Large-Scale Machine Translation between Arabic and Hebrew:
Available Corpora and Initial Results

Yonatan Belinkov and James Glass

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA

AMTA 2016 Workshop on Semitic Machine Translation
Austin, TX
November 1, 2016

MIT CSAIL

Why Arabic and Hebrew

- Culture and politics

Why Arabic and Hebrew

- Culture and politics
- Linguistic similarities
  - Orthography, Morphology, Syntax, lexicon

Figure from: http://www.lironlavi.com/2012/08/06/aravrit/

Why Arabic and Hebrew

- Culture and politics
- Linguistic similarities
  - Orthography, Morphology, Syntax, lexicon
- Challenges
  - Ambiguous orthography, rich morphology

Figure from: http://www.lironlavi.com/2012/08/06/aravrit/

Previous Work

- Lack of parallel corpora

Previous Work

- Lack of parallel corpora
  - Synchronous CFGs (Shilon+ 2012)
    - ➢ Manually-crafted, not robust
  - Pivoting via English (El Kholy+Habash 2014, 2015)
    - ➢ Under-specification of useful features

Previous Work

- Lack of parallel corpora
  - Synchronous CFGs (Shilon+ 2012)
    - ➢ Manually-crafted, not robust
  - Pivoting via English (El Kholy+Habash 2014, 2015)
    - ➢ Under-specification of useful features

- Tokenization/segmentation

Previous Work

- Lack of parallel corpora
  - Synchronous CFGs (Shilon+ 2012)
    - ➢ Manually-crafted, not robust
  - Pivoting via English (El Kholy+Habash 2014, 2015)
    - ➢ Under-specification of useful features

- Tokenization/segmentation
  - Important for Arabic-English (Badr+ 2008, Habash+Sadat 2006, El Kholy+Habash 2012, Devlin+ 2014, Almahairi+ 2016) and Hebrew-English (Lavie+ 2004, Lembersky+ 2012, Singh+Habash 2012)
  - Used in phrase-based, hybrid, and end-to-end neural MT

Parallel Corpora

Corpus	Sents	Arabic words	Hebrew words	Domain/genre
OpenSubtitles	14.6M	108M	111M	Movies, TV
OpenSubtitles (alt)	9.5M	71M	76M	Movies, TV
WIT^3	0.2M	3.4M	3.1M	TED talks
GNOME	0.6M	2.1M	2.6M	Localization
KDE	80.5K	0.5M	0.4M	Localization
Ubuntu	51.3K	0.2M	0.2M	Localization
Shilon et al.	1.6K	28K	25K	News
Tatoeba	0.9K	90K	0.6M	User-contributed
GlobalVoices	76	3.2K	3.7K	News

Parallel Corpora

Thanks to Mauro!
(Marcello's talk)

Corpus	Sents	Arabic words	Hebrew words	Domain/genre
OpenSubtitles	14.6M	108M	111M	Movies, TV
OpenSubtitles (alt)	9.5M	71M	76M	Movies, TV
WIT^3	0.2M	3.4M	3.1M	TED talks
GNOME	0.6M	2.1M	2.6M	Localization
KDE	80.5K	0.5M	0.4M	Localization
Ubuntu	51.3K	0.2M	0.2M	Localization
Shilon et al.	1.6K	28K	25K	News
Tatoeba	0.9K	90K	0.6M	User-contributed
GlobalVoices	76	3.2K	3.7K	News

The problem with rich morphology

The problem with rich morphology

- Large vocabulary size (and high ambiguity)

The problem with rich morphology

- Large vocabulary size (and high ambiguity)

- Standard approach: tokenize → translate → detokenize
  - Example: “in the house”

بالبيت	Tokenize	ب+ ال+ بيت	Translate	ב+ ה+ בית	Detokenize	בבית
bAlbyt		b+ Al+ byt		b+ h+ byt		bbyt

The problem with rich morphology

- Large vocabulary size (and high ambiguity)

- Standard approach: tokenize → translate → detokenize
  - Example: “in the house”

بالبيت	Tokenize	ب+ ال+ بيت	Translate	ב+ ה+ בית	Detokenize	בבית
bAlbyt		b+ Al+ byt		b+ h+ byt		bbyt

  - Requires language-specific tools
  - Tokenization scheme not tuned for MT

The problem with rich morphology

- Large vocabulary size (and high ambiguity)
- Standard approach: tokenize → translate → detokenize
- Recently, sub-word units in neural models (Sennrich+ 2016, Luong+Manning 2016, Kim+ 2016, …)
  - Byte-pair encoding: unsupervised, pre/post-processing
  - Subnet over characters: supervised, end-to-end (RNN, CNN, …)

Can character-based neural models replace morphology-aware tokenization?

Experimental Setup

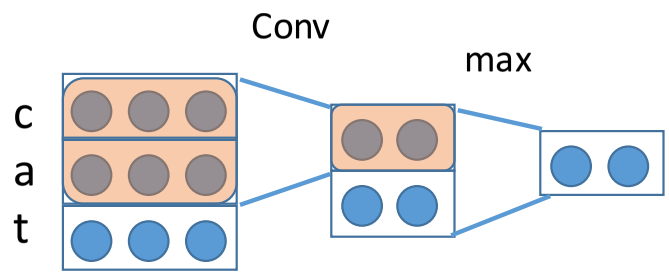
Experimental Setup

- MT systems
  - Moses for phrase-based MT (fast_align, 5-gram LM, MERT)
  - seq2seq-attn for neural MT (LSTM, attention, 50k vocabulary, beam search, Torch impl.)

Experimental Setup

- MT systems
  - Moses for phrase-based MT (fast_align, 5-gram LM, MERT)
  - seq2seq-attn for neural MT (LSTM, attention, 50k vocabulary, beam search, Torch impl.)

- Tokenization and sub-word models
  - MADAMIRA/Farasa for Arabic tokenization (Pasha+ 2014, Abdelali+ 2016)
  - (HTag for Hebrew tokenization (Adler 2009))
  - charCNN (Kim et al. 2016)
    - Replace word vectors with CNN over character vectors



Results (Ar-He)

		BLEU
a	PBMT	9.31
b	PBMT+Tok-Farasa	9.51
c	PBMT+Tok-MADAMIRA	9.63
d	NMT	9.91
e	NMT+UNK replace	10.12
f	NMT+charCNN	10.65
g	NMT+charCNN+UNK replace	10.86

Results (Ar-He)

- Tokenization helps PBMT (b>a, c>a)

		BLEU
a	PBMT	**9.31**
b	PBMT+Tok-Farasa	**9.51**
c	PBMT+Tok-MADAMIRA	**9.63**
d	NMT	9.91
e	NMT+UNK replace	10.12
f	NMT+charCNN	10.65
g	NMT+charCNN+UNK replace	10.86

Results (Ar-He)

- Tokenization helps PBMT (b>a, c>a)
- charCNN helps NMT (f>d)

		BLEU
a	PBMT	9.31
b	PBMT+Tok-Farasa	9.51
c	PBMT+Tok-MADAMIRA	9.63
d	NMT	**9.91**
e	NMT+UNK replace	10.12
f	NMT+charCNN	**10.65**
g	NMT+charCNN+UNK replace	10.86

Results (Ar-He)

- Tokenization helps PBMT (b>a, c>a)
- charCNN helps NMT (f>d)
- Replace UNK gives small boost (e>d, g>f)

		BLEU
a	PBMT	9.31
b	PBMT+Tok-Farasa	9.51
c	PBMT+Tok-MADAMIRA	9.63
d	NMT	**9.91**
e	NMT+UNK replace	**10.12**
f	NMT+charCNN	**10.65**
g	NMT+charCNN+UNK replace	**10.86**

Results (Ar-He)

- Tokenization helps PBMT (b>a, c>a)
- charCNN helps NMT (f>d)
- Replace UNK gives small boost (e>d, g>f)
- Char-based NMT works best (f, g)

		BLEU
a	PBMT	9.31
b	PBMT+Tok-Farasa	9.51
c	PBMT+Tok-MADAMIRA	9.63
d	NMT	9.91
e	NMT+UNK replace	10.12
f	NMT+charCNN	**10.65**
g	NMT+charCNN+UNK replace	**10.86**

More results

- Previous results used only source side, Arabic tokenization
- What about target side, Hebrew tokenization?

More results

- Previous results used only source side, Arabic tokenization
- What about target side, Hebrew tokenization?

- More complicated picture

	PBMT	Word NMT	Char NMT
Arabic Tok	✓	✓	×
Hebrew Tok	×	×	✓
Both	✓	✓	✓

  - Best combination: char-based NMT + Arabic tok + Hebrew tok (11.86 BLEU)

Example Translations

Input	السنة الماضية عرضت هاتين الشريحتين لكي أوضح أن الغطاء الجليدي القطبي ، الذي كان خلال الثلاثة ملايين سنة الماضية في حجم أقله ثمانية وأربعين ، قد تقلص بنسبة أربعين في المائة .
Ref	בשנה שעברה הצגתי את שתי השקופיות האלה שמראות כי כיפת הקרח הארקטי , אשר במשך רוב שלושת מיליון השנים האחרונות היתה שווה לגודלן של 48 המדינות התחתונות , התכווצה בכארבעים אחוזים .
English	Last year I showed these two slides so that demonstrate that the arctic ice cap, which for most of the last three million years has been the size of the lower 48 states, has shrunk by 40 percent.
PBMT	בשנה שעברה את שני الشريحتين כדי להדגים את כיפת הקרח של הקוטב , שהיה במהלך שלושת מיליון השנים האחרונות של אמרתי 48 , קמטי ב @-@ 40 אחוז .
NMT	בשנה שעברה הצגתי את שתי השורות האלה כדי להוכיח <unk> הקרח , שהיה במשך 3 מיליון השנים האחרונות של מה שאמרתי , שמונה <unk> , <unk> ב @-@ 40 אחוז .
NMT+unk	בשנה שעברה הצגתי את שתי השורות האלה כדי להוכיח את הקרח , שהיה במשך 3 מיליון השנים האחרונות של מה שאמרתי , שמונה כרגע , התכווץ ב @-@ 40 אחוז .
NMT+char	בשנה שעברה הראיתי את שתי השקופיות האלה כדי להראות <unk> הקרח , שהיה במהלך שלושת מיליוני השנים האחרונות , בערך ב @-@ 48 אחוז .
NMT+char+unk	בשנה שעברה הראיתי את שתי השקופיות האלה כדי להראות את הקרח , שהיה במהלך שלושת מיליוני השנים האחרונות , בערך ב @-@ 48 אחוז .

Example Translations

Input	ولكن هذا يقلل من خطورة هذه المشكلة بالذات لأنها لا تظهر ***سماكة*** الجليد .
Ref	אבל עובדה זו ממעיטה מחומרתה של הבעיה המסויימת הזאת כי היא אינה מראה את **עובי** הקרח .
English	But this understates the seriousness of this particular problem because it doesn't show the **thickness** of the ice.
PBMT	אבל זה להפחית את הסיכון הבעיה הזו , במיוחד כי לא **בעובי** הקרח .
NMT	אבל זה מפחית את הסיכון הזה מפני שהוא לא מראה **בעובי** של קרח .
NMT+unk	אבל זה מפחית את הסיכון הזה מפני שהוא לא מראה **בעובי** של קרח .
NMT+char	אבל זה מפחית את הסיכון לבעיה הספציפית הזו כי היא לא מראה את **הדגים** של קרח .
NMT+char+unk	אבל זה מפחית את הסיכון לבעיה הספציפית הזו כי היא לא מראה את **הדגים** של קרח .

"fish" "thickness"

Example Translations

Input	غطاء القطب الجليدي ، بمعنى ما ، هو القلب النابض لنظام المناخ العالمي . أنه يتمدد في الشتاء وينكمش في الصيف .
Ref	כיפת הקרח הארקטי היא , במובן מסויים , ליבה הפועם של מערכת מזג האויר הגלובלית . היא מתרחבת בחורף ומתכווצת בקיץ .
English	The arctic ice cap is, in a sense, the beating heart of the global climate system. It expands in winter and contracts in summer.
PBMT	כיסוי לקוטב , במובן מסויים , הוא הלב הפועם של מערכת האקלים הגלובלי . זה מתפשט בחורף وينكمش בקיץ .
NMT	<unk> הקוטב , במובן מסוים , הוא הלב הפועם של מערכת האקלים העולמית . הוא <unk> בחורף <unk> בקיץ .
NMT+unk	כיסוי הקוטב , במובן מסוים , הוא הלב הפועם של מערכת האקלים העולמית . הוא מתפשט בחורף מתארך בקיץ .
NMT+char	הקרח , במובן מסוים , הוא הלב של מערכת האקלים העולמית . הוא גדל בחורף <unk> בקיץ .
NMT+char+unk	הקרח , במובן מסוים , הוא הלב של מערכת האקלים העולמית . הוא גדל בחורף מתארך בקיץ .

Contributions

- Review existing large-scale Arabic-Hebrew corpora
- Evaluate state-of-the-art MT systems on Arabic-Hebrew translation
- Compare tokenization with character-based neural models

Results (Ar-He)

		BLEU	Meteor	PPL
a	PBMT	9.31	32.30	478.4
b	PBMT+Tok-Farasa	9.51	33.38	335.5
c	PBMT+Tok-MADAMIRA	9.63	32.90	342.5
d	NMT	9.91	30.55	2.275
e	NMT+UNK replace	10.12	31.84	2.275
f	NMT+charCNN	10.65	32.43	2.239
g	NMT+charCNN+UNK replace	10.86	33.61	2.239

System	BLEU	Meteor	PPL
PBMT	9.31	32.30	478.4
PBMT+Tok-Farasa	9.51	33.38	335.5
PBMT+Tok-MADAMIRA	**9.63**	32.90	342.5
NMT	9.91	30.55	2.275
NMT Large	9.92	30.46	2.214
NMT+UNK Replace	10.12	31.84	2.275
NMT+charCNN	10.65	32.43	2.239
NMT+charCNN+UNK Repl.	**10.86**	33.61	2.239
NMT+Tok-Farasa	10.36		
NMT+Tok-Farasa+UNK Replace	10.56		
NMT+Tok-MADAMIRA	10.15		
NMT+Tok-MADAMIRA+UNK Replace	10.36		
NMT+charCNN+Tok-Farasa	10.71		
NMT+charCNN+Tok-Farasa+UNK Replace	**10.89**		
NMT+charCNN+Tok-MADAMIRA	10.45		
NMT+charCNN+Tok-MADAMIRA+UNK Replace	10.46		
PBMT+Tok-HTAG	9.23	31.51	
PBMT+Tok-HTAG (reconstruct det)	9.07		
PBMT+Tok-HTAG (reconstruct+separate det)	8.89		
PBMT+Tok-HTAG-Farasa	10.13	32.72	
PBMT+Tok-HTAG-Farasa (reconstruct det)	9.95		
PBMT+Tok-HTAG-Farasa (reconstruct+separate det)	9.61		
PBMT+Tok-HTAG-MADAMIRA	**10.33**	32.82	
PBMT+Tok-HTAG-MADAMIRA (reconstruct det)	10.25		
PBMT+Tok-HTAG-MADAMIRA (reconstruct+separate det)	10.04		
NMT+Tok-HTAG	9.66		
NMT+Tok-HTAG (reconstruct det)	10.07		
NMT+Tok-HTAG (reconstruct+separate det)	10.31		
NMT+Tok-HTAG-Farasa	10.86		
NMT+Tok-HTAG-Farasa (reconstruct det)	10.95		
NMT+Tok-HTAG-Farasa (reconstruct+separate det)	11.26		
NMT+Tok-HTAG-MADAMIRA	10.58		
NMT+Tok-HTAG-MADAMIRA (reconstruct det)	11.11		
NMT+Tok-HTAG-MADAMIRA (reconstruct+separate det)	10.89		
NMT+charCNN+Tok-HTAG	11.43		
NMT+charCNN+Tok-HTAG (reconstruct det)	11.04		
NMT+charCNN+Tok-HTAG (reconstruct+separate det)	11.31		
NMT+charCNN+Tok-HTAG-Farasa	11.04		
NMT+charCNN+Tok-HTAG-Farasa (reconstruct det)	11.03		
NMT+charCNN+Tok-HTAG-Farasa (reconstruct+separate det)	11.14		
NMT+charCNN+Tok-HTAG-MADAMIRA	**11.86**		
NMT+charCNN+Tok-HTAG-MADAMIRA (reconstruct det)	11.63		
NMT+charCNN+Tok-HTAG-MADAMIRA (reconstruct+separate det)	11.10		