From Arabic Handcrafted Grammar to Statistical Parsing

Mohammed Attia
NCLT, DCU
Outline

- Introduction: Why linguistics?
- Handcrafted grammar, a quick overview
- Tokenization
- Morphological Analysis
- Multiword Expressions
- Handcrafted grammar evaluation
- Statistical Parsing evaluation
- Which is Better?
Introduction

- Why linguistic knowledge is needed?
Algorithms and Data Structure

DCU Annotation → Parsing → Training → Bikel Parser
Algorithms and Data Structure

DCU Annotation

Parsing

Training

Tokenizer

Morphological Analyser

Treebank

Linguistics
Why Linguistics

● Linguistic Data is a naughty blackbox:
  – You get non-deterministic answers
  – You can get wrong answers
  – For the same question, you can get a set of inconsistent answers

● We need to make the algorithms suite the data structure, and we also need to make sure that the data is structured properly.
Handcrafted Grammar: A Quick Overview

Sentence

ساعدة الهيئة الفلسطينية
sāʿ adat al-haiʿatu al-filistīniyyīn/ al-filistīniyyain
helped the-agency the-Palestinian.pl/ the-Palestinian.dual
‘The agency helped the Palestinians/ the two Palestinians.’

Tokenization

مساعدة الهيئة الفلسطينيين
helped@the@agency@the@Palestinians
Handcrafted Grammar: A Quick Overview

Morphological analysis

Helped: +verb+past+active+1pers
+verb+past+active+3pers+sg+fem
+verb+past+active+2pers+sg+fem
+verb+past+active+2pers+sg+masc

The: +defArt

Agency: +noun+nonhuman+fem+sg

Palestinians: +adj+mas+dua+accgen
+adj+mas+pl+accgen
+noun+human+mas+dua+accgen
+noun+human+mas+pl+accgen
Handcrafted Grammar: A Quick Overview

Lexicon (Lexical properties/subcategorization frames)

**helped**

\[ V \text{ XLE} (^{\text{GLOSS}})=\text{help} "\text{This verb has three different subcat frames}"
\]

\[ \{ (^{\text{PRED}})='\%stem'(^{\text{SUBJ}})(^{\text{OBJ}})(^{\text{COMP}})>'
\]

\[ (^{\text{COMP}} )^{\text{COMP-FORM}}=c \text{ أن } (^{\text{COMP}} )^{\text{COMP-TYPE}}=c \text{ verbal}
\]

\[ | (^{\text{PRED}})='\%stem'(^{\text{SUBJ}})(^{\text{OBJ}})(^{\text{OBL}})>' (^{\text{OBL}})\text{ OBJ PCASE}=c \text{ على}
\]

\[ | (^{\text{PRED}})='\%stem'(^{\text{SUBJ}})(^{\text{OBJ}})>'\}.
\]

**agency**

\[ N \text{ XLE} (^{\text{GLOSS}})=\text{agency} (^{\text{PRED}})=’\%stem’ (^{\text{PERS}})=3
\]

\[ \{ (^{\text{NUM}}) (^{\text{NUM}}) \sim = sg | (^{\text{NUM}}) = sg \} "\text{the default number is singular}".
\]

**Palestinian**

\[ N \text{ XLE} (^{\text{GLOSS}})=\text{Palestinian} (^{\text{PRED}})=’\%stem’ (^{\text{PERS}})=3
\]

\[ \{ (^{\text{NUM}}) (^{\text{NUM}}) \sim = sg | (^{\text{NUM}}) = sg \} "\text{the default number is singular}";
\]

\[ \text{ADJ} \text{ XLE} (^{\text{PRED}})=’\%stem’ (^{\text{GLOSS}})=’\text{Palestinian’}
\]

\[ \{ (^{\text{ATYPE}})=c \text{ predicative } | (^{\text{ATYPE}})=\text{ attributive}\}.
\]
Handcrafted Grammar: A Quick Overview

Grammar Rules: PS-rules and functional equations

MT ARABIC RULES (1.0)

S_Nonequational --> "There are three word orders permitted in Arabic: VSO, SVO and VOS"
    { VSO
         | SVO
         | VOS }.

VSO --> V: ^=!. @DefSTense (^ VTYPE)~= copular (^ COMP-TYPE)=verbal
    { (^ SUBJ PRED)=c 'pro' (^ SUBJ NUM) = (^ AGR NUM)
         | (^ SUBJ PRED)~= 'pro' (^ AGR NUM)=sg}
    (^ AGR GEND) = (^ SUBJ GEND) (^ AGR PERS) = (^ SUBJ PERS);
    {NP: (^SUBJ)=! (! FIRST-CONJ)=+}
        (! CASE)=nom (! PRON-TYPE) ~=pers
    | e: (^ SUBJ PRED)='pro' "ProDrop"
        (^ AGR PERS) = (! PERS) (^ AGR NUM) = (! NUM) (^ AGR GEND) = (! GEND) 
    (NP: (^OBJ)=! (! CASE)=acc).
Handcrafted Grammar: A Quick Overview

Output: c-structures and f-structures

helped

the agency

the Palestinians
Tokenization
Tokenization in XLE

沃尔رجل
walilrajuli
wa@li@al@rajuli
and@to@the@man

ولليشكرون
wasayashkurunahu
wa@sa@yashkuruna@hu
and@will@thank[they]@him

Conjunction
Comp/Tense Marker
Stem with Affixes
Object Pronoun

Noun
Conjunction
Preposition
Definite Article
Stem with Affixes
Genitive Pronoun

Proclitics
Enclitic
Proclitics
Enclitic
Tokenization in XLE

Deterministic Tokenizer

والنجل (walirrağul: and to the man)
@لل@ال@نجل wa@li@al@rağul@ and@to@the@man@

Non-Deterministic Tokenizer

والنجل (walirrağul: and to the man)
@لل@ال@نجل wa@li@al@rağul@ and@to@the@man@
@لل@ال@نجل wa@li@al@rağul@ and@to@the@man@
@ال@نجل wa@li@al@rağul@ and@to@the@man@
@نجل wa@li@al@rağul@ and@to@the@man@
@نجل wa@li@al@rağul@ and@to@the@man@
@نجل wa@li@al@rağul@ and@to@the@man@
Tokenization in Bikel

- English parser
  - Input sentence:
    The President led his country in reform.
  - Formatted sentence:
    (The President led his country in reform.)

    (VBZ has) (RB n't)
    (NNP Chicago) (POS 's)
Tokenization in Bikel

- **English parser**
  - **Output:**
    
    (S (NP (DT The) (NNP President)) (VP (VBD led) (NP (PRP$ his) (NN country)) (PP (IN in) (NP (NNP reform.))))))
  
  - **Tree**
Tokenization in Bikel

- Arabic parser

![Diagram showing tokenization process for Verb and Noun categories. Each category branches into Conjunction, Preposition, Definite Article, Stem with Affixes, and Genitive Pronoun.]
Tokenization in Bikel

- Arabic parser
  - Input sentence:
    الرئيس قاد بلده في الإصلاح
    The President let his country in reform.
  - Formatted sentence:
    - Alra\{iysu qAda baladahu fiy Al<iSlaAHi
    - Alra\{iysu qAda balada- -hu fiy Al<iSlaAHi
    - Al+ra\{iys+u qAd+a balad+a- -hu fiy Al+<iSlaAH+i
    - (Al+ra\{iys+u qAd+a balad+a- -hu fiy Al+<iSlaAH+i)
Tokenization in Bikel

- Arabic parser
  - Output:
    $$(S \ (NP \ (NN \ Al+ra}{iys+u)) \ (VP \ (VBD \ qAd+a) \ (NP \ (NN \ balad+a-) \ (PRP$ \ -hu)) \ (PP \ (IN \ fiy) \ (NP \ (NN \ Al+<iSlAJAH+i))))$$
  - Tree
Morphological Analysis
Morphological Analysis in XLE

- Rule-based, finite state technology
- Contains 10,799 lemmas and 2,818 multiword expressions
- Suitable for both analysis and generation
- Based on contemporary data (a corpus of news articles of 4.5 million words)
- Truly MSA-specialized morphological analyser
Buckwalter Morphological Analysis

- Contains 38,600 lemmas
- Not rule-based
- Not suited for generation
- Does not handle multiword expressions
- Includes classical senses
  
 setbacks Hosam/sword
# Buckwalter Morphological Analysis

- Includes classical entries

<table>
<thead>
<tr>
<th>#</th>
<th>Meaning</th>
<th>Classical Word</th>
<th>Google</th>
<th>MSA Word</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sully</td>
<td>قلعط qal‘at</td>
<td>8</td>
<td>لطخ laṭṭaḥa</td>
<td>29,600</td>
</tr>
<tr>
<td>2</td>
<td>caulk</td>
<td>قلفط qalfaṭ</td>
<td>9</td>
<td>أفسد 'afsada</td>
<td>205,000</td>
</tr>
<tr>
<td>3</td>
<td>wear</td>
<td>استكد 'istakadda</td>
<td>4</td>
<td>أنهك 'anhaka</td>
<td>37,100</td>
</tr>
<tr>
<td>4</td>
<td>fickle</td>
<td>غملج gambar</td>
<td>7</td>
<td>متقلب mutaqallib</td>
<td>189,000</td>
</tr>
<tr>
<td>5</td>
<td>erosion</td>
<td>انتكال 'i’tikāl</td>
<td>7</td>
<td>تأكل ta ’ākul</td>
<td>1,700,000</td>
</tr>
</tbody>
</table>
Buckwalter Morphological Analysis

- Includes classical entries (Chauser's Canterbury Tales)

<table>
<thead>
<tr>
<th>#</th>
<th>Meaning</th>
<th>Classical Word</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sully</td>
<td>قلعط qal`at</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>caulk</td>
<td>قلفط qalfaṭ</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>wear</td>
<td>استکد 'istakadda</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>fickle</td>
<td>غملج ģamlaga</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>erosion</td>
<td>انتقال `i’tikāl</td>
<td>7</td>
</tr>
</tbody>
</table>
Buckwalter Morphological Analysis

- Excessive application of spelling relaxation rules
- Neglecting grammar-lexis specifications (e.g. adjectives do not combine with genitive pronouns)
  
  معادي muʿādī (hostile/anti- + my)

- This makes it highly ambiguous

  مصري mishriyy ‘Egyptian’

Attia 2 solutions
Buckwalter 10 solutions
Multiword Expressions
Multiword Expressions in XLE

• Three types of MWEs
  – Fixed Expressions: Lexically, morphologically and syntactically rigid. A word with spaces.
    • *New York*
    • *United Nations*
  – Semi-Fixed Expressions: Lexically, or morphologically flexible
    • *Sweep somebody under the rug/carpet*
    • *Transitional period(s)*
  – Syntactically-flexible Expressions
    • *to let the cat out of the bag*
    • *The cat was let out of the bag.*
Multiword Expressions in XLE

- MWEs are important
  - High frequency in natural language (30-40%)
  - Important for MT, literal translation is usually wrong
  - When taken as a block, they relieve the parser from the burden of processing component words
  - We have 2818 MWEs in our system in addition to 10799 lemmas in the morphology
Multiword Expressions in XLE

The United States looks for Bin Laden.
Multiword Expressions in Bikel

- Compositional, yet detectable in the English treebank

  (NP (DT the) (NNP United) (NNP Kingdom) )

  (NP (NNP New) (NNP York) )

  (NP (DT the) (NNP Middle) (NNP East) )

  (NP (NNP Saudi) (NNP Arabia) )

  (NP (NNP Las) (NNP Vegas) )

  (NP (NNP Los) (NNP Angeles) )

  (CONJP (IN in) (NN addition) (TO to) )
Multiword Expressions in Bikel

- Compositional, undetectable, sometimes inconsistent, in Arabic treebank

Los Angeles لوس أنجليس
(NP (NOUN_PROP luws)
  (NOUN_PROP >anojiliys))

United States الولايات المتحدة
(NP (DET+NOUN+NSUFF_FEM_PL+CASE_DEF_NOM Al+wilAy+At+u)
  (DET+ADJ+NSUFF_FEM_SG+CASE_DEF_NOM Al+mut~aHid+ap+u))

The Middle East الشرق الأوسط
(NP (DET+NOUN+CASE_DEF_GEN Al+$aroq+i)
  (DET+ADJ+CASE_DEF_GEN Al+>awosaT+i))

in addition to إضافة إلى
(CONJP (NOUN+NSUFF_FEM_SG+CASE_INDEF_ACC <iDAf+ap+F) (PREP <ilaY))

(NP-ADV (NP (NOUN+NSUFF_FEM_SG+CASE_INDEF_ACC <-iDAf+ap+F)) (PP (PREP <ilaY) (NP (NOUN_PROP EarafAt))))
Multiword Expressions in Bikel

• Example

The United States looks for Bin Laden.

الولايات المتحدة تبحث عن بن لادن

(S (NP (NNS Al-wilAy-At-u) (JJ Al-mut~aHid+ap+u)) (VP (VBP ta-boHav+u) (PP (IN Ean) (NP (NNP bin) (NNP IAdin))))))
XLE Arabic Grammar Development
XLE Arabic Grammar Development

- **Stage 1: Toy Grammar**
  - A test suite of 175 made-up sentences

- **Stage 2: Bulk Selection**
  - 4 articles from Al-Jazeera are chosen as a reference for development

- **Stage 3: Discriminative Selection**
  - We focused on sentences with 10-15 words in length
XLE Arabic Grammar Development

Distribution of Sentence Lengths

- 1-9: 20%
- 10-15: 15%
- 16-20: 15%
- 21-25: 35%
- 26-30: 20%
- 31-35: 15%
- 36-40: 10%
- 41-50: 5%
- 51-100: 25%
XLE Arabic Grammar Testing and Evaluation

• For sentences in the range of 10-15 words
  – 92% Fragment parsing
  – 33% Complete parses
Why are handcrafted grammars slow to develop?

- There is usually a few people working in the grammar.
- Development is hampered by linguistic (philosophical) issues that pop up frequently.
- Speed also depends on what tools (tokenizers, morphological analysers) are already available.
- Grammar writers are usually researchers who are more interested in linguistic phenomena than in coverage.
- No formal guidelines, training, or project management.
How can Arabic handcrafted grammar coverage be improved?

- Tripling the size of the morphology now 10,000 entries + 3,000 MWEs
  - This can now be done using statistical tools

- 1195 verbs that subcategorize for prepositions (3500 sents treebank)

- 161 adjectives and nouns that function as adverbs
How can Arabic hand-crafted grammar be improved?

- Acquire statistics about the frequency of constructions
  - Adjectives that function as a predicate in a copula construction
How can Arabic handcrafted grammar be improved?
Bikel Arabic Parser Evaluation

• Coverage of the statistical parser on sentence <= 40 words:
  – Arabic: 75.4%
  – Chinese: 81%
  – English: 87.4%

    (Bikel, 2004)

– Arabic is “far below” the required standard.

    (Kulick et al., 2006)
Bikel Arabic Parser Evaluation

• Why Arabic performs poorly? (Kulick et al. 2006)
  – The ATB tag set is very large and dynamic, this is why they are mapped into 20 PTB tags. The tagset reduction is extreme and important information is lost.
  
  – Verb
    • IV3FS+IV+IVSUFF_MOOD:I
    • IV3MS+IV+IVSUFF_MOOD:J
    • PV+PV+SUFF_SUBJ:3MS
    • IV+SUFF_DO:3MP

  – Noun
    • NOUN+CASE_DEF_ACC
    • DET+NOUN+NSUFF_FEM_PL+CASE_DEF_GEN
    • NOUN+NSUFF_FEM_SG+CASE_DEF_GEN
Bikel Arabic Parser Evaluation

- Why Arabic performs poorly? (Kulick et al. 2006)
  - Average sentence length in Arabic is 32 compared to 23 in English
  - Significant number of POS tag inconsistencies, for example *lys* is tagged as NEG_PART and PV
  - 5% of VP in Arabic have non-verbal heads
  - Base Noun Phrases (NPB) are 30% in English compared to 12% in Arabic.
  - Construct states in Arabic *roughly* correspond to possession constructions in English
Bikel Arabic Parser Evaluation

• Why Arabic performs poorly? (Kulick et al. 2006)
  – Arabic has a much greater variance in sentence structure than English.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Arabic %</th>
<th>English %</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSO</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>SVO</td>
<td>17</td>
<td>90</td>
</tr>
<tr>
<td>No VP</td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>Subjectless VP</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

• Major revision of Arabic treebank guidelines 08
Which is better?
Which is better?

- Common wisdom: handcrafted grammars are:
  - Time-consuming
  - Expensive
  - Require considerable linguistic and computational expertise
  - Lack coverage and robustness

(Burke et al., 2004)
Which is better?

- Common wisdom is not entirely true.
  - Creating a treebank is:
    - a “Herculean task” (Charniak, 1997)
    - very time-consuming
    - expensive
    - requires considerable linguistic and computational expertise
Which is better?

- Arabic treebank annotation (2001-2008)
  - Guidelines authored by:
    - Mohamed Maamouri
    - Ann Bies
    - Sondos Krouna
    - Fatma Gaddeche
    - Basma Bouziri

  With contribution of
  - Seth Kulick
  - Wigdane Mekki
  - Tim Buckwalter
Which is better?

- Arabic treebank annotation (2001-2008)
  - List of annotators (Part 2, 2004: 4519 sentences)
    - Wigdan Mekki
    - Tasneem Ghandour
    - Ichraf Amghouz
    - Zohra Bentaouit
    - Nourredine Bessaidi
    - Rachida Fathallah
    - Niama Laadioui
    - Abid Labidi
    - Dalal Zakhary
    - Fatma Gaddeche
    - Basma Bouziri
Which is better?

• Arabic treebank annotation (2001-2008)
  • List of annotators (Part 1, 2003: 2591 sentences)
    - Wigdan El Mekki
    - Ichraf Amghouz
    - Zohra Bentaouit
    - Fatima Chebchoub
    - Fatima El Himyani
    - Rachida Fathallah
    - Alexa Firat
    - Tasneem Ghandour
    - Niama Laadioui
    - Mohamed Mansour
    - Sarah Tlili
    - Gordon Witty
    - Dalel Zakhary
Which is better?

- Arabic treebank annotation (2001-2008)
  - Logistical issues
    - Automation tools and templates
    - Tests to ensure inter-annotator agreements
    - Investigation of linguistic phenomena
    - Guidelines for consistency
Which is better?

• Common wisdom: statistical parsers are:
  – Shallow: They do not mark syntactic and semantic dependencies needed for meaning-sensitive applications

(Kaplan et al., 2004)
Which is better?

- XLE: “We parse the web.”
Which is better?

- Common wisdom is not entirely true.
- DCU: “We can also parse the web.”
Which is better?

• Summary
  – Handcrafted grammars are built on assumptions and intuitions. They depend on how good these assumptions are.
  – Handcrafted grammar can be improved by:
    • Effectively managing the development project
    • Making use of statistical facts (treebanks, and TIGERSearch)
Which is better?

- Statistical grammars are built on facts. They depend on how true these facts are.
- Statistical grammar can be improved by:
  - Improving the quality and size of treebanks.
Which is better?

- Statistical grammars are more efficient because:
  - there is a clear separation between the algorithm and the data structure
  - there is a clear division of labour, the linguists fight their battle, and the engineers fight their own battle
Which is better?

- Hybridization? Complementation? Cooperation?
  - Sttical parser is used to increase the efficiency of hand-crafted grammar (pruning the search space)
  - Hand-crafted grammars are used automate the creation of treebanks (Norwegian grammar)
  - Some languages do not have a treebank

- This is for the future to decide.