A/RESTRIC_PART</td>
<td>125</td>
</tr>
<tr>
<td>وما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;nm A</td>
<td>&lt;nm A/RESTRIC_PART</td>
<td>233</td>
</tr>
<tr>
<td>إنما انما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>يِفِما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;in A</td>
<td>&lt;in A/RESTRIC_PART</td>
<td>398</td>
</tr>
<tr>
<td>كما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;nm A</td>
<td>&lt;nm A/RESTRIC_PART</td>
<td>232</td>
</tr>
<tr>
<td>انما انما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>يِفِما</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;in A</td>
<td>&lt;in A/RESTRIC_PART</td>
<td>15</td>
</tr>
<tr>
<td>لما</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Stage 5: Manual corrections of automatic search results

- Searches targeting several types of potential inconsistency and annotation error
- Increased the number of error searches threefold during the revision process
- Run searches after annotation is complete
- Hand-correct all errors detected
A certain residual type of correction is not possible in this context

- Corrections that require too much human decision to be made automatically
- But that are too frequent or otherwise too time-consuming to be made manually

Example: highly complex and very frequent noun (NOUN) vs. adjective (ADJ) distinction in Arabic

Time and funding allowing, a manual revision of these cases in the Arabic Treebank will be undertaken in the future, using an appropriate combination of automatic and manual means.
Parsing Experiment: Significant Improvement using Revised Data

◆ New ATB and old ATB:
  ● Parsed ATB1,2,3 separately and ATB123 together
  ● Mona Diab’s train/dev/test split (<=40 words)
  ● Using gold tokenization and tags
  ● Two modes
    ■ Parser uses its own tags for “known” words
    ■ Parser forced to use given tags for all words
  ● LDC reduced TAG set (+DET)

◆ Penn (English) Treebank
  ● Made up training, test sets same size as ATB3, 123
Nice improvement, not at PTB level yet, but closer

Results not as good for test section

Dependency Analysis shows:
- Improvement in recovery of core syntactic relations
- Problem with PP attachment!

(Kulick, Gabbard, Marcus TILT 2006, Gabbard & Kulick 2008 ACL)
Concluding Remarks

- Revised and enhanced guidelines
- Revised annotation in existing data
- Increased consistency
- Improved parsing results
- Combined manual and automatic corrections crucial to the revision process
THANK YOU FOR YOUR ATTENTION

For more information or if you have any questions please contact
Dr. Mohamed MAAMOURI
<maamouri@ldc.upenn.edu>