

A Syntax-based Hebrew-to-Arabic MT System

Reshef Shilon

Linguistics, Tel Aviv University
Computer Science, University of Haifa

ISCOL 2010



MT Frameworks

- ▶ Rule based
 - Linguistically rich rules map SL and TL syntactic constructions
 - Handles morphology and syntax well
 - Doesn't scale up
- ▶ Statistical MT
 - Statistically map SL and TL phrases
 - Scales well
 - Requires large scale parallel corpora

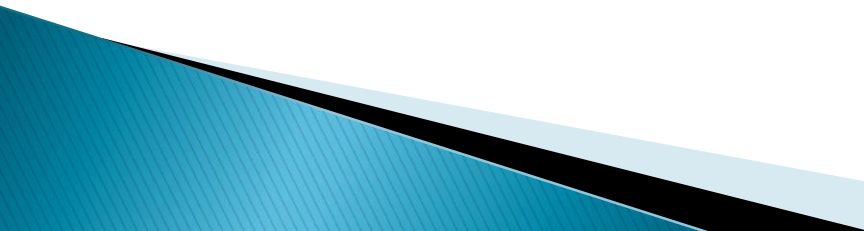
Stat-XFER

- ▶ Lavie et al. 2008
- ▶ Hybrid approach
 - Usage of morphological and syntactic components for analysis and generation
 - Rich formalism enables unification-augmented transfer rules
 - Rules - manually crafted or acquired from corpus
 - Statistical decoding enables scaling
 - Suited for linguistically-rich low-resource language pairs
 - Analysis → Transfer + Generation → Decoding

Hebrew and Arabic – similarities

- ▶ Orthography
 - omission of diacritics in common texts
- ▶ Word formation
 - prefixes (b+, l+) and suffixes (+im, +wn)
- ▶ Inflectional Morphology
 - similar root+template verbal system, similar noun inflection
- ▶ Syntax
 - similar agreement features - V-Subj (gender, person),
 - N-Adj (number, gender, definiteness)
 - Smikhut/Idafa
 - Pro-drop in 1st and 2nd person verbs

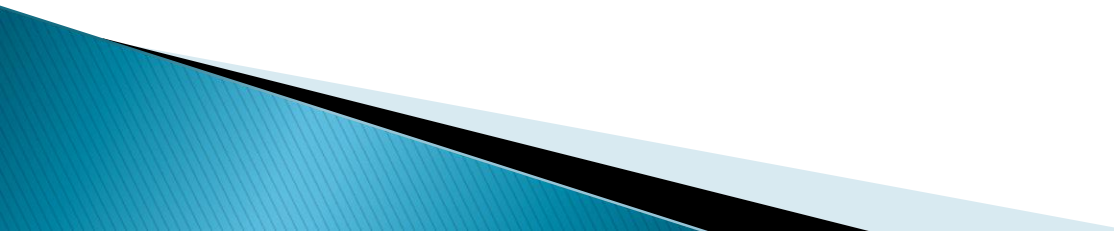
Hebrew and Arabic – differences

- ▶ Word order (SVO vs. VSO)
 - ▶ Arabic has nominal case and verbal mood
 - ▶ Different agreement constraints
 - V-Subj
 - Irrational plural nouns
 - ▶ Constructions that don't exist in the other language
 - Hebrew: shel, At, double genitive
 - Arabic: imperfective subjunctive/jussive, dual number
- 

Challenges

- ▶ Lexical - acquiring gender and rationality for Arabic nouns
- ▶ Morphological
 - Hebrew and Arabic surface forms are highly ambiguous
 - Different tokenization (rAiti Awtm vs. rAyt+hm)
- ▶ Syntactic
 - Mapping and generation of constructions that don't exist in other language
 - Enforcing long distance agreement in Arabic
 - Correctly translating preps according to the verb
- ▶ Computational
 - Lattice explosion due to number of possibilities
 - Some decisions can be taken only at a late stage

Our system

- ▶ Available resources
 - Hebrew morphological analyzer (Itai and Wintner, 2008)
 - Arabic morphological generator (Habash, 2004)
 - Medium coverage bilingual dictionary, no weights
 - Arabic LM based on GigaWord corpus
 - No large parallel corpora
 - ▶ Work in progress
 - ▶ End-to-end
 - ▶ Still not scalable
 - ▶ Current grammar - 42 rules
- 

Google's solution

- ▶ Using English as pivot
- ▶ Adds lexical ambiguity
 - Atm/Atn → Ant (you)
 - Tblh → Tawlp (table)
 - gdh → Albnk (bank)
 - idni → ktyb (manual)
- ▶ Adds syntactic ambiguity
 - mwrwm/mwrwt → mElmyn
 - mwrwt ipwt → AlmElmyn Aljmylp
 - mwrwt ipwt aklw -> Aklt AlmElmyn jmylp

Stat-XFER vs. Google (1)

- (a) *mkwniwt ipwt wšwtrwt ipwt*
car.pl.f.indef pretty.pl.f.indef and+policewomen.f.indef pretty.pl.def.indef
'pretty cars and pretty policewomen'
- (b) *syArAt jmylĥ wšrTyAt jmylAt*
car.pl.f.indef pretty.sg.f and+policewomen.f.indef pretty.pl.f.indef
'pretty cars and pretty policewomen'
- (c) *syArAt AlšrTĥ lTyf lTyf*
car.pl.f.def police.sg.f.def pretty.sg.m.indef pretty.sg.m.indef
'The police's cars pretty pretty'
- ▶ Rational/irrational plural noun
 - ▶ NP conjunction
 - ▶ Lexical translation

Stat-XFER vs. Google (2)

- (a) *hncigim* *šlkm* *nkxw* *bišibh*
representatives.m.def you.pl.m.poss attend.past.3.pl in+meeting.def
'your representatives attended the meeting'
- (b) *HDr* *mmvlwkm* *Aljlsħ*
attend.past.sg.m representatives.m.nom+you.pl.m.poss meeting.def
'your representatives attended the meeting'
- (c) *wmmvlwkm* *AlHADryn* *fy* *AlAjtmAç*
and+representatives.m.nom+you.pl.m.poss attend.ptcp.pl.m.def.acc/gen in meeting.def
'And your representatives that attended the meeting'

- ▶ V-Subj agreement
- ▶ V-Prep matching
- ▶ Case agreement
- ▶ Verb translation

Stat-XFER vs. Google (3)

(a) *mkwnith* *šl* *hmnhlt* *gdwlh*
car.sg.f.def+she.poss of principal.sg.f.def big.sg.f.indef

‘The principal’s car is big’

(b) *syArĥ* *Almdyrĥ* *kbyrĥ*
car.sg.f.indef principal.sg.f.def big.sg.f.indef

‘The principal’s car is big’

(c) *Alf* *AlrAysy* *llsyArAt*
thousand main.sg.m.indef to+the+car.pl.f.def

‘The cars’ thousand main’

- ▶ Double genitive
- ▶ Verbless sentence
- ▶ Definiteness matching

Thank you

