HKUST Statistical Machine Translation Experiments for IWSLT 2007

Yihai SHEN   Chi-kiu LO   Marine CARPUAT   Dekai WU

HKUST
Human Language Technology Center
Department of Computer Science
University of Science and Technology
Hong Kong

{shenyh, jackielo, marine, dekai}@cs.ust.hk

The HKUST submission
Goals for our second IWSLT participation

- Experiment with the open-source Moses decoder, focusing primarily on Chinese-English text translation
  - on various data sets and input conditions
    - Chinese-English text translation task
    - Challenge task on spontaneous speech cancelled by organizers
  - on various language pairs from different language families
    - Arabic-English, Chinese-English, Italian-English, Japanese-English
- Systematically compare Moses against the closed-source Pharaoh decoder
  - used by HKUST for IWSLT-2006

The HKUST submission
Secondary goals for contrastive experiments

- Obtain preliminary indications on performance with…
  - **(semantics)** integration of our recent WSD-for-SMT model [Carpuat & Wu 2007] with Moses (not Pharoah)
  - **(syntax)** our BITG decoder [Wu 1996] substituted for Moses

… while holding all else constant

Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder

System description
Experiments using several SMT decoders

- Decoders
  - Pharaoh [Koehn 2004]
  - Moses [Koehn 2007]
  - Moses [Koehn 2007] + WSD-for-SMT [Carpuat & Wu 2007]
  - Bracketing ITG [Wu 1996]

- Common assumptions of the controlled experiments
  - Phrasal bilexicon
  - Log-linear model
  - Phrases/words represented using surface forms only
    - did not use Moses' factored representation option

System description
Common phrasal bilexicons used

- Learned from bidirectional IBM4 word alignments
  - produced by GIZA++ [Och & Ney 2002]
- Base features used [Koehn 2003]:
  - conditional translation probabilities in both directions
  - lexical weights derived from word translation probabilities
- Allowed phrase lengths up to 20 words
  - short sentences in a well-defined domain

System description
Common phrasal bilexicons used

- Compared two phrase extraction methods:
  - intersection
    - uses strict intersection of bidirectional word alignments
  - grow-diag-final
    - expands alignment by adding directly neighboring alignment points in diagonal neighborhood

- grow-diag-final produced better BLEU scores
  - typically around 0.5 points higher

System description
Language model

- Standard n-gram language models
  - trained using SRI LM toolkit [Stolcke 2002]
- Chinese-English: mixture*
  - 4-gram LM trained on BTEC English
  - 3-gram LM trained on English Gigaword
- Arabic-English, Italian-English, Japanese-English:
  - 3-gram LM trained on BTEC English
- Same LMs used for all experiments*

*except that BITG decoding used only a 3-gram LM trained on BTEC English

Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder

Experimental setup
IWSLT tasks

- Chinese-English text translation only
  - Challenge task (correct recognition vs. read speech vs. spontaneous speech) was cancelled by the organizers
- Text and read speech translation
  - Arabic-English
  - Italian-English
  - Japanese-English

Experimental setup
Minimal language-specific preprocessing

- **English** data was tokenized and case-normalized
- **Italian** data was processed as if it were English
- **Chinese** data was word segmented using LDC segmenter
- **Japanese** data was used directly as provided
- **Arabic**
  - Converted to Buckwalter romanization scheme
  - Tokenized with ASVMT Morphological Analysis toolkit [Diab 2005]

Experimental setup
Improving the sentence segmentation

- The original sentence segmentation is not optimal for training
- Re-segmenting the sentences consistently improves BLEU score

IWSLT-07 data set	# sentences	# sentences after resegmentation	BLEU with original sentences	BLEU after resegmentation
CE devtest1	506	546	41.09	**42.05**
CE devtest2	500	543	42.43	**43.76**
CE devtest2	506	558	51.86	**53.51**

Experimental setup
Training corpus statistics

- Corpora for Chinese and Japanese are twice as large as for Arabic and Italian
- The English side of corpus for Arabic and Italian is a subset

Training data statistics	Chinese-English	Arabic-English	Italian-English	Japanese-English
Number of bisentences	39,953	19,972	19,972	39,953
Vocabulary size (input language)	11,178	25,152	17,917	12,535
Vocabulary size (English output)	18,992	13,337	13,337	18,992

Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder

Results
Official (buggy) results

- Submitted runs were buggy
  (arising from accidental errors in combining models and parameters)

IWSLT07 task	Clear Transcription	ASR Output
Chinese-English	34.26	N/A
Arabic-English	19.51	14.20
Italian-English	17.02	17.02
Japanese-English	40.51	32.49

- Chinese-English: 34.26
  (range among 9 primary submissions: 19.34 - 40.77)

Results
Updated results after removing bugs

IWSLT07 data set	BLEU buggy submitted	BLEU	NIST	METEOR	METEOR no synonyms	TER	WER	PER	CDER
CE devtest1 (buggy)		45.49	7.78	66.11	64.50	36.13	41.68	36.25	**37.10**
CE devtest1		**46.23**	**8.00**	**68.01**	**66.41**	**36.18**	**41.35**	**36.12**	37.14
CE devtest2 (buggy)		48.23	8.32	68.98	67.22	34.99	40.78	34.45	35.43
CE devtest2		**49.77**	**8.82**	**71.88**	**69.85**	**34.47**	**40.12**	**33.41**	**34.58**
CE devtest3 (buggy)		56.44	9.26	76.57	74.47	29.40	34.16	28.86	33.02
CE devtest3		**58.29**	**9.61**	**78.48**	**76.28**	**28.29**	**32.76**	**27.62**	**29.15**
CE test (buggy)	34.26	34.04	6.18	58.28	56.50	45.53	49.15	44.17	41.53
CE test		**35.12**	**6.51**	**60.47**	**58.57**	**44.89**	**48.30**	**43.40**	**41.50**

Results
Updated results after removing bugs

IWSLT07 data set	BLEU buggy submitted	BLEU	NIST	METEOR	METEOR no synonyms	TER	WER	PER	CDER
CE devtest1 (buggy)		45.49	7.78	66.11	64.50	36.13	41.68	36.25	**37.10**
CE devtest1		**46.23**	**8.00**	**68.01**	**66.41**	**36.18**	**41.35**	**36.12**	37.14
CE devtest2 (buggy)		48.23	8.32	68.98	67.22	34.99	40.78	34.45	35.43
CE devtest2		**49.77**	**8.82**	**71.88**	**69.85**	**34.47**	**40.12**	**33.41**	**34.58**
CE devtest3 (buggy)		56.44	9.26	76.57	74.47	29.40	34.16	28.86	33.02
CE devtest3		**58.29**	**9.61**	**78.48**	**76.28**	**28.29**	**32.76**	**27.62**	**29.15**
CE test (buggy)	34.26	34.04	6.18	58.28	56.50	45.53	49.15	44.17	41.53
CE test		**35.12**	**6.51**	**60.47**	**58.57**	**44.89**	**48.30**	**43.40**	**41.50**

our own scoring tools give lower BLEU scores than the official IWSLT scoring

Results
Moses almost always outperforms Pharoah

- Varied many settings and pre-/post-processing steps (bilexicons, LMs, …) to obtain experimental runs under many conditions

Run No.	Pharaoh	Moses
1	41.14	**41.17**
2	41.65	41.70
3	42.05	42.16
4	43.40	**43.55**
5	41.92	**42.26**
6	42.80	**43.19**
7	43.76	**44.28**
8	44.17	**44.64**
9	51.64	**52.19**
10	52.15	**52.59**
11	53.51	**53.64**
12	**53.87**	53.53

Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder

Contrastive experiments (semantics)
Phrase Sense Disambiguation: WSD for SMT

- Today's SMT makes little use of source-language context
- In contrast, WSD approaches generalize across rich contextual features to assign **context-dependent** probabilities to senses

- Earlier negative results: [Carpuat & Wu 2005]
  - Surprisingly, Senseval WSD models do not help translation quality when integrated into a word-based SMT model

- New: Using **PSD**, we repurpose the WSD models for SMT in our newer fully phrasal model: [Carpuat & Wu EMNLP, MT-Summit, TMI 2007]
  - Words are phrasal, just as in traditional lexicography
  - WSD "senses" are exactly same as SMT translation candidates
  - WSD training data is exactly same as SMT training data
  - WSD scores are added to log linear model feature set
  - Feature engineering is <u>exactly</u> inherited from Senseval WSD models

Contrastive experiments (semantics) The HKUST WSD System

- Proved highly effective at Senseval-3
  - Placed first on Chinese lexical sample
  - Placed second on Multilingual lexical sample (translation)
  - 71.4% on English lexical sample (median 67.2, best 72.9)

- Classifier ensemble:
  - naïve Bayes [Yarowsky & Florian 2002]
  - maximum entropy [Klein & Manning 2002]
  - boosting [Carreras *et al.* 2002; Wu *et al.* 2002]: we use boosted decision stumps
  - Kernel PCA model [Wu *et al.* 2004]

Contrastive experiments (semantics) Contextual features in HKUST WSD system

- Feature set includes:
  - Bag-of-words context
  - Position sensitive local collocational features
  - Syntactic features

- A WSD model using these features yielded the best classification accuracy in Yarowsky & Florian [2002]

Contrastive experiments (semantics) PSD improved Moses… just like Pharoah

- Encouraging preliminary indication
- Consistent with our larger EMNLP-CoNLL results [Carpuat & Wu 2007]

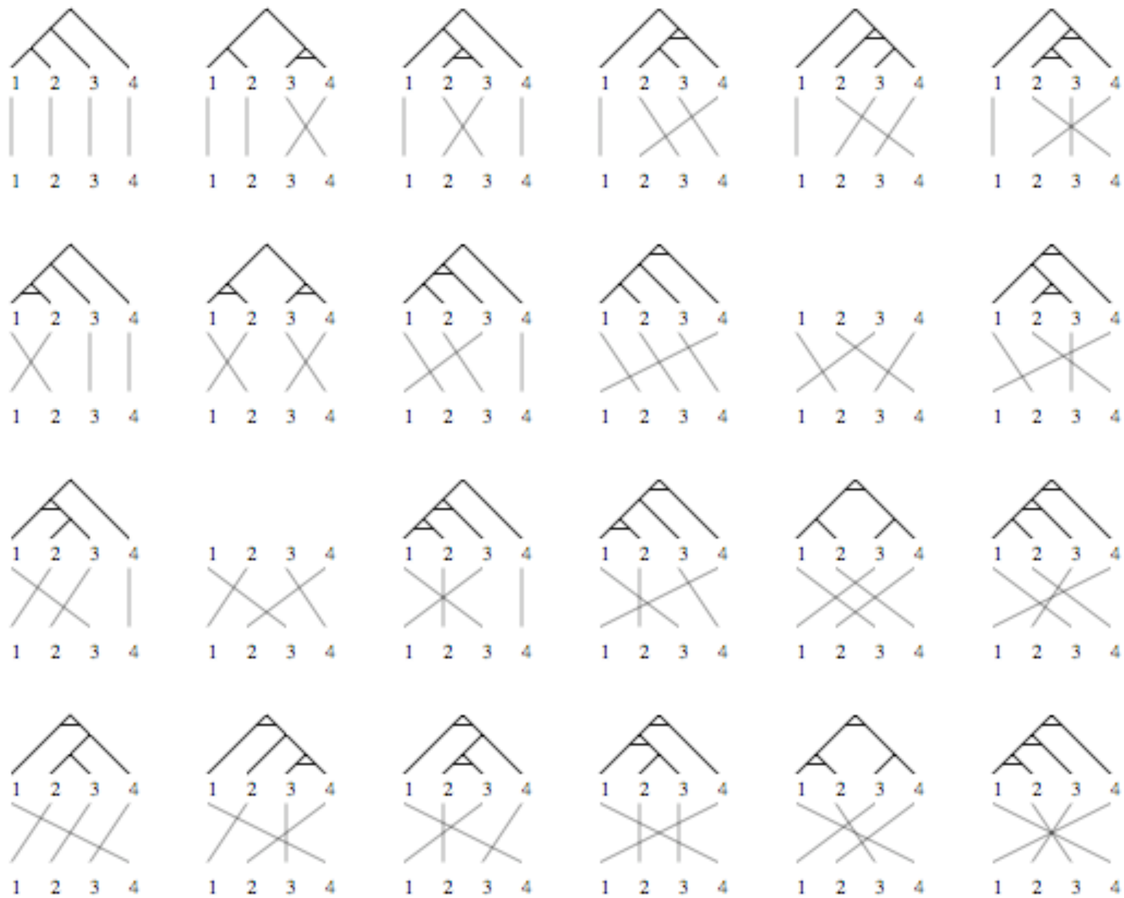
Run No.	**Pharaoh**	**Moses**	**WSD**
1	41.14	**41.17**	
2	41.65	41.70	**43.47**
3	42.05	42.16	
4	43.40	**43.55**	
5	41.92	**42.26**	
6	42.80	**43.19**	
7	43.76	**44.28**	
8	44.17	**44.64**	
9	51.64	**52.19**	
10	52.15	**52.59**	
11	53.51	**53.64**	
12	**53.87**	53.53	

Outline

- System description
- Experimental setup
  - Chinese-English
  - Other language pairs
- Results
- Contrastive experiments
  - (semantics) Phrase Sense Disambiguation: WSD for SMT
  - (syntax) Bracketing ITG decoder

Contrastive experiments (syntax)
Decoding under the ITG Hypothesis

- Intrinsically imposes ITG constraints on permutations/reorderings [Wu 1995]



Contrastive experiments (syntax)
Bracketing ITG decoder

- Basic decoding algorithm is polynomial-time $O(n^7)$ [Wu 1996]
- Current version uses beam search
- Current version integrates trigram LM
  - Note: did not use 4-gram LM or Gigaword 3-gram LM, so has less information than the Moses and Pharoah models
- Phrase-based SMT's distortion feature replaced by BITG permutation score
- All other factors controlled to be the same as Moses and Pharoah
  - Note: did not yet take advantage of any additional syntactic or other information naturally integrated into ITGs

Contrastive experiments (syntax)
BITG decoding competitive with Moses

- Again, encouraging preliminary indications

Run No.	Pharaoh	Moses	WSD	BITG
1	41.14	**41.17**		
2	41.65	41.70	**43.47**	
3	42.05	42.16		**43.04**
4	43.40	**43.55**		
5	41.92	**42.26**		
6	42.80	**43.19**		
7	43.76	**44.28**		
8	44.17	**44.64**		
9	51.64	**52.19**		
10	52.15	**52.59**		
11	53.51	**53.64**		
12	**53.87**	53.53		

Conclusion

- We have described experiments at HKUST focusing primarily on the Chinese-English task
  - also reported results on 3 other language pairs from different language families
- On Chinese-English, both our Pharaoh and Moses based systems achieved good performance
- Moses almost always outperforms Pharaoh
  - across a wide variety of experimental conditions
- Preliminary indications from contrastive experiments:
  - our WSD-for-SMT model improves Moses too
  - plain vanilla BITG decoding appears competitive with Moses