The Impact of Arabic Morphological Segmentation on Broad-Scale Phrase-based SMT

Alon Lavie and **Hassan Al-Haj**

Language Technologies Institute
Carnegie Mellon University

With contributions from

Kenneth Heafield, Silja Hildebrand and Michael Denkowski

lti

Carnegie Mellon

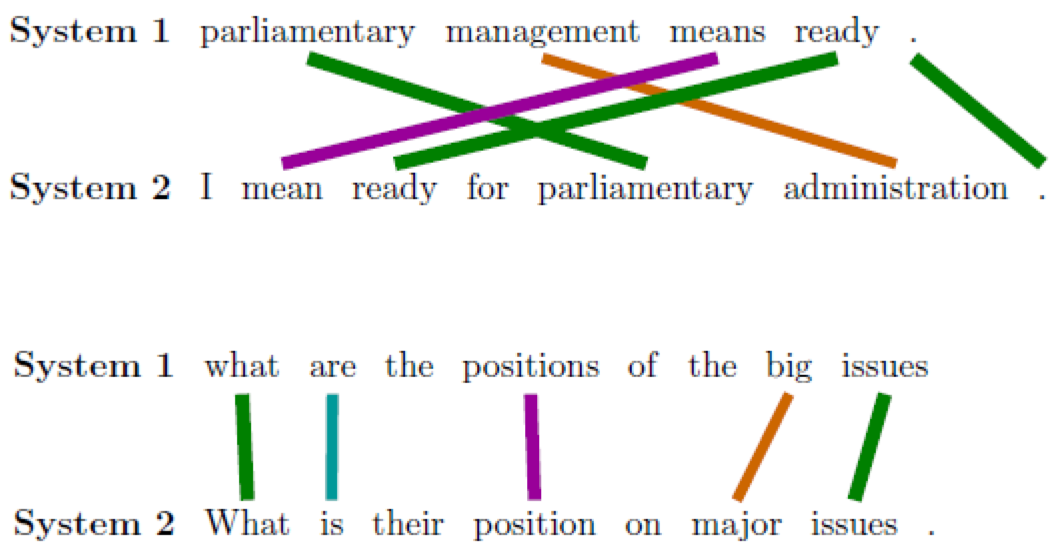
Prelude

- Among the things I work on these days:
  - METEOR
  - MT System Combination (MEMT)
  - Start-Up: Safaba Translation Solutions
- Important Component in all three:
  - METEOR Monolingual Knowledge-Rich Aligner

The METEOR Monolingual Aligner

- Developed as a component in our METEOR Automated MT Evaluation system
- Originally word-based, extended to phrasal matches
- Finds maximal one-to-one alignment match with minimal "crossing branches" (reordering)
- Allows alignment of:
  - Identical words
  - Morphological variants of words (using stemming)
  - Synonymous words (based on WordNet synsets)
  - Single and multi-word Paraphrases (based on statistically-learned and filtered paraphrase tables)
- **Implementation:** efficient search algorithm for best scoring weighted string match

The Monolingual Aligner Examples:

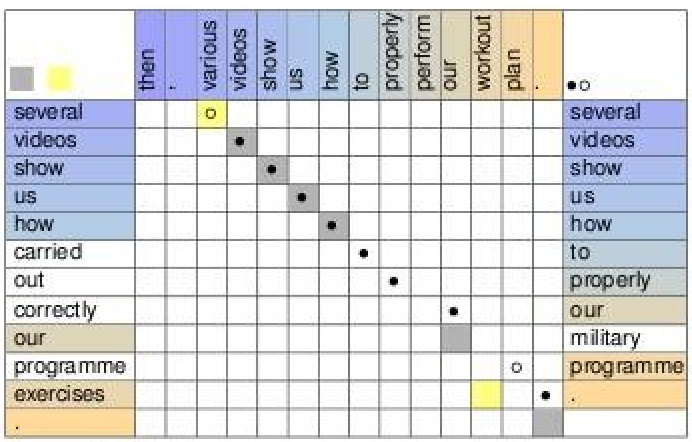


Multi-lingual METEOR

- Latest version METEOR 1.2
- Support for:
  - English: exact/stem/synonyms/paraphrases
  - Spanish, French, German: exact/stem/paraphrases
  - Czech: exact/paraphrases
- METEOR-tuning:
  - Version of METEOR for MT system parameter optimization
  - Preliminary promising results
  - Stay tuned…
- METEOR is free and Open-source:
  - www.cs.cmu.edu/~alavie/METEOR

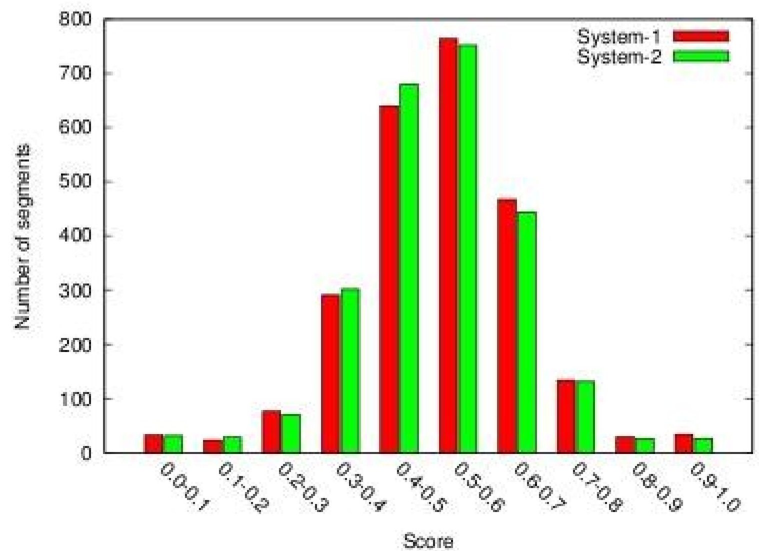
METEOR Analysis Tools

- METEOR v1.2 comes with a suite of new analysis and visualization tools called METEOR-XRAY



Segment 2001

P:	0.633	vs	0.873	:	0.239
R:	0.543	vs	0.686	:	0.143
Frag:	0.231	vs	0.170	:	-0.061
Score:	0.433	vs	0.601	:	0.168



- And now to our Feature Presentation…

Motivation

- Morphological segmentation and tokenization decisions are important in phrase-based SMT
  - Especially for morphologically-rich languages
- Decisions impact the entire pipeline of training and decoding components
- Impact of these decisions is often difficult to predict in advance
- **Goal:** a detailed investigation of this issue in the context of phrase-based SMT between English and Arabic
  - Focus on segmentation/tokenization of the Arabic (not English)
  - Focus on translation from English into Arabic

Research Questions

- Do Arabic segmentation/tokenization decisions make a significant difference even in large training data scenarios?
- English-to-Arabic vs. Arabic-to-English
- What works best and why?
- Additional considerations or impacts when translating into Arabic (due to detokenization)
- Output Variation and Potential for System Combination?

Methodology

- Common large-scale training data scenario (NIST MT 2009 English-Arabic)
- Build a rich spectrum of Arabic segmentation schemes (nine different schemes)
  - Based on common detailed morphological analysis using MADA (Habash et al.)
- Train nine different complete end-to-end English-to-Arabic (and Arabic-to-English) phase-based SMT systems using Moses (Koehn et al.)
- Compare and analyze performance differences

Arabic Morphology

- Rich inflectional morphology with several classes of clitics and affixes that attach to the word
- conj + part + art + base + pron

CONJ	w+ (*and*), f+ (*then*)
PART	l+ (*to/for*), b+ (*by/with*), k+ (*as/such*) s+ *will/future*.
DET	Al+(*the*)
PRON	+h (+O:3MS, +P:3MS) +hA (+O:3FS,+P:3FS) +hm (+O:3MP,+P:3MP) +hmA (+O:3D,+P:3D) +hn (+O:3FP, +P:3FP) +k (+O:2FS,+P:2FS,+O:2MS,+P:2MS) +km (+O:2MP,+P:2MP) +kmA (+O:2D,+P:2D) +kn (+O:2FP,+P:2FP) +nA (+O:1P,+P:1P) +y (+O:1S,+P:1S)

Table 1. Arabic clitics divided to 4 classes.

Arabic Orthography

- Deficient (and sometimes inconsistent) orthography
  - Deletion of short vowels and most diacritics
  - Inconsistent use of أ,آ,إ,ا
  - Inconsistent use of ى , ي
- Common Treatment (Arabic→English)
  - Normalize the inconsistent forms by collapsing them
- Clearly undesirable for MT into Arabic
  - Enrich: use MADA to disambiguate and produce the full form
  - Correct full-forms enforced in training, decoding and evaluation

Arabic Segmentation and Tokenization Schemes

- Based on common morphological analysis by MADA and tokenization byTOKAN (Habash et el.)
- Explored nine schemes (coarse to fine):
  - UT: unsegmented (full enriched form)
  - S0: w + REST
  - S1: w|f + REST
  - S2: w|f + part|art + REST
  - S3: w|f + part/s|art + base + pron-MF
  - S4: w|f + part|art + base + pron-MF
  - S4SF: w|f + part|art + base + pron-SF
  - S5: w|f + part +art + base + pron-MF
  - S5SF: w|f + part + art + base + pron-SF

Arabic Segmentation and Tokenization Schemes

- Based on common morphological analysis by MADA and tokenization byTOKAN (Habash et el.)
- Explored nine schemes (coarse to fine):
  - UT: unsegmented (full enriched form)
  - S0: w + REST
  - S1: w|f + REST
  - S2: w|f + part|art + REST
  - **S3:** w|f + part/s|art + base + **pron-MF**
  - **S4:** w|f + part|art + base + **pron-MF**
  - S4SF: w|f + part|art + base + pron-SF
  - **S5:** w|f + part +art + base + **pron-MF**
  - S5SF: w|f + part + art + base + pron-SF

Morphological Forms!

Arabic Segmentation and Tokenization Schemes

- Based on common morphological analysis by MADA and tokenization byTOKAN (Habash et el.)
- Explored nine schemes (coarse to fine):
  - UT: unsegmented (full enriched form)
  - S0: w + REST
  - S1: w|f + REST
  - S2: w|f + part|art + REST
  - S3: w|f + part/s|art + base + pron-MF
  - S4: w|f + part|art + base + pron-MF
  - **S4SF:** w|f + part|art + base + **pron-SF**
  - S5: w|f + part +art + base + pron-MF
  - **S5SF:** w|f + part + art + base + **pron-SF**

Surface Forms!

Arabic Segmentation and Tokenization Schemes

- Based on common morphological analysis by MADA and tokenization byTOKAN (Habash et el.)
- Explored nine schemes (coarse to fine):
  - UT: unsegmented (full enriched form)
  - S0: w + REST
  - S1: w|f + REST
  - S2: w|f + part|art + REST
  - **S3: w|f + part/s|art + base + pron-MF** **Original PATB**
  - **S4: w|f + part|art + base + pron-MF** **ATBv3**
  - S4SF: w|f + part|art + base + pron-SF
  - S5: w|f + part + art + base + pron-MF
  - S5SF: w|f + part + art + base + pron-SF

Arabic Segmentation Schemes

Input	wbAlnsbp lAyTAlyA fAnh yEny AnhA sttSrf kdwlp Sgyrp ttxlY En ms&wlyAthA
Gloss	and regarding to italy this means that it will act as a country small giving up its responsibilities
English	And regarding Italy, this mean that it will act as a small country giving up its responsibilities
UT	wbAlnsbp l<yTAlyA f>nh yEny >nhA sttSrf kdwlp Sgyrp ttxlY En ms&wlyAthA
S0	w+ bAlnsbp l<yTAlyA f>nh yEny >nhA sttSrf kdwlp Sgyrp ttxlY En ms&wlyAthA
S1	w+ bAlnsbp l<yTAlyA f+ >nh yEny >nhA sttSrf kdwlp Sgyrp ttxlY En ms&wlyAthA
S2	w+ b+ Alnsbp l+ <yTAlyA f+ >nh yEny >nhA s+ ttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAthA
S3	w+ b+ Alnsbp l+ <yTAlyA f+ >n +O:3MS yEny >n +O:3FS sttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAt +P:3FS
S4	w+ b+ Alnsbp l+ <yTAlyA f+ >n +O:3MS yEny >n +O:3FS s+ ttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAt +P:3FS
S5	w+ b+ Al+ nsbp l+ <yTAlyA f+ >n +O:3MS yEny >n +O:3FS s+ ttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAt +P:3FS
S5SF	w+ b+ Al+ nsbp l+ <yTAlyA f+ >n +h yEny >n +hA s+ ttSrf k+ dwlp Sgyrp ttxlY En ms&wlyAt +hA

Table 2. The different tokenization schemes exemplified on the same sentence.

S	Token#	Type #	OOV#
UT	136,280,410	653,584	85
S0	145,826,275	566,024	76
S1	146,162,567	552,150	76
S2	154,974,999	475,335	68
S3	160,194,619	425,645	62
S4	160,599,031	418,832	62
S5	199,179,300	391,190	59

Table 3. tokens, and types count of the Arabic side of the training data for the different schemes and the out-of-vocabulary tokens on NIST MT02 test set.

MT02 Test Set:
- 728 sentences
- 18277 unsegmented words

Previous Work

- Most previous work has looked at these choices in context of Arabic→English MT
  - Most common approach is to use PATB or ATBv3
- (Badr et al. 2006) investigated segmentation impact in the context of English→Arabic
  - Much smaller-scale training data
  - Only a small subset of our schemes

Arabic Detokenization

- English-to-Arabic MT system trained on segmented Arabic forms will decode into segmented Arabic
  - Need to put back together into full form words
  - Non-trivial because mapping isn't simple concatenation and not always one-to-one
  - Detokenization can introduce errors
  - The more segmented the scheme, the more potential errors in detokenization

Arabic Detokenization

- We experimented with several detokenization methods:
  - C: simple concatenation
  - R: List of detokenization rules (Badr et al. 2006)
  - T: Mapping table constructed from training data (with likelihoods)
  - T+C: Table method with backoff to C
  - T+R: Table method with backoff to R
  - T+R+LM: T+R method augmented with a 5-gram LM of full-forms and viterbi search for max likelihood sequence.

Arabic Detokenization

- Evaluation set: 50K sentences (~1.3 million words) from NIST MT 2009 training data
- Rest of NIST MT 2009 training data used to construct mapping table T and train LM
- Evaluated using sentence error rate (SER)

Tok.	C	R	T	T+C	T+R	T+LM+R
S0	3.30	3.37	1.07	0.41	0.48	0.49
S1	4.41	4.48	1.32	0.55	0.60	0.60
S2	36.66	11.30	2.28	1.10	1.09	1. 10
S3	50.26	23.93	3.00	1.76	1.59	1.47
S4	50.59	24.51	3.21	1.94	1.77	1.64
S5	53.52	30.04	3.73	2.40	2.25	1.99
S4SF	50.59	24.51	3.20	1.96	1.79	1.65

Table 6. SER for different tokenization scheme using the six different detokenization scheme.

Experimental Setup

- NIST MT 2009 constrained training parallel-data for Arabic-English:
  - ~5 million sentence-pairs
  - ~150 million unsegmented Arabic words
  - ~172 million unsegmented English words
- Preprocessing:
  - English tokenized using Stanford tokenizer and lower-cased
  - Arabic analyzed by MADA, then tokenized using scripts and TOKAN according to the nine schemes
- Data Filtering: sentence pairs with > 99 tokens on either side or ratio of more than 4-to-1 were filtered out

Tuning and Testing Data

- Use existing NIST MT02, MT03, MT04, MT05 test sets developed for Arabic→English
  - Four English translation references for each Arabic sentence
  - Create English→Arabic sets by selecting First English reference
  - Use MT02 for tuning
  - Use MT03, MT04 and MT05 for testing

	#Sentences	#Tokens	Genres
MT02	728	18277	Newswire
MT03	663	16369	Newswire
MT04	1353	35870	707 Newswire 646Speech/editorial
MT05	1056	28399	Newswire

Table 7. Number of sentences, unsegmented tokens and genres of the tuning and test sets we use.

Training and Testing Setup

- Standard training pipeline using Moses
  - Word Alignment of tokenized data using MGIZA++
  - Symetrized using grow-diag-final-and
  - Phrase extraction with max phrase length 7
  - Lexically conditioned distortion model conditioned on both sides
- Language Model: 5-gram SRI-LM trained on tokenized Arabic-side of parallel data (152 million words)
  - Also trained 7-gram LM for S4 and S5
- Tune: MERT to BLEU-4 on MT-02
- Decode with Moses on MT-03, MT-04 and MT-05
- Detokenized with T+R method
- Scored using BLEU, TER and METEOR on detokenized output

English-to-Arabic Results

System	BLEU	TER	METEOR
UT	35.66	50.76	51.21
S0	**36.25**	**50.98**	**51.60**
S1	35.74	51.47	50.98
S2	*35.05*	*53.16*	*49.81*
S3	36.19	50.49	51.75
S4	**36.22**	**50.61**	**51.58**
S5	34.93	51.77	49.96
S4SF	35.83	50.88	51.48
S5SF	*33.64*	*52.73*	*48.90*
S4,7gram	35.81	50.92	51.26
S5,7gram	34.84	51.88	50.10

MT03

System	BLEU	TER	METEOR
UT	31.53	56.15	45.55
S0	**31.80**	**56.26**	**45.87**
S1	31.46	57.08	45.17
S2	*29.89*	*59.49*	*44.03*
S3	31.73	56.25	45.81
S4	**31.90**	**55.86**	**45.90**
S5	30.87	57.56	44.52
S4SF	31.99	55.90	45.84
S5SF	*30.06*	*57.83*	*43.67*
S4,7gram	31.46	56.04	45.60
S5,7gram	30.91	57.31	44.47

MT04

System	BLEU	TER	METEOR
UT	38.40	47.94	53.96
S0	**38.83**	**48.42**	**54.13**
S1	38.29	48.84	53.40
S2	*37.29*	*51.00*	*52.72*
S3	38.55	48.22	54.33
S4	**38.55**	**48.01**	**54.21**
S5	37.72	49.65	52.94
S4SF	38.15	48.28	54.01
S5SF	*36.80*	*49.91*	*52.00*
S4,7gram	38.32	48.19	54.07
S5,7gram	37.72	49.23	52.81

MT05

Analysis

- Complex picture:
  - Some decompositions help, others don’t help or even hurt performance
- Segmentation decisions really matter – even with large amounts of training data:
  - Difference between best (S0) and worst (S5SF)
    - **On MT03 : +2.6 BLEU, -1.75 TER, +2.7 METEOR points**
- Map Key Reminder:
  - S0: w+REST, S2: conj+part|art+REST, S4: (ATBv3) split all except for the art, S5: split everything (pron in morph. form)
- S0 and S4 consistently perform the best, are about equal
- S2 and S5 consistently perform the worst
- S4SF and S5SF usually worse than S4 and S5

Analysis

- Simple decomposition S0 (just the "w" conj) works as well as any deeper decomposition
- S4 (ATBv3) works well also for MT into Arabic
- Decomposing the Arabic definite article consistently hurts performance
- Decomposing the prefix particles sometimes hurts
- Decomposing the pronominal suffixes (MF or SF) consistently helps performance
- 7-gram LM does not appear to help compensate for fragmented S4 and S5

Analysis: Phrase Tables

Scheme	#Phrase Pairs	#Source Phrases`	PTE	ANTP1	ANTP2	ANTP3	ANTP4	ANTP5	ANTP6	ANTP7
UT	15 ,111,038	29,678	3.411	3317.58	436.15	98.15	41.62	18.69	7.68	5.71
S0	15,575,350	29,870	3.371	3483.66	434.05	95.48	40.43	17.62	7.32	5.43
S1	15,641,938	29,849	3.372	3498.44	435.38	96.73	40.46	17.54	7.38	5.43
S2	16,180,001	29,983	3.332	3674.34	439.06	95.01	39.39	17.46	6.93	4.95
S0PRON	16,489,620	29,896	3.402	3705.43	455.44	99.93	41.41	18.15	7.20	5.57
S3	16,906,278	29,971	3.367	3847.85	455.76	98.37	40.82	17.83	7.03	5.45
S3T	16,910,558	29,949	3.364	3842.83	458.02	98.31	40.80	17.89	7.20	5.13
S4	16,937,625	29,984	3.363	3856.77	455.86	98.47	40.95	17.90	7.12	5.26
S4SF	16,923,937	30,008	3.361	3849.77	457.36	98.62	41.07	17.76	6.92	5.01
S5SFT	20,273,498	29,266	3.611	4776.88	517.70	103.68	40.26	16.63	5.59	3.88
S5SF	20,580,967	29,080	3.634	4877.90	521.23	103.68	39.42	15.86	5.34	3.82
S5	20,596,688	29,045	3.635	4883.26	520.62	103.13	39.68	16.16	5.29	3.69

Table 13. All the features calculated for the different phrase tables of the various segmentation schemes.

- Phrase table filtered to MT03 test set (source side matches)
- PTE = Phrase Table Entropy
- ANTPn = average number of translations for source phrases of length n

Analysis

- Clear evidence that splitting off the Arabic definite article is bad for English→Arabic
  - S4→S5 results in 22% increase in PT size
  - Significant increase in translation ambiguity for short phrases
  - Inhibits extraction of some longer phrases
  - Allows ungrammatical phrases to be generated:
    - Middle East → **Al**$rq **Al**>wsT
    - Middle East → $rq >qsT
    - Middle East → $rq **Al**>wsT

Output Variation

- How different are the translation outputs from these MT system variants?
  - Upper-bound: Oracle Combination on the single-best hypotheses from the different systems
    - Select the best scoring output from the nine variants (based on posterior scoring against the reference)
  - Work in Progress - actual system combination:
    - Hypothesis Selection
    - CMU Multi-Engine MT approach
    - MBR

Oracle Combination

MT03

System	BLEU	TER	METEOR
Best Ind. (S0)	36.25	50.98	51.60
Oracle Combination	**41.98**	**44.59**	**58.36**

MT04

System	BLEU	TER	METEOR
Best Ind. (S4)	31.90	55.86	45.90
Oracle Combination	**37.38**	**50.34**	**52.61**

MT05

System	BLEU	TER	METEOR
Best Ind. (S0)	38.83	48.42	54.13
Oracle Combination	**45.20**	**42.14**	**61.24**

Output Variation

- Oracle gains of 5-7 BLEU points from selecting among nine variant hypotheses
  - Very significant variation in output!
  - Better than what we typically see from oracle selections over large n-best lists (for n=1000)

Arabic-to-English

- Running similar set of experiments in the Arabic→English direction
  - Use all four English references for Tuning and testing
  - Single same English LM for all systems
- Intuitive prediction on magnitude of differences between systems?
  - Smaller, same, or larger?

Arabic-to-English Results

	BLEU	**TER**	**METEOR**
UT	49.55	42.82	72.72
S0	49.27	43.23	72.26
S1	*49.17*	*43.03*	*72.37*
S2	49.97	42.82	73.15
S3	49.15	43.16	72.49
S4	49.70	42.87	72.99
S5	**50.61**	**43.17**	**73.16**
S4SF	49.60	43.53	72.57
S5SF	49.91	43.00	72.62

MT03

Analysis

- Results are preliminary
- Still some significant differences between the system variants
  - Less pronounced than for English→Arabic
- Segmentation schemes that work best are different than in the English→Arabic direction
- S4 (ATBv3) works well, but isn't the best
- More fragmented segmentations appear to work better
- Segmenting the Arabic definite article is no longer a problem
  - S5 works well now
- We can leverage from the output variation
  - Preliminary hypothesis selection experiments show nice gains

Conclusions

- Arabic segmentation schemes has a significant impact on system performance, even in very large training data settings
  - Differences of 1.8-2.6 BLEU between system variants
- Complex picture of which morphological segmentations are helpful and which hurt performance
  - Picture is different in the two translation directions
  - Simple schemes work well for English→Arabic, less so for Arabic→English
  - Splitting off Arabic definite article hurts for English→Arabic
- Significant variation in the output of the system variants can be leveraged for system combination

Current and Future Work

- System combination experiments
  - Hypothesis selection, MEMT and MBR
  - Contrast with lattice decoding (Dyer, 2008) and combining phrase-tables
- Arabic-to-English Experiments
- Better way to do this for other languages?

References

- Al-Haj, H. and A. Lavie. "The Impact of Arabic Morphological Segmentation on Broad-coverage English-to-Arabic Statistical Machine Translation". In Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA-2010), Denver, Colorado, November 2010.
- Al-Haj, H. and A. Lavie. "The Impact of Arabic Morphological Segmentation on Broad-coverage English-to-Arabic Statistical Machine Translation". MT Journal Special Issue on Arabic MT. Under review.