2012 AMTA 20 Years

The Tenth Biennial Conference of the
Association for Machine Translation in the Americas

MT and Arabic Language Issues

Nizar Habash
Columbia University

Arabic poses many interesting challenges to machine translation: ambiguous orthography, rich morphology, complex morpho-syntactic behavior, and numerous dialects. In this tutorial, we introduce the most important themes of challenges and solutions for people working on translation from/to Arabic or any of its dialects. The tutorial is intended for researchers and developers working on MT. The discussion of linguistic issues and how they are addressed in MT will help linguists and professional translators understand the issues machine translation faces when dealing with Arabic and other morphologically rich languages. The tutorial does not expect the attendees to be able to speak/read/write Arabic.

Presenter

- Nizar Habash, PhD, Research Scientist at the Center for Computational Learning Systems at Columbia University.

AMTA 2012

San Diego, October 28, 2012

Machine Translation and Arabic Language Issues

Dr. Nizar Habash
Columbia University
Center for Computational Learning Systems
habash@ccls.columbia.edu

1

- Introduction
  - What is Arabic?
  - Relevant SMT concepts
- Arabic Orthography
- Arabic Morphology
- Arabic Syntax
- Arabic Dialects

2

What is Arabic?

- Arabic: the **script** and the **language**
- Arabic **script** is used to write a number of languages
  - Arabic, Farsi, Urdu, Pashto, etc.
- The Arabic **language** has many forms
  - Classical Arabic (CA)
    - Classical Historical & Liturgical texts
  - Modern Standard Arabic (MSA)
    - News media & formal speeches and settings
    - Only written standard
  - Dialectal Arabic (DA)
    - Predominantly spoken vernaculars
    - No written standards
- Dialect vs. Language
  - Linguistics vs. Politics

3

What is Arabic?

- ~300M people worldwide speak Arabic
  - Largest living Semitic language
- Arabic is the/an official language of 23 countries
- No native speakers of CA nor MSA
- In the Arabic speaking world, MSA and CA are the only forms of Arabic taught in schools

4

What is Arabic?

- Arabic Diglossia
  - Diglossia is where two forms of the language exist side by side
  - MSA is the formal public language
    - Perceived as “language of the mind”
  - Dialectal Arabic is the informal private language
    - Perceived as “language of the heart”
- General Arab perception: dialects are a deteriorated form of Classical Arabic

5

- Introduction
  - What is Arabic?
  - **Relevant SMT Concepts**
- Arabic Orthography
- Arabic Morphology
- Arabic Syntax
- Arabic Dialects

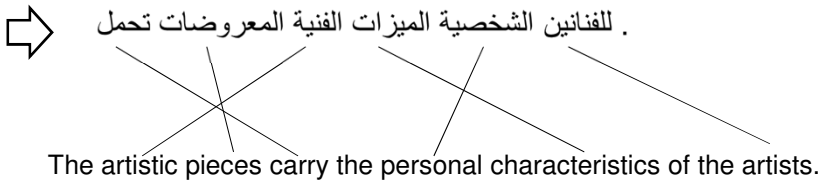
6

Relevant SMT Concepts

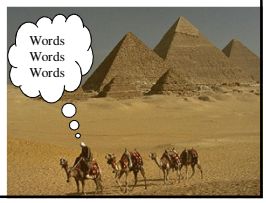
- Parallel Corpora
- Automatic Word Alignment
- Phrase Table Extraction
- Language Model
- Decoding
- Automatic Evaluation Metrics
- MT Pivoting
- Out-of-Vocabulary (OOV)

7

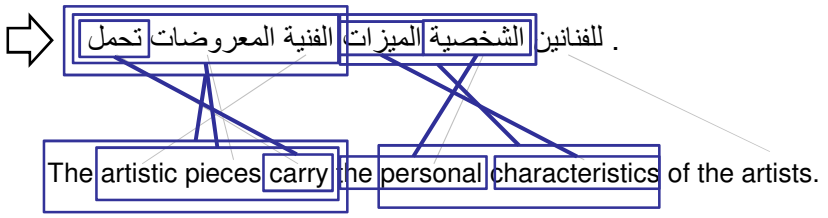
Word Alignment



- Statistical Word Alignment
- GIZA++
  - A statistical machine translation toolkit used to train word alignments.
  - Uses Expectation-Maximization with various constraints to bootstrap alignments



Phrase Table Extraction



تحمل	carry	<probabilities>
الفنية المعروضات تحمل	artistic pieces carry	<probabilities>
الفنية المعروضات تحمل	the artistic pieces carry	<probabilities>
الفنية المعروضات تحمل	the artistic pieces carry the	<probabilities>
الميزات	characteristics	<probabilities>
الميزات	characteristics of	<probabilities>
الشخصية	personal	<probabilities>
الشخصية	the personal	<probabilities>
الشخصية الميزات	personal characteristics	<probabilities>
الشخصية الميزات	the personal characteristics	<probabilities>
...	...	...

SMT Decoding

- SMT Decoders take
  - Input source-language sentence
  - Translation model: phrase table with probabilities mapping between source and target languages
  - Target language model
  - Tuned weights for the log-linear combination of the various probabilities
- … and they produce a ranked list of translations in the target language
- Popular decoders: Moses (Koehn et al., 2007), cdec (Dyer et al., 2010), Joshua (Li et al., 2009), Portage (Sadat et al, 2005) and others.

10

Automatic Evaluation Metrics

- BLEU (Papineni et al, 2001)
  - *BiLingual Evaluation Understudy*
  - Modified n-gram precision with length penalty
  - Quick, inexpensive and language independent
  - Bias against synonyms and inflectional variations
  - Most commonly used MT metric
  - Official metric of the NIST Open MT Evaluation

Google translate

- Google offers translations between over 3,000 language pairs
  - Statistical MT
  - English Pivoting: Arabic→Hebrew = Arabic →English→Hebrew

Afrikaans	Croatian	Georgian	Italian	Portuguese	Turkish
Albanian	Czech	German	Japanese	Romanian	Ukrainian
Arabic	Danish	Greek	Korean	Russian	Urdu
Armenian	Dutch	Haitian Creole	Latvian	Serbian	Vietnamese
Azerbaijani	English	Hebrew	Lithuanian	Slovak	Welsh
Basque	Estonian	Hindi	Macedonian	Slovenian	Yiddish
Belarusian	Filipino	Hungarian	Malay	Spanish	
Bulgarian	Finnish	Icelandic	Maltese	Swahili	
Catalan	French	Indonesian	Norwegian	Swedish	
Chinese	Galician	Irish	Polish	Thai	

- Introduction
- **Arabic Orthography**
  - **Arabic Script**
  - Phonology and Spelling
  - Encoding Issues
- Arabic Morphology
- Arabic Syntax
- Arabic Dialects

13

Arabic Script

Arabic script is a right-to-left alphabet with allographic variants, optional zero-width diacritics and common ligatures.

الخَطُّ العَرَبِيُّ

Arabic script is used to write many languages: Arabic, Persian, Kurdish, Urdu, Pashto, etc.

14

Arabic Script

Alphabet (MSA)

- letters (form+mark)
  - Distinctive

ش س ث ت ب

/ʃ/ /s/ /θ/ /t/ /b/

- Non-distinctive

ا أ إ آ ى ئ ؤ ء

/ʔ/

glottal stop aka hamza

15

Arabic Script

Letter Shapes

- No distinction between print and handwriting
- No capitalization
- Ambiguous shapes
- Connective letters
- Disconnective letters (ا ؤ و ذ د ز ر ى ة) cause word-internal visual spacing

ز	د	ا	ن	ب	ك	م	ش	غ	**Stand alone**
			ز	بـ	كـ	مـ	شـ	غـ	**initial**
ـز	ـد	ـا	ـن	ـب	ـك	ـم	ـش	ـغ	**medial**
			ن	ب	ك	م	ش	غ	**final**

16

Arabic Script

Letter shaping

كتب = كتب ← ب ت ك

/katab/ b t k

to write

كتاب = كتاب ← ب ا ت ك

/kitāb/ b ā t k

book

17

Arabic Diacritics

Vowel				Nunation			Gemination
بَ	بُ	بِ	بْ	بً	بٌ	بٍ	بّ
/ba/	/bu/	/bi/	/b/	/ban/	/bun/	/bin/	/bb/

- Optional Zero-width characters
  - Full, partial or no diacritics
  - كتب كُتب كُتِبَ /kutiba/ *it was written*
- Combinable: بُّ /bbu/

اسبانيا تنفي تجميد المساعدة الممنوحة للمغرب
مدريد 1 - 11 (اف ب)- اكد رئيس الحكومة الاسبانية خوسيه ماريا اثنار اليوم الخميس ان اسبانيا لم توقف المساعدة التي تقدمها للمغرب خلافا لما اكده امس الاربعاء وزير الشؤون الخارجية والتعاون المغربي محمد بن عيسى امام مجلس النواب المغربي .
وقال رئيس الحكومة الاسبانية في مؤتمر صحافي ان التعاون بين اسبانيا والمغرب لم يتوقف ابدا ولم يجمد.

اِسْبانِيا تَنْفِي تَجْمِيدَ المُساعَدَةَ المَمْنُوحَة لِلمَغْرِبِ
مَدْرِيد 1 - 11 (اِف ب)- اَكَّدَ رَئِيسُ الحُكُومَةِ الإسْبانِيَّةُ خُوسِيه ماريا اثنار اليَوْمَ الخَمِيسَ اَنَّ اِسْبانِيا لِمَ تَوَقَّفَ المُساعَدَةُ الَّتِي تُقَدِّمُها لِلمَغْرِبِ خِلافاً لِما اَكَّدَهُ اَمْسِ الأرْبِعاءَ وَزِيرَ الشُؤُونِ الخارِجِيَّةِ وَالتَعاوُنِ المَغْرِبِيِّ مُحَمَّد بن عِيسَى اَمامَ مَجْلِسِ النُوّابِ المَغْرِبِيِّ .
وَقالَ رَئِيسُ الحُكُومَةِ الإسْبانِيَّةِ فِي مُؤْتَمَرٍ صَحافِيٍّ اَنَّ التَعاوُنَ بَيْنَ اِسْبانِيا وَالمَغْرِبِ لِمَ يَتَوَقَّفْ اَبَداً وَلِمَ يُجَمِّدْ.

19

Arabic Script

Putting it together

Simple combination

Arab /ʕarab/ ع َ ر َ ب ← عَرَب = عرب

West /ʁarb/ غ َ ر ْ ب ← غَرْب = غرب

Ligatures

Peace /salām/ س ل ا م ← سلام سلام

20

Arabic Script

Tatweel

- 'elongation'
- aka kashida
- used for text highlight and justification

حقوق الانسان

حقـوق الانسـان

حقـــوق الانســـان

human rights /ħuqūq alʔinsān/

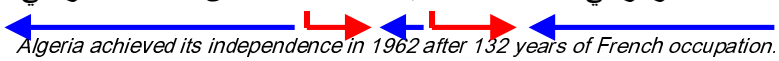
21

Arabic Script

"Arabic" Numerals

- Decimal system
- Numbers written left-to-right in right-to-left text

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.



Algeria achieved its independence in 1962 after 132 years of French occupation.

- Three systems of enumeration symbols that vary by region

Western Arabic *Tunisia, Morocco, etc.*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Indo-Arabic *Middle East*	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Eastern Indo-Arabic *Iran, Pakistan, etc.*	٠	١	٢	٣	۴	۵	۶	٧	٨	٩

- Introduction
- **Arabic Orthography**
  - Arabic Script
  - **Phonology and Spelling**
  - Encoding Issues
- Arabic Morphology
- Arabic Syntax
- Arabic Dialects

23

MSA Phonology and Spelling

- Phonological profile of Standard Arabic
  - 28 Consonants
  - 3 short vowels, 3 long vowels, 2 diphthongs
- Arabic spelling is mostly phonemic …
  - Letter-sound correspondence

ي و ه ن م ل ك ق ف غ ع ظ ط ض ص ش س ز ر ذ د خ ح ج ث ة ت ب ا ى ئ ؤ إ آ أ ء

ī j ū w h n m l k q f ʁ ʕ ð ṭ ḍ ṣ ʃ s z r ð d x ħ ʤ θ t b ā ʔ

24

MSA Phonology and Spelling

Except that …

- Diacritics are optional in most written text
  - 6.8 diacritizations/word
  - Only 1.5% of words have at least one diacritic in news text
- Medial short vowels only appear as diacritics
- Dual use of ا, و, ي as consonant and long vowel
  - ا (/ʾ/,/ā/) و (/w/,/ū/) ي (/y/,/ī/)
- Feminine marker ة (Ta Marbuta)
  - كبير /kabīr/ (big ♂) كبيرة /kabīra/ (big ♀)
- Derivation marker ى (Alif Maqsura)
  - /ʕaṣa/ (to disobey عصى) (a stick عصا)
- Hamza variants (6 characters for one phoneme!)
  - (ء أآإؤئ) بهاء بهاؤه بهائه /baha'/ + 3MascSing (his glory)

25

MSA Phonology and Spelling

- Spelling ambiguity
  - Optional diacritics
    - كاتب: /kātib/ writer , /kātab/ to correspond
  - Suboptimal spelling
    - Hamza dropping: أ, إ → ا
    - Undotted ta-marbuta: ة → ه
    - Undotted final ya: ي → ى
  - Generating the correct Hamza form in context can be hard
- Arabic spelled in other scripts
  - Roman script, Hebrew, etc.
  - Different conventions

26

MSA Phonology and Spelling

- Arabic spelling can be ambiguous
  - optional diacritics and dual use of letter
- But how ambiguous? Really?
- Classic example

  ths s wht n rbc txt lks lk wth n vwls

  this is what an Arabic text looks like with no vowels
- Not exactly true
  - Long vowels are always written
  - Initial vowels are represented by an ا ‘alef’
  - Some final short vowels are represented

  ths is wht an Arbc txt lks lik wth no vwls

Will revisit ambiguity in more detail again under morphology discussion

27

Proper Name Transliteration

- The Qadafi-Schwarzenegger problem
  - Foreign Proper name spelling is often ad hoc
  - Multiplicity of spellings causes increased sparsity

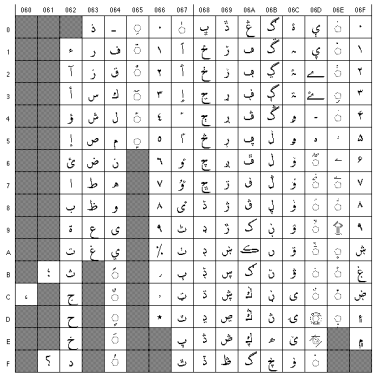
قذافي	→	Gadafi Gaddafi Gaddfi Gadhafi Ghaddafi Kadaffy Qaddafi Qadhafi …
شوارزنيغر شوارزنغر شوارزنيجر شوارتزنجر	←	Schwarzenegger

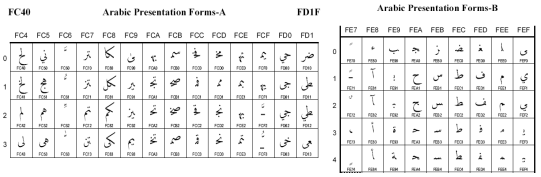
28

- Introduction
- **Arabic Orthography**
  - Arabic Script
  - Phonology and Spelling
  - **Encoding Issues**
- Arabic Morphology
- Arabic Syntax
- Arabic Dialects

Encodings

- Unicode
  - The standard way to encode Arabic today
  - 2-byte characters
  - Widely supported input/display
  - Supports extended Arabic characters
  - Multi-script representation
  - Supports presentation forms (shapes and ligatures)
- Other less common encodings: CP-1256, ISO 8859-6





Encoding Issues
Arabic Display

- Memory (logical order)

ÔÇÑßÊ ÝáÓØíä (Palestine) Ýí ÇæáãÈíÇÏ (Olympics) 2000 æ 2004.

ﺵﺍﺭﻛﺖ ﻓﻠﺴﻄﻴﻦ (Palestine) ﻓﻲ ﺍﻭﻟﻤﺒﻴﺎﺩ (Olympics) 2000 ﻭ 2004.

- Display (visual order)
  - Bidirectional (BiDi) support
    - Numbers and Roman script

ﺵﺍﺭﻛﺖ ﻓﻠﺴﻄﻴﻦ (Palestine) ﻓﻲ ﺍﻭﻟﻤﺒﻴﺎﺩ (Olympics) 2000 ﻭ 2004.

  - Letter and ligature shaping

ﺷﺎﺭﻛﺖ ﻓﻠﺴﻄﻴﻦ (Palestine) ﻓﻲ ﺍﻭﻟﻤﺒﻴﺎﺩ (Olympics) 2000 ﻭ 2004.

31

Encodings

Buckwalter Transliteration

- Romanization
  - One-to-one mapping to Arabic script spelling
  - Left-to-right
  - Easy to learn/use
  - Human & machine compatible
- Commonly used in NLP
  - Penn Arabic Tree Bank
- Some characters can be modified to allow use with XML and regular expressions
- Roman input/display
- Monolingual encoding (can't do English and Arabic)
- Minimal support for extended Arabic characters
- Safe Buckwalter(s)

ء	'	C	ذ	*	V	ل	l	
آ	|	M	ر	r		م	m	
أ	>	O	ز	z		ن	n	
ؤ	&	W	س	s		ه	h	
إ	<	I	ش	$	c	و	w	
ئ	}	Q	ص	S		ى	Y	
ا	A		ض	D		ي	y	
ب	b		ط	T		ً	F	
ة	p		ظ	Z		ٌ	N	
ت	t		ع	E		ٍ	K	
ث	v		غ	g		َ	a	
ج	j		ـ	_		ُ	u	
ح	H		ف	f		ِ	i	
خ	x		ق	q		ّ	~	
د	d		ك	k		ْ	o	

32

Checklist
Arabic Orthography & MT

- Encoding issues
  - Clean up encoding
    - http://www.unicode.org/faq/normalization.html
    - MADA clean-utf8 utility for Arabic
  - Do not mix different encodings!
  - To romanize or not to romanize?
    - If yes, use same romanization scheme!
- Diacritization
  - Most people strip all diacritics to minimize sparsity. See However (Diab et al, 2009; Zbib et al, 2010)

33

Checklist
Arabic Orthography & MT

- Remove Tatweel/Kashida (-)
- Normalization
  - Reductive (internal models; non-Arabic target)
    - Alif forms ا ← آ إ أ **A ← > < |**
    - Ya forms ي ← ى **y ← Y**
    - Ta Marbuta ه ← ة **h ← p**
    - Hamzas ء ← ئ ؤ أ إ آ **‘ ← | < > W & {**
  - Enriching (internal models; Arabic target)
    - Select the appropriate form in context, e.g., using MADA (Habash&Rambow, 2005)
- Use tools for transliteration and spell correction to address OOV, e.g., REMOOV (Habash, 2008)

34

- Introduction
- Arabic Orthography
- **Arabic Morphology**
  - **Arabic Morphology sketch**
  - Morphological Ambiguity
  - MT Tokenization
  - OOV Reduction
- Arabic Syntax
- Arabic Dialects

35

Arabic Morphology Sketch

- Form
  - Concatenative: prefix, suffix, circumfix
  - Templatic: root+pattern
- Function
  - Derivational
    - Creating new words
    - *Mostly templatic*
      - *Exception: +iy~ (Ya of Nisba)*
  - Inflectional
    - Modifying features of words
      - Tense, number, person, mood, aspect
    - *Mostly concatenative*
      - *Notable exceptions include broken plurals*

36

Derivational Morphology

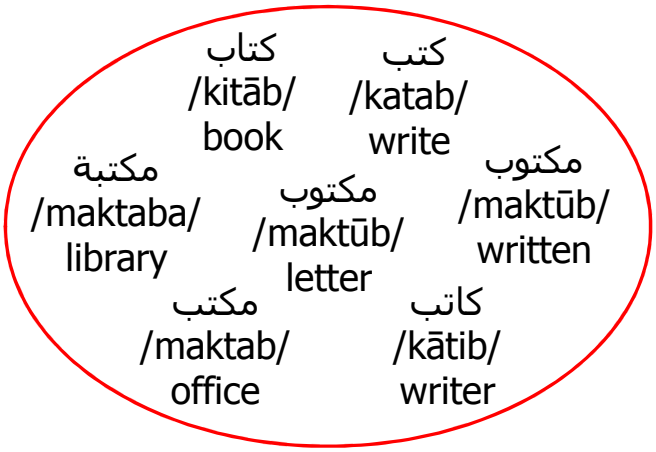
- Templatic Morphology
  - Root: ك ت ب
    k=1 t=2 b=3
  - Pattern: ma12ū3 *passive participle*; 1ā2i3 *active participle*
  - Lexeme: مكتوب maktūb *written*; كاتب kātib *writer*

*Lexeme.Meaning = (Root.Meaning+Pattern.Meaning)*Idiosyncrasy.Random*

37

Derivational Morphology
Root Meaning

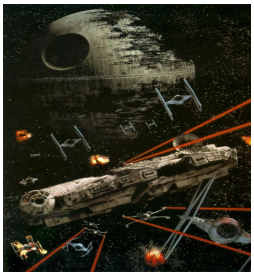
- ك ت ب KTB = notion of *"writing"*



38

Root Polysemy

LHM-1	LHM-2	LHM-3
"meat"	"battle"	"soldering"
/laħm/	/malħama/	/laħam/
Meat	Fierce battle	Weld, solder, stick, cling
/laħħām/	Massacre	
Butcher	Epic	



39

Inflectional Morphology

- Derivational Morphology
  - Lexeme ≈ Root + Base Pattern ≈ Set of all inflectionally related word forms
  - Lemma = Label for Lexeme set
  - Part-of-speech
    - *Traditional*: Noun, Verb, Particle
    - *Computational*: N, PN, V, Adj, Adv, P, Pron, Num, Conj, Det, Aux, Pun, IJ, and others
- Inflectional Morphology
  - Inflected Word = Lexeme + Features
- Cliticization Morphology
  - Word = Inflected Word + Clitics

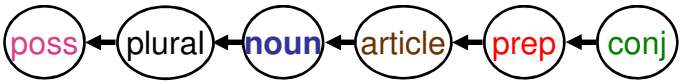
40

Inflectional Morphology

- Inflectional Features
  - Number: singular, dual, plural [Noun,Verb]
  - Gender: masculine, feminine [Noun,Verb]
  - State: definite, indefinite, construct [Noun]
  - Case: nominative, accusative, genitive [Noun]
  - Aspect: perfective, imperfective, imperative [Verb]
  - Voice: active, passive [Verb]
  - Mood: indicative, subjunctive, jussive [Verb]
  - Person: 1, 2, 3 [Verb]
- Clitics (Cliticization Features)
  - Single-letter conjunction proclitic [all]
  - Single-letter preposition proclitic [Noun]
  - Future marker *s+* [Verb]
  - Definite Article *Al+* [Noun]
  - Possessive pronoun enclitic [Noun]
  - Object pronoun enclitic [Verb,Prep]

41

Inflectional Morphology
Nouns

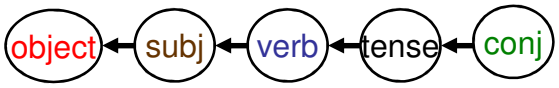


وللمكتبات	وكبيوتنا
/walilmaktabāt/	/wakabiyūtinā/
و+ل+ال+مكتبة+ات	و+ ك + بيوت + نا
wa+li+al+maktaba+āt	wa+ka+biyūt+nā
and+for+the+library+plural	and+like+houses+our
And for the libraries	*And like our houses*

- Morphotactics (e.g. ل+ال → لل)
- Arabic *Broken Plurals* (templatic)

42

Inflectional Morphology
Verbs



وسنقولها
/wasanaqūluhā/
و+ س+ ن+ قول + ها
wa+sa+na+qūl+u+hā
and+will+we+say+it
And we will say it

فقلناها
/faqulnāhā/
ف+ قال+ نا+ ها
fa+qul+na+hā
so+said+we+it
So we said it

- Morphotactics
- Subject conjugation (suffix or circumfix)

43

Inflectional Morphology

- Perfect verb subject conjugation (*suffixes only*)

	Singular	Dual	Plural
1	كتبتُ katabtu	كتبنا katabnā	
2	كتبتَ katabta	كتبتما katabtumā	كتبتم katabtum
3	كتبَ kataba	كتبا katabā	كتبوا katabtū

- Imperfect verb subject conjugation (*prefix+suffix*)

	Singular	Dual	Plural
1	اكتبُ aktubu	نكتبُ naktubu	
2	تكتبُ taktubu	تكتبان taktubān	تكتبون taktubūn
3	يكتبُ yaktubu	يكتبان yaktubān	يكتبون yaktubūn

44

Feminine form and other verb moods not shown

Arabic Morphology Representations

- Natural token وللـمـكتبـــات wl_l_m_kt_b___At
  - White space separated strings (as is)
  - Can include extra characters (e.g. tatweel/kashida)
- Word وللمكتبات wllmktbAt
- Segmented word
  - Can include any degree of morphological analysis
  - Pure segmentation: و ل لمكتبات w l lmktbAt
  - Normalized segmentation
    - Arabic Treebank tokens (with recovery of some deleted/modified letters): و ل المكتبات w l AlmktbAt
  - Tokenization schemes

45

Arabic Morphology Representations

- Prefix + Stem + Suffix
  - وللـ+مكتب+ات wll+mktb+At
  - Can create more ambiguity
- Lexeme + Features
  - مكتبة[+Plural +Def +و+ ل]
- Root + Pattern + Features
  - ك ت ب + a3a21a م + [+Plural +Def +ل +و]
  - Very abstract
- The more complex the representation the harder the disambiguation
- Preprocessing: technique (how) vs scheme (what)

46

- Introduction
- Arabic Orthography
- **Arabic Morphology**
  - Arabic Morphology sketch
  - **Morphological Ambiguity**
  - MT Tokenization
  - OOV Reduction
- Arabic Syntax
- Arabic Dialects

47

Morphological Ambiguity

- Morphological richness
  - Token Arabic/English = 80%
  - Type Arabic/English = 200%
- Morphological ambiguity
  - Each word: 12.3 analyses and 2.7 lemmas
  - Derivational ambiguity
    - قاعدة: basis/principle/rule, military base, Qa'ida/Qaeda/Qaida
  - Inflectional ambiguity
    - تكتب: you write, she writes
    - Segmentation ambiguity
      - وجد: he found; و+جد: and+grandfather
      - للغة: ل+لغة: for a language; ل+اللغة: for the language

48

Morphological Ambiguity

- Spelling ambiguity
  - Optional diacritics
    - كاتب: /kātib/ writer , /kātab/ to correspond
  - Suboptimal spelling
    - Hamza dropping, undotted ta-marbuta, undotted final ya
- Multiple sources of ambiguity

  بين

– /bayyana/	Verb	*he demonstrated*
– /bayyanna/	Verb	*they [feminine] demonstrated*
– /bayyin/	Adj	*clear/evident/explicit*
– /bayna/	Prep	*between/among*
– /biyin/	Proper Noun	*in Yen*
– /biyn/	Proper Noun	*Ben*

49

Morphological Disambiguation *in English*

- Select a morphological tag that fully describes the morphology of a word
- Complete English morphological tag set (Penn Treebank): 48 tags

  Verb:

VB	VBD	VBG	VBN	VBP	VBZ
go	went	going	gone	go	goes

- Same as “POS Tagging” in English

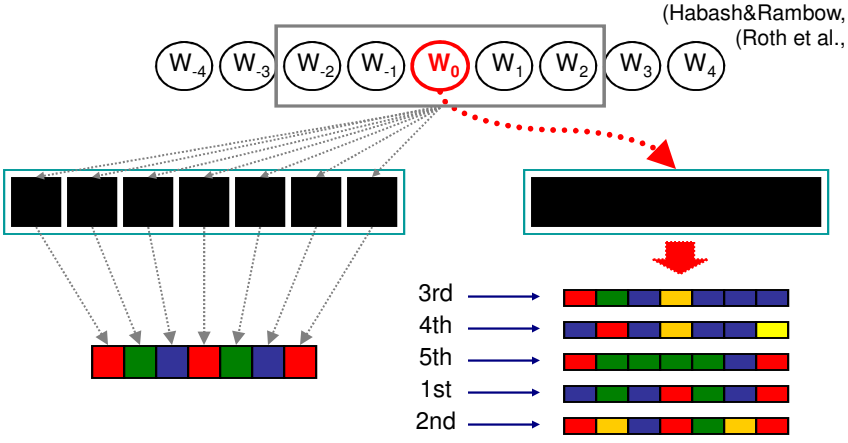
50

Morphological Disambiguation *in Arabic*

- Morphological tag has 14 subtags corresponding to different linguistic categories
  - Example:Verb
    Gender(2), Number(3), Person(3), Aspect(3), Mood(3), Voice(2), Pronominal clitic(12), Conjunction clitic(3)
- 22,400 possible tags
- 2,200 appear in Arabic Tree Bank Part 1 (140K words)
- Example solution: MADA (Habash&Rambow 2005)

51

MADA (Habash&Rambow, 2005)
(Habash&Rambow, 2007)
(Roth et al., 2008)



MORPHOLOGICAL CLASSIFIERS	RANKER	MORPHOLOGICAL ANALYZER
• Multiple independent classifiers • Corpus-trained	• Heuristic or corpus-trained	• Rule-based • Human-created

52

(Habash&Rambow, 2005)
(Habash&Rambow, 2007)
(Roth et al., 2008)

MADA

Morphological Analysis and Disambiguation for Arabic

;;; SENTENCE AsbAnyA ln **twqf** AlmsAEdp Alty tqdmhA **llmgrb**
;;WORD **twqf**
;;SVM_PREDICTIONS: twqf asp:i cas:na enc0:0 gen:f mod:s num:s per:3 pos:verb prc0:0 prc1:0 prc2:0 prc3:0 stt:na vox:a
*0.944308 diac:tuwqifa lex:>awoqaf_1 bw:tu/IV3FS+wqif/IV+a/IVSUFF_MOOD:S
gloss:detain;make_stand pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:3 asp:i vox:a mod:s gen:f num:s stt:na cas:na enc0:0
_0.932187 diac:tawaq~afa lex:tawaq~af_1 bw:+tawaq~af/PV+a/PVSUFF_SUBJ:3MS
gloss:stop;halt;depend_on pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:3 asp:p vox:a mod:i gen:m num:s stt:na cas:na enc0:0
_0.875109 diac:tuwqifa lex:>awoqaf_1 bw:tu/IV2MS+wqif/IV+a/IVSUFF_MOOD:S
gloss:detain;make_stand pos:verb prc3:0 prc2:0 prc1:0 prc0:0 per:2 asp:i vox:a mod:s gen:m num:s stt:na cas:na enc0:0
_0.733784 diac:tawaq~ufi lex:tawaq~uf_1 bw:+tawaq~uf/NOUN+i/CASE_DEF_GEN
gloss:stop;halt pos:noun prc3:0 prc2:0 prc1:0 prc0:0 per:na asp:na vox:na mod:na gen:m num:s stt:c cas:g enc0:0
…

;;WORD **llmgrb**
;;SVM_PREDICTIONS: llmgrb asp:na cas:g enc0:0 gen:m mod:na num:s per:na pos:noun_prop prc0:Al_det
prc1:li_prep prc2:0 prc3:0 stt:d vox:na
*1.000083 diac:lilmagoribi lex:magorib_1 bw:li/PREP+Al/DET+magorib/NOUN_PROP+i/CASE_DEF_GEN
gloss:Morocco pos:noun_prop prc3:0 prc2:0 prc1:li_prep prc0:Al_det per:na asp:na vox:na mod:na gen:m num:s stt:d cas:g enc0:0
_0.965421 diac:lilmagoribi lex:magorib_2 bw:li/PREP+Al/DET+magorib/NOUN+i/CASE_DEF_GEN
gloss:sunset pos:noun prc3:0 prc2:0 prc1:li_prep prc0:Al_det per:na asp:na vox:na mod:na gen:m num:s stt:d cas:g enc0:0
_0.910288 diac:lilmugar~abi lex:mugar~ab_1 bw:li/PREP+Al/DET+mugar~ab/ADJ+i/CASE_DEF_GEN
gloss:exiled pos:adj prc3:0 prc2:0 prc1:li_prep prc0:Al_det per:na asp:na vox:na mod:na gen:m num:s stt:d cas:g enc0:0
…

53

(Habash, 2007)

TOKAN

- A generalized tokenizer
- Assumes disambiguated morphological analysis
  - a la MADA
- Declarative specification of tokenization scheme

```
wsyktbhA=[katab_1 pos:verb prc2:wa_conj prc1:sa_fut prc0:0 per:3
    asp:i vox:a mod:i gen:m num:s stt:na cas:na enc0:3fs_dobj]
```

Example	Scheme	Specification
w+ syktbhA	D1	w+ f+ REST
w+ s+ yktbhA	D2	w+ f+ b+ k+ l+ s+ REST
w+ s+ yktb +hA	D3	w+ f+ b+ k+ l+ s+ Al+ REST +P: +O:
w+ syktb +hA	TB	w+ f+ b+ k+ l+ REST +P: +O:
w+ s+ ktb/VBZ S:3MS +hA	EN	w+ f+ b+ k+ l+ s+ Al+ LEXEME + BIESPOS +S:

- Uses the ALMOR morphological generator (Habash 2004)

54

- Introduction
- Arabic Orthography
- **Arabic Morphology**
  - Arabic Morphology sketch
  - Morphological Ambiguity
  - **MT Tokenization**
  - OOV Reduction
- Arabic Syntax
- Arabic Dialects

55

Arabic, English and MT

	Arabic	English
Orthographic ambiguity	More	Less
Orthographic inconsistency	More	Less
Morphological inflections	More	Less
Morpho-syntactic complexity	More	Less
Word order freedom	More	Less

- We consider next results from Arabic-to-English (Habash&Sadat, 2006) and English-to-Arabic SMT (El Kholy& Habash, 2012)
- We focus on preprocessing/postprocessing
  - schemes & techniques
  - (de)normalization & (de)tokenization

(Habash&Sadat, 2006)

Preprocessing Schemes

- ST Simple Tokenization
- D1 Decliticize CONJ+
- D2 Decliticize CONJ+, PART+
- D3 Decliticize all clitics
- BW Morphological stem and affixes
- EN D3, Lemmatize, English-like POS tags, Subj

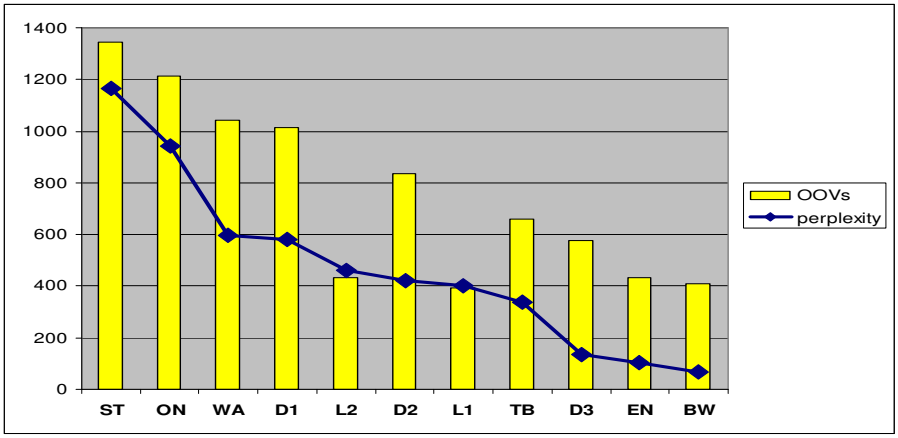
Input:		wsyktbhA?	'and he will write it?'
	ST	wsyktbhA ?	
	D1	w+ syktbhA ?	
	D2	w+ s+ yktbhA ?	
	D3	w+ s+ yktb +hA ?	
	BW	w+ s+ y+ ktb +hA ?	
	EN	w+ s+ ktb/VBZ S:3MS +hA ?	

Preprocessing Schemes

- ST Simple Tokenization
- D1 Decliticize CONJ+
- D2 Decliticize CONJ+, PART+
- D3 Decliticize all clitics
- BW Morphological stem and affixes
- EN D3, Lemmatize, English-like POS tags, Subj
- ON Orthographic Normalization
- WA wa+ decliticization
- TB Arabic Treebank
- L1 Lemmatize, Arabic POS tags
- L2 Lemmatize, English-like POS tags

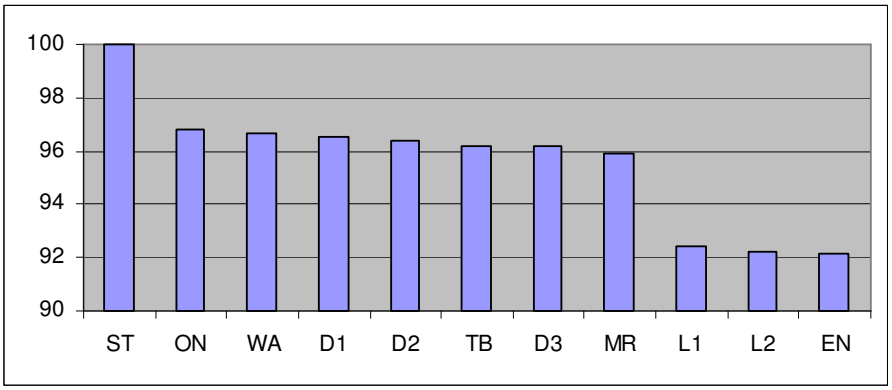
Preprocessing Schemes

- OOVs and Perplexity
  - MT04,1353 sentences, 36000 words



Preprocessing Schemes

- Scheme Accuracy
  - Measured against Penn Arabic Treebank



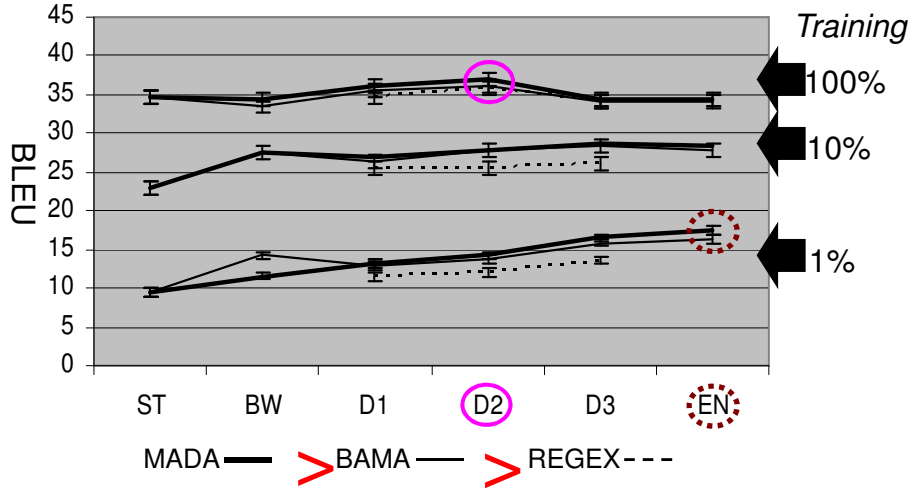
Preprocessing Techniques

- **REGEX**: Regular Expressions
- **BAMA**: Buckwalter Arabic Morphological Analyzer (Buckwalter 2002; 2004)
  - Pick first analysis
  - Use TOKAN (Habash, 2007)
- **MADA**: *Morphological Analysis and Disambiguation for Arabic* (Habash&Rambow, 2005)
  - Multiple SVM classifiers + combiner
  - Selects BAMA analysis
  - Use TOKAN

Experiments

- Portage Phrase-based MT (Sadat et al., 2005)
- Training Data: parallel 5 Million words only
  - All in News genre; Learning curve: 1%, 10% and 100%
- Language Modeling: 250 Million words
- Development Tuning Data: MT03 Eval Set
- Test Data MT04
  - Mixed genre: news, speeches, editorials
- Each experiment
  - Select a preprocessing scheme
  - Select a preprocessing technique
  - Some combinations do not exist, e.g., REGEX and EN

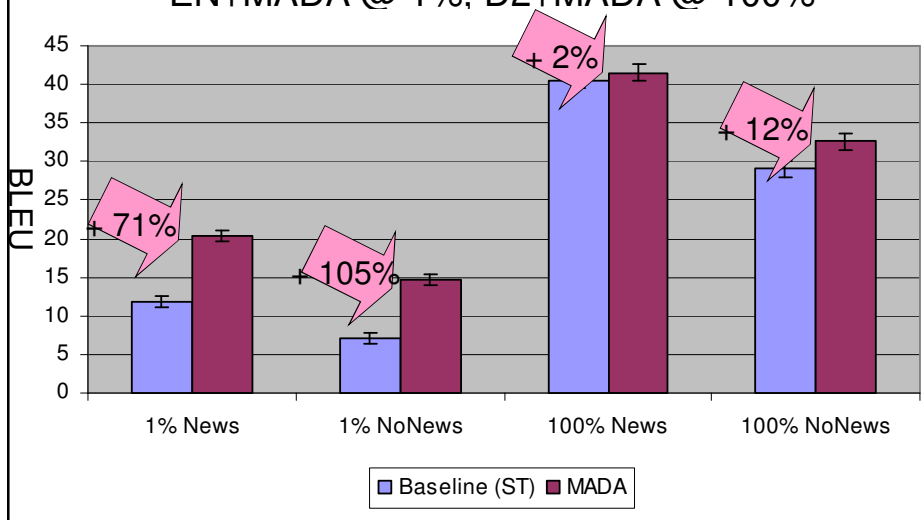
MT04 Results



MT04 Genre Variation

Best Schemes + Technique

EN+MADA @ 1%, D2+MADA @ 100%



Lessons Learned

- For large amounts of training data, splitting off conjunctions and particles performs best
- For small amount of training data, following an English-like tokenization performs best
- Suitable choice of preprocessing scheme and technique yields an important increase in BLEU score if
  - there is little training data
  - there is a change in genre between training and test
- Using MADA+TOKAN provides a framework no more.
- Differences in MT approach, data genre and size require the developers to study the behavior under different settings.
  - For Phrase-based MT, D2/ATB does best; Other approaches do better with D3

Arabic-to-English VS English-to-Arabic

- Arabic-to-English SMT
  - Tokenization and normalization help

    (Lee, 2004; Habash & Sadat, 2006; Zollmann et al., 2006)
- English-to-Arabic SMT
  - What tokenization scheme?

    (Badr et al., 2008; Al Kholy & Habash, 2010; Al-Haj & Lavie, 2010)
  - Output **Detokenization** and **Denormalization**
    - Anything less is comparable to all lower-cased English or uncliticized and undiacritized French

Detokenization

- We compare several detokenization techniques on
  - D3 tokenization scheme (most verbose)
  - Enriched (ENR): Proper Arabic form
  - Reduced (RED): Reductively normalized form
  - Results in Sentence-level Error Rate (SER) (%)

Detokenization	Input & Output form	
	ENR	RED
Simple (S)	38.36	35.53
Rule Based (R)	1.41	3.03
Table Based (T)	1.37	1.54
(T) + back off to (R)	0.79	0.95
(T) + Language Model (LM)	1.2	1.29
(T) + (LM) + back off to (R)	**0.62**	**0.71**

Lesson: Use (T+LM+R) technique for detokenization

Denormalization

- We compare two techniques for denormalization
  - Use MADA (Habash & Rambow, 2005)
  - Jointly detokenize and denormalize (Joint-Detok-ENR)
- Mapping RED tokenized form to ENR untokenized form
  - Results in Sentence-level Error Rate (SER) (%)

Detokenization	Denormalization	
	MADA-ENR	Joint-DeTok-ENR
T+LM+R	7.39	**5.89**

Lesson: Use Joint-DeTok-ENR technique for denormalization

Machine Translation Experiments

End-to-End

- Compare various tokenization and normalization conditions
- Moses Phrase-based MT (Koehn et al., 2008)
- Using our optimal detokenization settings (T+LM+R) and (Joint-DeTok-ENR)
- BLEU scores

Train Data (4M)	ENR	RED
Output	ENR	ENR
D0	24.63	24.66
D1	25.92	26.06
D2	26.41	26.06
TB	**26.46**	**26.73**
S2	25.71	26.11
D3	25.68	25.03

Lesson: Best System is trained on RED text and TB tokenization Scheme

Improving & Scaling Up

- Use MADA 3.0
  - Higher accuracy
- Scaling up by 15 times
- Still can see improvement over baseline
- TB-filtering shows that 25% of the difference between D0 and TB is due to biased filtering

Data	Bleu
D0-4m	26.00
TB-4m	**27.25**
D0-60m	31.30
TB-60m	**32.24**
D0-60m (TB-filtered)	30.98

Lesson: use latest version of MADA

Lesson: Tokenization helps even with larger data sets

Lesson: beware of interactions between different SMT components and tokenization choices

- Introduction
- Arabic Orthography
- **Arabic Morphology**
  - Arabic Morphology sketch
  - Morphological Ambiguity
  - MT Tokenization
  - **OOV Reduction**
- Arabic Syntax
- Arabic Dialects

71

(Habash, 2008)

REMOOV

- Out-Of-Vocabulary (OOV)
  - Test words that are not modeled in training
  - May be in training data but not in phrase table
  - May be in phrase table but not matchable
- A persistent problem
  - Arabic in ATB tokenization with orthographic normalization:

  Increasing the training data by 12 times
  - → 66% reduction in Token/Type OOV
  - → 55% reduction in Sentence OOV (sentences with at least 1 OOV word)

	Medium			**Large**		
Word count	**4.1M**			**47M**		
	MT03	**MT 04**	**MT 05**	**MT03**	**MT 04**	**MT 05**
Token OOV	**2.5%**	**3.2%**	**3.0%**	**0.8%**	**1.1%**	**1.1%**
Type OOV	**8.4%**	**13.32%**	**11.4%**	**2.7%**	**4.6%**	**4.0%**
Sentence OOV	**40.1%**	**54.47%**	**48.3%**	**16.9%**	**25.6%**	**22.8%**

Profile of OOVs in Arabic

- Proper nouns (40%)
  - Different origins: Arabic, Hebrew, English, French, Italian, and Chinese
- Other parts-of-speech (60%)
  - Nouns (26.4%), Verbs (19.3%) and Adjectives (14.3%)
  - Less common morphological forms such as the dual form of a noun or a verb
- Orthogonally, spelling errors appear in (6%) of cases and tokenization errors appear in (7%) of cases

Proper Noun	**40%**	روثبين، جفعاتايم، هوكايدو
Noun/Adjective	**41%**	قريتين، مدرستا
Verb	**19%**	سيلتقيان، تر، مررنا
Spelling Error	**13%**	اشحاض، باكتسان، لروثبين

OOV Reduction Techniques

- Two strategies for online handling of OOVs by phrase table extension
  - Recycle Phrases
    - Expand the phrase table online with recycled phrases
      - Relate OOV word to INV (in-vocabulary) word
      - Copy INV phrases and replace INV word with OOV word
      - Example: add misspelled variant of a word in phrase table
        - » كناب *knAb* → book
      - Using unigram and bigram phrases was optimal for BLEU
  - Novel Phrases
    - Expand the phrase table online with new phrases
      - Example: باستور *bAstwr* is OOV
      - Use transliteration software to produce possible translations
        - » Pasteur, Pastor, Pastory, Bostrom, etc.

REMOOV Techniques

- MorphEx (morphological expansion)
- DictEx (dictionary expansion)
- SpellEx (spelling expansion)
- TransEx (name transliteration)

	Morphology	**No Morphology**
Recycled Phrases	*MorphEx*	*SpellEx*
Novel Phrases	*Dictex*	*TransEx*

REMOOV *Toolkit is available for research*
Contact habash@ccls.columbia.edu

Morphology Expansion

- Model target-irrelevant source morphological variations
  - Cluster Arabic translations of English words
    - book ← (كتاب, الكتاب, كتابا)
    - write ← (... يكتب تكتب نكتب يكتبون يكتبن سيكتبن)
  - Learn mappings of morphological features for words sharing lexemes in the same cluster
    - [POS:V +S:3MS] == [POS:V +S:3FS]
    - [POS:N Al+ +PL] == [POS:N +PL]
    - [POS:N +DU] == [POS:N +PL]
- Map OOV word to INV word using a morphology rule:
- الجماعتين → [POS:N Al+ +DU] == [POS:N +PL] → جماعات

Spelling Expansion

- Relate an OOV word to an INV word through:
  - Letter deletion فلسطني → فلسطيني
  - Letter Insertion فليسطيني → فلسطيني
  - Letter inversion فلسيطني → فلسطيني
  - Letter substitution قلسطيني → فلسطيني
  - Substitution in Arabic was limited to 90 cases (as opposed to 1260)
    - Shape alternations ر <> ز
    - Phonological alternations س <> ص
    - Dialectal variations أ <> ق
- *No modification of the probabilities in the recycled phrases*

Transliteration Expansion

- Use a similarity metric (Freeman et al 2006) to match Arabic spelling to English spelling of proper names
  - Expand forms by mapping to Double Metaphones (Philips, 2000)
- Assign very low probabilities that are adjusted to reflect similarity metric score

المتنبي	→	MTNP	→	Al-Mutannabi Al-Mutanabi
باستور	→	PSTR	→	Pasteur Pastor Pastory Pasturk Bistrot Bostrom
شوارزنغر شوارزنيجر شوارتزنجر	→	XFRTSNKR	→	Schwarzenegger
قذافي	→	KTF	→	Qadhafi Gadafi Gaddafi Kadafi Ghaddafi Qaddafi Katif Qatif

Dictionary Expansion

- OOV word is analyzable by BAMA (Buckwalter 2004)
- Add phrase table entries for OOV translating to all inflected forms of the BAMA English gloss
- Assign equal very low probabilities to all entries

الموسيقيون	→	موسيقي	→	musical	→	musical musicals
			→	musician	→	musician musicians
المخطئة	→	مخطئ	→	mistaken	→	mistaken
			→	at fault	→	at fault at faults
جلستم	→	جلس	→	sit	→	sit sits sat sitting

REMOOV Evaluation

- Medium Set
  - 4.1 M words
  - Average token OOV is 2.9%
- All techniques improve on baseline
  - TransEx < MorphEx < DictEx < SpellEx
- Combinations improve on combined techniques
  - Least improving combination (on average): MorphEx+DictEx
  - Most improving combination (on average): DictEx+TransEx
- Combining all improves most

BLEU Scores

	MT03	MT04	MT05
BASELINE	44.20	40.60	42.86
TRANSEX	44.83	40.90	43.25
MORPHEX	44.79	41.18	43.37
DICTEX	44.88	41.24	43.46
SPELLEX	45.09	41.11	43.47
MORPHEX+DICTEX	45.00	41.38	43.54
SPELLEX+dMORPHEX	45.28	41.40	43.64
SPELLEX+TRANSEX	45.43	41.24	43.75
DICTEX+TRANSEX	45.30	41.43	43.72
ALL	**45.60**	**41.56**	**43.95**
Absolute improvement	*1.4*	*0.96*	*1.09*
Relative improvement	*3.17*	*2.36*	*2.54*

REMOOV Evaluation

- Learning Curve Evaluation
  - Different techniques do better under different size conditions
  - Even with 10 times data, OOV handling techniques still help
- Error Analysis
  - Hardest cases are Names
  - 60% of time, OOV handling is acceptable

MT04 BLEU Scores

	1%	10%	100%	1000%
Baseline	13.40	31.07	40.60	42.06
TransEX	13.80	31.78	40.90	42.10
SpellEX	14.02	31.85	41.11	42.25
MorphEX	15.06	32.29	41.18	42.16
DictEx	**20.09**	**33.56**	41.24	42.14
ALL	18.17	33.41	**41.56**	**42.29**
Best Absolute	6.69	2.49	0.96	0.23
Best Relative	49.93	8.01	2.36	0.55

	PN	NOM	V	
Good	26 (40%)	41 (73%)	17 (85%)	*60%*
Bad	39 (60%)	15 (27%)	3 (15%)	*40%*
	46%	*40%*	*14%*	*100%*

OOV Handling Examples

- Foreign name
  - Before: … and president of ecuador lwt$yw gwtyryz .
  - After: … and president of ecuador lucio gutierrez .
- Dual noun
  - Before: … headed the mission to grytyn in the north .
  - After: … headed the mission to villages in the north .
- Dual verb
  - Before: … baghdad and riyadh , which qTEtA their diplomatic relations …
  - After: … baghdad and riyadh , which sever their diplomatic relations …
- Spelling error
  - Before: … but mHAdtAt between palestinian factions …
  - After: … but talks between palestinian factions …

- Introduction
- Arabic Orthography
- Arabic Morphology
- **Arabic Syntax**
  - **Arabic Syntax Sketch**
  - Verb-Subject Reordering
- Arabic Dialects

83

Sentence Structure

- Verbal sentences
  - Verb agreement with gender only
    - Default singular number
    - كتب الولد\الاولاد wrote$_{3MascSing}$ the-boy/the-boys
    - كتبت البنت\البنات wrote$_{3FemSing}$ the-girl/the-girls
  - Pronominal subjects are conjugated (pro-dropped)
    - كتبتُ wrote-you$_{MascSing}$
    - كتبتم wrote-you$_{MascPlur}$
    - كتبوا wrote-they$_{MascPlur}$
  - Common structural ambiguity
    - Verb$_{3MascSingular}$ Noun$_{Masc}$

      Verb subject=he object=Noun || Verb subject=Noun

84

Sentence Structure

- Copular sentences
  - [Topic Complement]
    Definite Topic, Indefinite Complement
    - الولد شاعر
      the-boy poet
      The boy is a poet
  - [Auxiliary Topic Complement]
    Auxiliaries (*kāna and her sisters*)
    - Tense, Negation, Transformation, Persistence
    - كان الولد شاعرا was the-boy poet *The boy was a poet*
    - ليس الولد شاعرا is-not the-boy poet *The boy is not a poet*
  - Inverted order is expected in certain cases
    - Indefinite topic
      عندي كتاب /ʕindi kitābun/ at-me a-book *I have a book*

85

Sentence Structure

- Copular sentences
  - Types of complements
    - Noun/Adjective/Adverb
      - الولد ذكي the-boy smart *The boy is smart*
    - Prepositional Phrase
      - الولد في المكتبة the-boy in the-library *The boy is in the library*
    - Copular-Sentence
      - الولد كتابه كبير [the-boy [book-his big]] *The boy, his book is big*
    - Verb-Sentence << complex sentences
      - الاولاد كتبوا الاشعار

        [the-boys [wrote$_{3rdMascPlur}$ the-poems]] The boys wrote the poems
      - Full agreement in this order (SVO)
      - الاشعار كتبها الاولاد

        [the-poems [wrote$_{3rdMascSing}$-them the boys]] The poems, the boys wrote

86

Phrase Structure

- Noun Phrase
  - Determiner Noun Adjective PostModifier
    - هذا الكاتب الطموح القادم من اليابان
      this the-writer the-ambitious the-arriving from Japan
      This ambitious writer from Japan
  - Noun-Adjective agreement
    - number, gender, definiteness
      - الكاتبة الطموحة the-writer$_{FemSing}$ the-ambitious$_{FemSing}$
      - الكاتبات الطموحات the-writer$_{FemPlur}$ the-ambitious$_{FemPlur}$
    - Exception: Plural non-persons
      - definiteness agreement; feminine singular default
      - المكتب الجديد the-office$_{MascSing}$ the-new$_{MascSing}$
      - المكتبة الجديدة the-library$_{FemSing}$ the-new$_{FemSing}$
      - المكاتب الجديدة the-offices$_{MascBPlur}$ the-new$_{FemSing}$
      - المكتبات الجديدة the-libraries$_{FemPlur}$ the-new$_{FemSing}$

87

Phrase Structure

- Noun Phrase
  - Idafa construction (اضافة)
    - **Noun1** ***of*** **Noun2** encoded structurally
    - Noun1-indefinite Noun2-definite
    - ملك الاردن
      king Jordan
      the king of Jordan / Jordan's king
  - Noun1 becomes definite
    - Agrees with definite adjectives
  - Idafa chains
    - $N^1_{indef}\ N^2_{indef}\ \ldots\ N^{n-1}_{indef}\ N^n_{def}$
    - ابن عم جار رئيس مجلس ادارة الشركة
      son uncle neighbor chief committee management the-company
      The cousin of the CEO's neighbor

88

Computational Resources

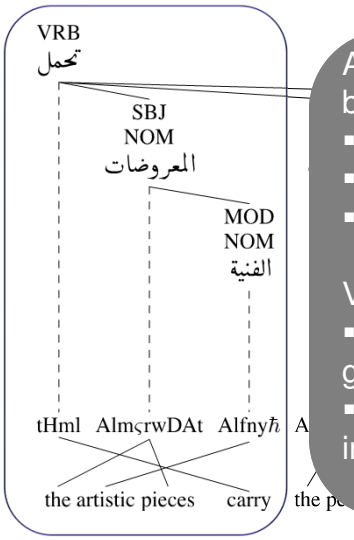
- Multiple Treebanks
  - PATB : Penn Arabic Treebank (Maamouri et al., 2004)
  - PADT : Prague Arabic Dependency Treebank (Smrž et al., 2008)
  - CATIB : Columbia Arabic Treebank (Habash & Roth, 2009)
  - Quran Corpus: Quranic Arabic Treebank (Dukes et al., 2010)
- Parser for Arabic
  - Bikel's parser (Bikel, 2003)
  - Nivre's MALT parser (Nivre et al., 2006)
  - Stanford parser (Green & Manning, 2010)
  - Easy First parser (Goldberg & Elhadad, 2010)
  - Morphological features for Arabic parsing (Marton et al., 2010,2011)
- AMIRA Base-phrase Chunking
  - (Diab et al, 2004; Diab et al. 2007)

89

- Introduction
- Arabic Orthography
- Arabic Morphology
- **Arabic Syntax**
  - Arabic Syntax Sketch
  - **Verb-Subject Reordering**
- Arabic Dialects

90

Why are Arabic verb subjects hard to translate into English?



Arabic verb-subjects might be
- VS post-verbal
- SV pre-verbal
- NS pro-drop

Verbs agree with subject
- in person, number and gender in SV order
- in number and gender only in VS order

(Carpuat et al, 2012)

How should we handle Arabic VS in phrase-based SMT?

- Phase-based SMT
  - state-of-the-art results in Arabic-English benchmarks
  - but not well suited to integrating syntax
- Previous attempts at explicitly modeling Arabic VS were not conclusive
  - Pre-reordering improves AER, but not BLEU (Habash, 2007)
  - Using VS subject boundaries in decoding slightly hurts BLEU (Green et al., 2010)

How are Arabic Verb Subjects reordered in gold translations?

We study gold reorderings in the Arabic-English parallel treebank using

- Gold Arabic parses
- Manual word alignments

Findings:

- **Almost all SV are translated monotonically**
- **27% of VS too, surprisingly!**
- **Matrix VS are inverted twice as much as non-matrix.**

Construction	Gold Reorder	All verbs %	Matrix %	Non Matrix %
SV	monotone	98.2	98.4	98.0
SV	inverted	0.5	0	0.7
SV	overlap	1.3	1.6	1.3
SV	total	100	100	100
VS	monotone	27.3	13.6	40.8
VS	inverted	64.7	81.4	48.1
VS	overlap	8.0	5.0	11.1
VS	total	100	100	100

How well does our parser detect Arabic Verb Subjects?

State-of-the-art Arabic dependency parser

- MaltParser v1.3 (Nivre, 2008)
- Trained on the Penn Arabic Treebank 3 (Maamouri et al., 2004) converted to CATiB format (Habash & Roth, 2009) using an extended tag set (Marton et al. 2010)

Evaluation

- *Combined* detection of verb constructions and their subjects

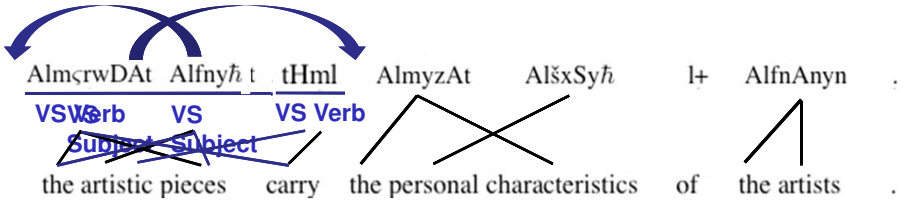
Construction	All	Matrix	Non-Matrix
All (VS+SV+null)	74.11	66.50	75.48
VS	62.81	65.25	57.69
SV	71.68	64.57	72.44
V+null subjects	83.45	69.41	84.27

F-scores

Why are VS harder to identify?

- **VS subject's end boundary**
- **Verb the subject belongs to**

Reordering Arabic VS for alignment



Reordering VS for alignment improves both BLEU and TER on two SMT systems

Evaluation on a large test set:
4432 sentences:
MT03+MT04+MT05+ MT08-nw+MT08-wb

Two systems:

- using Standard Moses SMT system, incl. GIZA++
- medium-scale bitext (12M Arabic words)
- large-scale bitext (64M Arabic words)

System	BLEU r4n4 (%)	TER (%)
Medium baseline	44.35	48.34
+ VS reordering	44.65 (+0.30)	47.78 (-0.56)
+Matrix VS reordering	44.96 (+0.61)	47.52 (-0.82)
Large baseline	51.45	42.45
+ VS reordering	51.70 (+0.25)	42.21 (-0.24)
+Matrix VS reordering	51.80 (+0.35)	42.11 (-0.34)

Lessons Learned

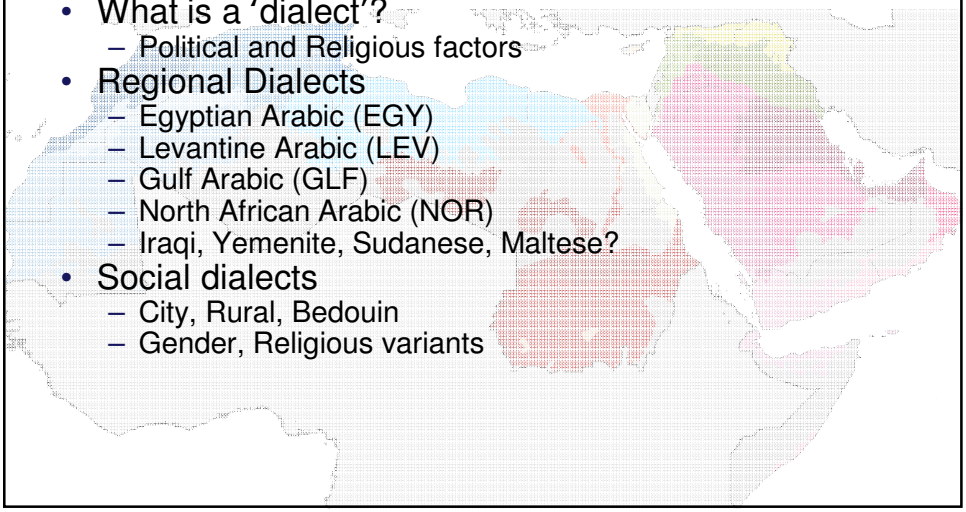
- Reordering patterns of Arabic VS are more ambiguous than expected
  - Only 64.7% of Arabic VS are inverted when translated into English; 81.4% of matrix VS
- Automatic detection of Arabic VS is noisy even with our state-of-the-art parser
  - 62.8% F-score for VS detection; 65.25 % F-score for matrix VS
- Even in these conditions, we can improve Arabic-English SMT by reordering Arabic VS for alignment only
  - Stat. sig. gains in BLEU and TER on medium and large scale baselines

- Introduction
- Arabic Orthography
- Arabic Morphology
- Arabic Syntax
- **Arabic Dialects**
  - **Dialectal Issues**
  - Dialect Resources
  - Dialect MT

98

General Definitions

- Official language: Modern Standard Arabic (MSA)
  - ➢ No one's native language
- What is a 'dialect'?
  - Political and Religious factors
- Regional Dialects
  - Egyptian Arabic (EGY)
  - Levantine Arabic (LEV)
  - Gulf Arabic (GLF)
  - North African Arabic (NOR)
  - Iraqi, Yemenite, Sudanese, Maltese?
- Social dialects
  - City, Rural, Bedouin
  - Gender, Religious variants



Arabic Diglossia

	Formal	Informal
MSA	**Typical** MSA	*Telenovela Arabic MSA L2*
Dialect	Formal Spoken Arabic	**Typical** Dialect

Phonological Variations

- Major variants

MSA		Dialects
ق	/q/	/q/, /k/, /ʔ/, /g/, /ʤ/
ث	/θ/	/θ/, /t/, /s/
ذ	/ð/	/ð/, /d/, /z/
ج	/ʤ/	/ʤ/, /g/

- Some of many limited variants
  - /l/ →/n/ MSA: /burtuqāl/ → LEV: /burtʔān/ 'orange'
  - /ʕ/ → /ħ/ MSA: /kaʕk/ → EGY: /kaħk/ 'cookie'
  - Emphasis add/delete: MSA: /fustān/ → LEV: /fuṣṭān/ 'dress'

101

Non-Standard Letters used in Dialect Areas

	IRQ	LEV	EGY	TUN	MOR
/ʤ/	ج	ج	چ	ج	ج
/g/	گ	چ	ج	ق	ڤ
/ʧ/	چ	تش	تش	تش	تش
/p/	پ	پ	پ	پ	پ
/v/	ڤ	ڤ	ڤ	ڤ	ڤ

- Historical variants: MSA (ف, ق) = MOR (ڤ, ف)

Spelling Inconsistency

في البدايا خلق الله السمّا والأرض. والأرض
كانت خَربانِي وفاضيي وعلى وُشْ الغمق عتمِي وروح
الله يرفرق على وُشْ المويّي. وقال الله خلّي يصير ضَوء
وصار ضوء. وشاف الله الضَّو انُّو شي ظريف وفرَّق
الله بين الضَّوء والعتمِي. وسمَّى الله الضَّوء نهار
والعتمِي سمَّاها ليل وكان مَسا وكان صباح يوم واحد.

وقال الله خلّي يصير جَوْ في وسط المويّي ويصير
فَاصل بين المُويّيْ ومُوّيي. وعمل الله الجَوْ وفرَّق بين
الموَّيّيْ اللّي تحت الجوْ والمُوّيِيْ فوقَ الجَو وهيك صار.
وسَمَّى الله الجَو سماء وكان مساء وكان صباح يوم تاني.

http://www.language-museum.com/a/arabic-north-levantine-spoken.php

Lexical Variation

- Arabic Dialects vary widely lexically

English	Table	Cat	Of	I want	There is	There isn't
MSA	Tāwila طاولة	qiTTa قطة	*idafa* Ø	ʻuridu اريد	yūjadu يوجد	lā yujadu لا يوجد
Moroccan	mida ميدة	qeTTa قطة	dyāl ديال	bγīt بغيت	kāyn كاين	mā kāynš ما كاينش
Egyptian	Tarabēza طربيزة	ʻoTTa قطة	bitāç بتاع	çāwez عاوز	fī في	mafīš مفيش
Syrian	Tāwle طاولة	bisse بسة	tabaç تبع	biddi بدي	fī في	mā fi ما في
Iraqi	mēz ميز	bazzūna بزونة	māl مال	ʻarīd اريد	aku اكو	māku ماكو

- Arabic script allows for consolidation of some variations

Lexical Variation

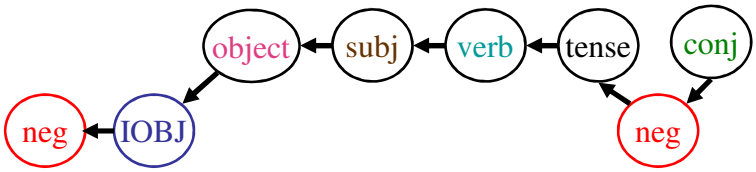
- خلف EGY:reproduce – GLF: give condolences
- مكوى EGY:press iron – GLF:buttocks
- براد EGY:kettle - LEV:fridge
- مرا EGY:prostitute - LEV:woman
- ماشي EGY/LEV:okay – MOR:not
- بسط EGY/LEV:make happy – IRQ:beat up
- العافية EGY/LEV:health – MOR:hell fire
- بلش LEV:start – SUD:end

105

Morphological Variation

- Nouns
  - No case marking
    - Word order implications
  - Paradigm reduction
    - Consolidating masculine & feminine plural
- Verbs
  - Paradigm reduction
    - Loss of dual forms
    - Consolidating masculine & feminine plural (2nd,3rd person)
    - Loss of morphological moods
- Other aspects increase in complexity
  - e.g. additional clitics

Morphological Variation
Verb Morphology



MSA
ولم تكتبوها له
walam taktubūhā lahu
wa+lam taktubū+hā la+hu
and+not_past write_you+it for+him

EGY
وماكتبتوهالوش
wimakatabtuhalūʃ
wi+ma+katab+tu+ha+lū+ʃ
and+not+wrote+you+it+for_him+not

And you didn't write it for him

107

Morphological Variation

	Perfect	***Imperfect***			
	Past	Subjunctive	Present habitual	Present progressive	Future
MSA	كتب /kataba/	يكتب /yaktuba/	يكتب /yaktubu/		سيكتب /sayaktubu/
LEV	كتب /katab/	يكتب /yiktob/	بيكتب /byoktob/	عم بيكتب /ʕam byoktob/	حيكتب /ħayiktob/
EGY	كتب /katab/	يكتب /yiktib/	بيكتب /byiktib/		هيكتب /hayiktib/
IRQ	كتب /kitab/	يكتب /yiktib/	ديكتب /dayiktib/		رح يكتب /raħ yiktib/
MOR	كتب /kteb/	يكتب /yekteb/	كيكتب /kyekteb/		غيكتب /ʁayekteb/

Dialect Switching

MSA and Dialect mixing in speech
- phonology, morphology and syntax

MSA-LIKE LEV

MSA
LEV

لا أنا ما بعتقد لأنه عملية اللي عم بيعارضوا اليوم تمديد للرئيس لحود هم اللي طالبوا بالتمديد للرئيس الهراوي وبالتالي موضوع منه موضوع مبدئي على الأرض أنا بحترم أنه يكون في نظرة ديمقراطية للأمور وأنه يكون في احترام للعبة الديمقراطية وأن يكون في ممارسة ديمقراطية وبعتقد إنه الكل في لبنان أو أكثرية ساحقة في لبنان تريد هذا الموضوع، بس بدي يرجع لحظة على موضوع إنجازات العهد يعني نعم نحكي عن إنجازات العهد لكن هل النظام في لبنان نظام رئاسي النظام في لبنان من بعد الطائف ليس نظام رئاسي وبالتالي السلطة هي عمليا بيد الحكومة مجتمعة والرئيس لحود أثبت خلال ممارسته الأخيرة بأنه لما بيكون في شخص مسؤول في منصب معين وأنا عشت هذا الموضوع شخصيا بممارستي في موضوع الاتصالات لما بياخد مواقف صالحة ضمن خطاب ومبادئ خطاب القسم هو إلى جانبه إنما مش مطلوب من رئيس جمهورية هو يكون رئيس السلطة التنفيذية لأنه منه بقى في لبنان ما بعد إتفاق الطائف رئيس السلطة التنفيذية عليه التوجيه عليه إبداء الملاحظات عليه القول ما هو خطأ وما هو صح عليه تثمير جهود الوطنية الشاملة كي يظل في مصالحة وطنية كي يظل في توافق ما بين المسلم والمسيحي في لبنان يحتضن أبناء هذا البلد ما يترك المسار يروح باتجاه الخطأ نعم إنما خطاب القسم كان موضوع مبادئ طرحت هو ملتزم فيها اللي مشيوا معه وآمنوا فيها التزموا فيها أنا أثبت خلال الأربع سنوات بالممارسة الحكومية أني التزمت فيها ولما التزمنا بهذا الموضوع كان الرئيس لحود إلى جنبنا في هذا الموضوع، أما الموضوع الديمقراطي أنا بتفهم تماما هذا هالوجهة النظر بس ما ممكن نقول إنه الدستور أو تعديله هو أو إمكانية فتح إعادة انتخاب ديمقراطي ضمن المجلس والتصويت إلى ما هنالك لرئيس جمهورية بولاية ثانية هو مسح هيئة في جوهر الديمقراطية هذا بالأقل يعني قناعتي في هذا الموضوع.

109

Aljazeera Transcript http://www.aljazeera.net/programs/op_direction/articles/2004/7/7-23-1.htm

Arabic Dialect Computational Challenges

- General Arabic Challenges
  - Morphologically complex language
    - Rich inflections + clitics
  - Ambiguity due to omitted diacritics
- Dialect-Specific Challenges
  - No standard orthographies
  - Sparse resources (even unannotated corpora)
  - Limited number of morphological analysis and disambiguation tools compared to MSA
  - Substantial Dialect-MSA differences impede direct application of MSA NLP tools

- Introduction
- Arabic Orthography
- Arabic Morphology
- Arabic Syntax
- **Arabic Dialects**
  - Dialectal Issues
  - **Dialect Resources**
  - Dialect MT

111

Dialect Resources

- Lots of resources for Automatic Speech Recognition
  - Speech/transcript corpora (from LDC, ELDA, Appen)
- **BOLT**: Broad Operational Language Translation (DARPA)
  - Egyptian Arabic-to-English MT
  - Supporting the creation of dialect tools and annotations
    - Linguistic Data Consortium
    - Columbia Arabic Dialect Modeling (CADIM) group
- **COLABA**: Cross-lingual Arabic Blog Alerts
  NADIA: NLP tools for Arabic Dialects
  (TSWG: Technical Support Working Group)
  - Supporting the creation of dialect tools and annotations
    - Columbia Arabic Dialect Modeling (CADIM) group
- **TransTac**: Translation System for Tactical Use (DARPA)
  - Iraqi <> English speech-to-speech MT

112

CADIM Tools

(Columbia Arabic Dialect Modeling Group)

- **AIDA 1.4.1:** Automatic Identification of Dialectal Arabic
- **CALIMA 0.4.2:** Columbia Arabic Language Morphological Analyzer -- Egyptian Arabic
- **CODAFY 0.3:** Automatic mapper into the Conventional Orthography of Dialectal Arabic
- **MADA-ARZ 0.3:** Morphological Analysis and Disambiguation for Arabic -- Egyptian Arabic
- **ELISSA 1.0:** Dialectal Arabic to Modern Standard Arabic Translation System
- **DIRA 2.0:** Dialectal Information Retrieval Assistant
- **Tharwa Dictionary:** Egyptian-MSA-English machine readable dictionary (~55K entries)

CODA:
Conventional Orthography for Dialectal Arabic

- Developed by CADIM for computational processing
- Objectives
  - CODA covers all DAs, minimizing differences in choices
  - CODA is easy to learn and produce consistently
  - CODA is intuitive to readers unfamiliar with it
  - CODA uses Arabic script
- Inspired by previous efforts from the LDC and linguistic studies

114

CODA Examples

Phenomenon	Original	CODA
Spelling Errors **Typos** **Speech effects** **Merges** **Splits**	الاجابه شبب كبيييييير اليومبريستيج المع روف	الإجابة سبب كبير اليوم بريستيج المعروف
MSA Root Cognate	هاذا، هاظ هدا، آلب، كلب	هذا قلب
Dialectal Clitic Guidelines	عهلبيت مشفناش	عهالبيت ما شافناش
Unique Dialect Words	بردو، برضو	برضه

115

- Introduction
- Arabic Orthography
- Arabic Morphology
- Arabic Syntax
- **Arabic Dialects**
  - Dialectal Issues
  - Dialect Resources
  - **Dialect MT**

116

Arabic Dialect Machine Translation

- Problems
  - Limited resources
    - Small Dialect-English corpora & no Dialect-MSA corpora
  - Non-standard orthography
  - Morphological complexity
- Solutions
  - Rule-based segmentation (Riesa et al. 2006)
  - Minimally supervised segmentation (Riesa and Yarowsky 2006)
  - Dialect-MSA lexicons (Chiang et al. 2006, Maamouri et al. 2006)
  - Pivoting on MSA (Sawaf 2010, Salloum and Habash, 2011)
    - Elissa 1.0 (Salloum & Habash, 2012)
  - Crowdsourcing Dialect-English corpora (Zbib et al., 2012)

117

Elissa 1.0

- Dialectal Arabic to MSA MT System
- Output
  - MSA top-1 choice, n-best list or map file
- Components
  - Dialectal morphological analyzer (ADAM) (Salloum and Habash, 2011)
  - Hand-written morphological transfer rules & dictionaries
  - MSA language model
- Evaluation (DA-English MT)
  - MADA preprocessing (ATB scheme)
  - Moses trained for MSA-English MT
  - 64 M words training data
  - Best system only processes MT OOVs and ADAM dialect-only words
  - Top-1 choice of MSA
  - Results in BLEU

System	Dev. Set	Blind Test
Baseline	37.20	38.18
Elissa + Baseline	37.86	38.80

Elissa 1.0: DA to MSA translation

Direct Translation of Dialectal Arabic (DA)	
Dialectal Arabic	بهالحالة ماحيكتبولو شي عحيط صفحتو لأنو ماخبرهن يوم اللي وصل عالبلد
DA-English Human Transaltion	In this case, they will not write on his page wall because he did not tell them the day he arrived to the country.
Arabic-English Google Translate	Bhalhalh Mahiketbolo Shi Ahat Cefhto to Anu Mabrhen day who arrived Aalbuld.

Pivoting on Modern Standard Arabic (MSA) using Elissa	
DA-MSA Elissa Translation	في هذه الحالة لن يكتبوا شي علي حائط صفحته لانه لم يخبرهم يوم الذي وصل الي البلد
Arabic-English Google Translate	In this case it would not write something on the wall yet because he did not tell them the day arrived in the country.

Relevant Books

- Soudi, A., S. Vogel, G. Neumann and A. Farghaly, eds. Challenges for Arabic Machine Translation. John Benjamins. 2012.
- Habash, N. and H. Hassan, eds. Machine Translation for Arabic. Special Issue of MT Journal. 2012.
- Habash, N. Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool. 2010.
- Soudi, A., A. van den Bosch, and G. Neumann, eds. Arabic Computational Morphology. Springer, 2007.

120