Developing a Robust Arabic Morphological Transducer/Tokenizer, and Integration with XLE

By
Mohammed A. Attia
Ph.D. Student,
School of Informatics,
The University of Manchester

Introduction

Available Arabic Morphological Analyzers:

- Xerox Finite State Arabic Morphological Analyzer
- Buckwalter Arabic Morphological Analyzer

Introduction

Arabic Morphological Peculiarities

- Large number of prefixes and suffixes to show person, number and gender with verbs, and number and gender with nouns
- Separated Dependencies
- Clitics

A New Arabic Transducer - Why?

- Specific domain – News
- Specific language – MSA
- Specific purpose – MT
- Compatibility – XLE
- Native script
- Maintenance and update
- Owning tools: customizability in form and content

Development Decision

- Using finite state technology with the Advantages:
  - Handling concatenative and non-concatenative morphotactics
  - Fast and efficient
  - Unicode support
  - Multi-platform support

Development Decision

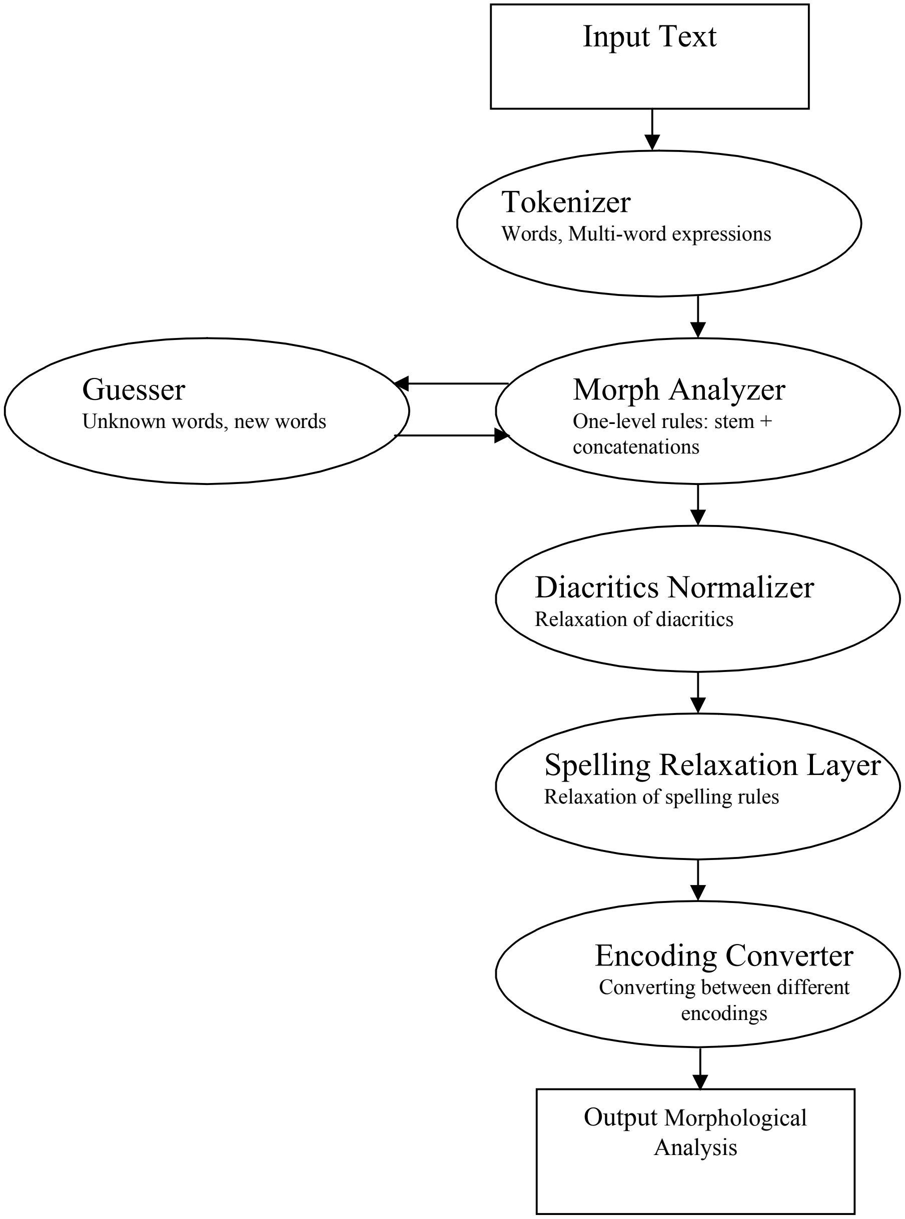
- Using the stem as the base form, which makes the solution:
  - Easier and faster to develop
  - More suitable for translation

Development Decision

- Separating the task of the developer and the lexicographer
- Making no account of diacritics
- Generating valid surface forms
- Developing a guesser to prevent the system from failing

System Architecture

- Tokenizer
- Morphological Transducer
- Guesser
- Diacritics Normalizer
- Spelling Relaxation Layer


Input Text
Tokenizer
Words, Multi-word expressions
Guesser
Unknown words, new words
Morph Analyzer
One-level rules: stem + concatenations
Diacritics Normalizer
Relaxation of diacritics
Spelling Relaxation Layer
Relaxation of spelling rules
Encoding Converter
Converting between different encodings
Output Morphological Analysis

Verb Morphotactics

Possible Concatenations

(Conjunctions or question Article)	(Complementizers)	Tense Prefixes	Verb Stem	Tense Suffixes	(Clitic Object Pronouns)
Conjunctions و "wa" (and) or ف "fa" (then)	ل "li" (to)	Present tense prefixes (5)	Stem	Present tense suffixes (10)	First person object pronoun (2)
Question word أ "a" (does or did)	س "sa" (will)	Past tense prefix (1)		Past tense suffixes (12)	second person object pronoun (5)
	ل "la" (then)	Imperative prefix (2)		Imperative suffixes (5)	Third person object pronoun (5)

Verb Morphotactics

- Statistically these (unconstrained) concatenations can generate up to: 33,696 Forms
  3 * 4 * 8 * 27 * 13
- Flag Diacritics are used to handle separated dependencies (constrained concatenations)
- 2,552 well-formed forms for transitive verbs

Verb Morphotactics

Alternation Rules

- Over 60 replace rules to handle alternation rules with “weak letters”
  - Verbs with an initial glottal stop, long vowel or glide
  - Verbs with a medial glottal stop, long vowel or glide. With verbs more than three letters long, their position inside the word can have effective difference.
  - Verbs with a final glottal stop, long vowel or glide.
  - Verbs that contain a doubled letter in the second, third, fourth, fifth or sixth position.

Noun Morphotactics

Possible Concatenations

(Conjunction or question Article)	(Preposition)	(Definite Article)	Noun Stem	(Suffixes)	(Clitic Genitive Pronoun)
Conjunctions و "wa" (and) or ف "fa" (then)	Feminine Mark (1)	ال "al" (the)	Stem	Masc Dual (4)	First person pronoun (2)
				Fem Dual (4)	
Question word أ "a" (does or did)				Masculine regular plural (4)	second person pronoun (5)
				Third person pronoun (5)	Third person pronoun (5)

Noun Morphotactics

- Statistically these (unconstrained) concatenations can generate up to 6,240 forms
  4 * 4 * 2 * 15 * 13
- Constrained concatenations generate 519 valid forms

Noun Morphotactics

Noun Types according to gender and number

- 13 Types
- Valid inflections must be specified in the lexicon

	Masculine Singular	Feminine Singular	Masculine Dual	Feminine Dual	Regular Masculine Plural	Regular Feminine Plural	Broken Plural
1	jahil (ignorant)	jahilah	jahilan	jahilatan	jahilun	jahilat	juhala’
2	mu’allim (teacher)	mu’allimah	mu’alliman	mu’allimatan	mu’allimuun	mu’allimat	X
3	talib (student)	talibah	taliban	talibatan	X	Talibat	tullab
4	ta’limi (educational)	Ta’limiah	ta’limian	ta’limiatan	X	X	X
5	imtihan (exam)	X	Imtihanan	X	X	Imtihanat	X
6	kitab (book)	X	kitaban	X	X	X	kutub
7	X	shajarah (tree)	X	shajaratan	X	shajarat	shajar
8	X	hamsah (whisper)	X	hamsatan	X	hamasat	X
9	X	shams (sun)	X	shamsan	X	X	shumus
10	tanazul (waiver)	X	X	X	X	tanazulat	X
11	khuruj (exit)	X	X	X	X	X	X
12	Mohammed	X	X	X	X	X	X
13	X	Zainab	X	X	X	X	X

Noun types according to number and gender

	Masculine Singular	Feminine Singular	Masculine Dual	Feminine Dual	Regular Masculine Plural	Regular Feminine Plural	Broken Plural
1	Yes	Yes	Yes	Yes	Yes	Yes	Yes
2	Yes	Yes	Yes	Yes	Yes	Yes	No
3	Yes	Yes	Yes	Yes	No	Yes	Yes
4	Yes	Yes	Yes	Yes	No	No	No
5	Yes	No	Yes	No	No	Yes	No
6	Yes	No	Yes	No	No	No	Yes
7	No	Yes	No	Yes	No	Yes	Yes
8	No	Yes	No	Yes	No	Yes	No
9	No	Yes	No	Yes	No	No	Yes
10	Yes	No	No	No	No	Yes	No
11	Yes	No	No	No	No	No	No
12	Yes (Prop)	No	No	No	No	No	No
13	No	Yes (Prop)	No	No	No	No	No

Noun Morphotactics

Broken plurals are not handled in a rule-based approach. The problem with broken plural:

- 30 singular noun templates served by 39 broken plural templates
- Broken plural forms are fossilized
- They are to be entered by hand

Function Words Morphotactics

- Conjunctions
- Pronouns
- Prepositions
- Modal Verbs
- Question Words

- Demonstratives
- Relatives
- Particles
  - Confirmation
  - Negation
  - Exception
  - Complementization
  - Future
  - Condition

Function Words Morphotactics

Function words take either:

- No prefix or suffix
  - Independent conjunctions
- Conjunction prefixes and no suffix
  - Independent Pronouns
- Conjunction prefixes and a pronoun prefix
  - Modals
- Conjunction and preposition prefixes and no suffix
  - Demonstrative pronouns

Analysis

- Ambiguities
  - Active vs. Passive vs. Imperative
    - كرم
      - Karrama (Active)
      - Kurrima (Passive)
      - karrim (Imperative)
  - 2nd Person Masc vs. 3rd Person Fem
    - تشكر
      - tashkur (2nd Person Masc)
      - tashkur (3rd Person Fem)
  - 1st Person sg vs. 3rd person fem
    - شكرت
      - shakartu (1st Person sg)
      - Shakarat (3rd person fem)

Analysis

- Ambiguities
  - Different Entries
    - أقال
      - aqala (+QuestionParticle [qala])
      - aqala
  - Different POS
    - شكر
      - shakara (verb)
      - shukr (noun)

Analysis

معلم	+3pers+noun+masc[معلم]
طالب	+3pers+noun+masc[طالب]
امتحن	+imp[امتحن]+2pers+masc+sg
امتحن	+past+active[امتحن]+3pers+sg+masc
امتحن	+past+active[امتحن]+3pers+pl+fem
امتحن	+past+pass[امتحن]+3pers+sg+masc
امتحن	+past+pass[امتحن]+3pers+pl+fem
شكر	+past+active[شكر]+3pers+sg+masc
شكر	+past+pass[شكر]+3pers+sg+masc
فهم	+conj+pron+3pers+pl+masc
فهم	+conj+obj3+them
علم	+past+active[علم]+3pers+sg+masc
علم	+past+pass[علم]+masc+sg
انهزم	+imp[انهزم]+2pers+masc+sg
انهزم	+past+active[انهزم]+3pers+sg+masc
انهزم	+past+pass[انهزم]+masc+sg
استعان	+past+active[استعان]+3pers+sg+masc

Generation

- Generating valid forms
- Eliminating ill-formed forms
- Accommodating spelling variation and common spelling errors in analysis but not in generation

Tokenization

Whereas the morphological transducer provides analysis, The tokenizer is responsible for identifying:

- Word boundaries
- Multi-word expressions
- Punctuation
- Abbreviations
- Clitics

Tokenization and Analysis: First Approach – 2 in 1

Why they are inseparable in dealing with Arabic clitics (prepositions, pronouns, conjuctions, etc.)

- Clitics can be concatenated one after the other.
- Clitics undergo assimilation with words.
- Without complete morphological knowledge, you cannot tell whether some initial or final letters are part of the word or only clitics.

Tokenization and Analysis: First Approach – 2 in 1

Implementation

- Tokenizer is responsible for deciding word boundaries, clitic boundaries as well as analysis
- Morphological analyzer: accepts the output of the tokenizer as is

In fact the core morphological analyzer is part of the tokenizer

Tokenization and Analysis: First Approach – 2 in 1

Implementation – Tokenizer output: +morph feature @token boundary

- وللرجل (and to the man)
  و+conj@ل+prep@ال+defArt@+nounرجل+masc@

- ولمعلمهم (and to their teacher)
  و+conj@ل+prep@+nounمعلم+masc@هم+genpron@

- وشكر (and he thanked/is thanked)
  و+conj@+verb+past+activeشكر+3pers+sg+masc@
  و+conj@+verb+past+passشكر+3pers+sg+masc@

- وليشكرهم (and to thank them)
  و+conj@ل+comp@+verb+pres+active+3persشكر+masc+sg@هم+objpron@

Tokenization and Analysis: Second Approach – Clitics Guesser

Step 1: Developing a guesser for Arabic words with all possible clitics, and accommodating possible assimilations. This guesser is then used by the tokenizer to mark clitic boundaries. There will be no analyses, but there will be increased tokenization ambiguities.

وللرجل (and to the man)

@رجل@ال@ل@و
@الرجل@ل@و
@للرجل@و
@وللرجل

Tokenization and Analysis: Second Approach – Clitics Guesser

Step 2: Developing a lexc transducer for clitics only, treating them as separate words. Then a morphological transducer is created by applying rules to remove all paths that contain any clitics from the core morphology. The output is then unioned with the clitics transducer.

Tokenization and Analysis: Second Approach – Clitics Guesser

Advantages:

1. Keeping the core morphology intact
2. Following the usual rule of separating the tokenizer and the analyzer.
3. Trees display more nicely in XLE.

Disadvantages:

1. The system has to deal with tokenization ambiguities. For a simple sentence of 3 words, I get 8 different tokenziation solutions.
2. I have to write stricter sublexical rules.
3. Treating clitics as free morphemes will create ambiguities with some originally free morphemes. Sometimes there will be an ambiguity also regarding whether this clitic belongs to the previous or the following word.

Integration

Integration with XLE: 4 Steps

- Adding a morphology section in the grammar file and referring to it in the grammar configuration section
- Setting the character encoding UTF-8 in the configuration section and in the test file
- Writing sublexical rules
- Writing sublexical entries

Integration

Problems with Arabic in XLE:

- Arabic fonts do not display correctly in trees and charts.
- when printing postscript for any chart, Arabic fonts disappear.
- You cannot write Arabic on the shell under Mac OS, and when you do under Linux the encoding is not interpreted correctly.

Conclusion

"Linguistic development is an endless round of observation, theorizing, formalizing and testing; and the goal, for a lexical transducer, is to create a system that correctly analyzes and generates a language that looks as much like the real natural language as possible."

Beesley and Karttunen, Finite State Morphology. P. 287

Conclusion

- FST is fast, efficient and reliable.
- Development time can be reduced significantly for Arabic if we take the stem as the base form and ignore diacritics.