

The NUS Statistical Machine Translation System for IWSLT 2009



Preslav Nakov, Chang Liu, Wei Lu, Hwee Tou Ng

Dept. of Computer Science
National University of Singapore

IWSLT'2009 – Tokyo, Japan
December 2, 2009

Overview



Overview

- Chinese-English BTEC task
- Statistical Phrase-Based Machine Translation
- Focus
 - Different Chinese Word Segmentations
 - System Combination (Re-ranking)
 - Retrain on the Development Data after Tuning

Pre-processing



Pre-processing Steps

1. ASCII-ization
2. Sentence Breaking
3. Capitalization
4. English Re-tokenization
5. Number Translation

Step 1: ASCII-ization

- Convert to ASCII all full-width English letters and digits on the Chinese side

- Example:

你好 。 我 是 I c h i r o T a n a k a 。 我 想 改 变 我
5 月 1 7 号 的 预 订 。



你好 。 我 是 **Ichiro Tanaka** 。 我 想 改 变 我 **5 月 1
7** 号 的 预 订 。

- Add Chinese-side ASCII tokens to both sides of the bi-text

Step 2: Sentence Breaking

□ Split multi-sentence lines

- Should be consistent on both sides of the bi-text
- 16% more BTEC sentence pairs: from 19,972 to 23,110.

□ Example pair:

- 你好 。 我 是 Ichiro Tanaka 。 我 想 改 变 我 5 月 1 7 号 的 预 订 。
- Hello . This is Ichiro Tanaka . I'd like to change my reservation for May seventeenth .



你好 。 -- Hello .

我 是 Ichiro Tanaka 。 -- This is Ichiro Tanaka .

我 想 改 变 我 5 月 1 7 号 的 预 订 。 -- I'd like to change my reservation for May seventeenth .

Step 3: Capitalization

- Convert to lowercase

- Example:

我 是 Ichiro Tanaka 。 || This is Ichiro Tanaka .



我 是 **i**chiro **t**anaka 。 || **t**his is **i**chiro **t**anaka .

- In the final submission, we use a recaser.

Step 4: English Re-tokenization

- Split tokens with internal apostrophes

- reduces data sparseness

- Examples

- important

the shower water's too hot . → the shower **water 's** too hot .

my wallet's been stolen . → my **wallet 's** been stolen .

my name's kurosawa . → my **name 's** kurosawa .

- maybe not so important

let's hurry , or we'll be late . → **let 's** hurry , or **we 'll** be late .

where's the jal counter ? → **where 's** the jal counter ?

i'd like some crayfish . → **i 'd** like some crayfish .

Step 5: Number Translation (1)

□ Numbers

- hard to translate
- abundant
- inconsistent

□ training (BTEC corpus)

- Chinese side: digits, e.g., 3
- English side: words, e.g., *three*

□ tuning: almost no digits; all spelled as words

□ Solution: translate all numbers (both digits and words) on the Chinese-side to English

- 1 8 美元 正 。 → **eighteen** 美元 正 。
- 差 二 十 分钟 八 点 。 → 差 **twenty** 分钟 **eight** 点 。
- 在 三 层 什 么 地 方 ？ → 在 **third** 层 什 么 地 方 ？

Step 5: Number Translation (2)

- Manual study to identify the types of numbers on the English-side:
 1. *Integer*, e.g., *size twenty-two*.
 2. *Digits*, e.g., *flight number one one three*.
 3. *Series*, e.g., *March nineteen ninety-nine*.
 4. *Ordinal*, e.g., *July twenty-seventh*.
 5. *Others*: all other cases, e.g., — (‘one’) translated as a/an in English.

Step 5: Number Translation (3)

- Chinese-side numbers translation
 1. Choose a category (*Integer, Digits, Series, Ordinal, Others*)
 - Maximum entropy classifier
 - Features:
 - (1) number of digits in the numerical form;
 - (2) the numerical value of the number;
 - (3) the preceding word;
 - (4) the preceding character;
 - (5) the following word;
 - (6) the following two words; and
 - (7) preceding + following word.
 2. Translate to English using category-specific manual rules
 - no translation for *Others*
 3. Add the English words for numbers (e.g., *ten, three*) to both sides of the bi-text.

Chinese Word Segmentation and System Combination



Chinese Word Segmentation

□ Problem:

- Words are not well-defined in Chinese
 - Character-Word-Phrase continuum
- Various word segmentation standards
 - AS: Academia Sinica
 - CTB: UPenn Chinese Treebank
 - CITYU: City University of Hong Kong
 - PKU: Peking University
 - MSR: Microsoft Research

□ Our solution

- Train seven systems – one for each segmentation:
 - above five + ICTCLAS + default
- Combine their outputs

System Combination (1)

□ System combination

■ Training

□ Input data:

- Run all systems on the development set
- For each test input sentence
 - take the best translation from each system
 - extract features
 - calculate a bi-gram BLEU score
(4-gram BLEU is often zero at the sentence-level)

□ Classifier:

- Maximum entropy
 - selects the candidate with the highest oracle BLEU

System Combination (2)

□ System combination

■ Classifier's features

- 13 scores from the decoder:
 - 5 from the distortion model
 - 2 from the phrase translation model
 - 2 from the lexical translation model
 - 1 for the language model
 - 1 for the phrase penalty
 - 1 for the word penalty
 - 1 for the overall translation score
- + a global 14th score: *repetition count*

the most
important
feature



Training Methodology



Training Methodology: Phase I

□ Dev-time training

1. Build an SMT model

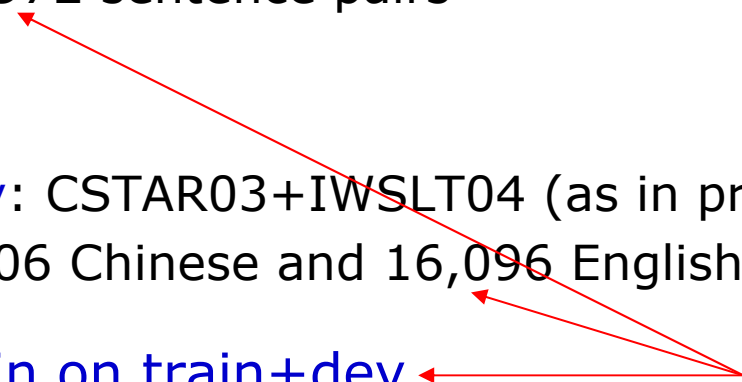
- Train: BTEC corpus
- 19,972 sentence pairs

2. Tune

- Dev: CSTAR03+IWSLT04 (as in prev. research)
- 1,006 Chinese and 16,096 English sentences

2.5. Retrain on train+dev

double the training data
(on the English-side)



3. Evaluate

- Test: IWSLT05, IWSLT07, and IWSLT08 (independently)
- 1,502 Chinese and 19,142 English sentences (in total)

Training Methodology: Phase II

- Re-ranker training

1. Run Phase I for all segmenters.
2. Train a re-ranker
 - Cross-validation on IWSLT05, IWSLT07, IWSLT08
 - Optimize the average BLEU score
 - Do feature selection
3. Re-train the re-ranker
 - IWSLT05+IWSLT07+IWSLT08


Training Methodology: Phase III

□ Training for final submission

1. Build an SMT Model

- Data: BTEC+IWSLT05+IWSLT07+IWSLT08
- 39,114 sentence pairs:
 - 19,972 BTEC pairs
 - 19,142 English sentences + 1,502 Chinese (repeated)

double
training
data



2. Tune the parameters

- Data: CSTAR03+IWSLT04
- 1,006 Chinese and 16,096 English sentences

3. Retrain

- Data:
BTEC+IWSLT05+IWSLT07+IWSLT08+CSTAR03+IWSLT04
- 55,210 sentence pairs:
 - 19,972 BTEC pairs
 - 35,238 English sentences + 2,508 Chinese (repeated)

triple
training
data



4. Test

- Run on the actual test data

5. Combine the system outputs (using the re-ranker)

Some Parameter Settings

□ Non-standard parameter settings

(Meta-)Parameter	Standard Setting	Our Setting
Language model order	3	5
Language modeling toolkit	SRILM	IRSTLM
Word aligner	GIZA++	Berkeley aligner
Alignment combination heuristic	grow-diag-final	intersection
Phrase reordering model	distance	monotonicity-bidirectional-f
Maximum phrase length	7	8
BLEU reference length used in MERT	shortest	closest
Miscellaneous	–	drop unknown words

Training-time Evaluation



Effect of Pre-processing

- **Excluding** each of the pre-processing steps

Excluded Pre-processing Step	IWSLT05	IWSLT07	IWSLT08	Average	
ASCII-ization	0.5286	0.3104	0.4438	0.4276	-0.43
Sentence breaking	0.5260	0.3100	0.4535	0.4298	-0.21
Number translation	0.5272	0.2941	0.4262	0.4158	-1.61
English re-tokenization	0.5244	0.3102	0.4439	0.4262	-0.57
<i>Keep all (i.e., exclude none)</i>	0.5189	0.3264	0.4503	0.4319	

lost
Bleu
points



Using the default segmentation

Effect of Non-standard Settings

□ Reverting settings to default values

Our Setting	Revert to	IWSLT05	IWSLT07	IWSLT08	Average	
Berkeley aligner	GIZA++	0.5246	0.3101	0.4342	0.4230	-0.89
mono-bidirectional-f	distance	0.5230	0.2983	0.4333	0.4182	-1.37
drop unknown words	–	0.5157	0.3023	0.4278	0.4153	-1.66
<i>Keep all (i.e., revert nothing)</i>		0.5189	0.3264	0.4503	0.4319	

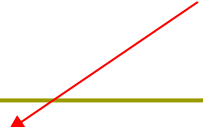
lost
Bleu
points



Using the default segmentation

Effect of Re-training

On Phase III,
(final submission)
data is tripled!



Effect of retraining on phase II (doubled data)

Segmentation	Re-training	IWSLT05	IWSLT07	IWSLT08	Average	Bleu points won
Default	before	0.5161	0.2887	0.4241	0.4096	
	after	0.5189	0.3264	0.4503	0.4319	+2.23
ICTCLAS	before	0.5296	0.2923	0.4149	0.4123	
	after	0.5394	0.3129	0.4539	0.4354	+2.31
AS	before	0.5321	0.2901	0.4053	0.4092	
	after	0.5272	0.3074	0.4458	0.4268	+1.76
CITYU	before	0.5390	0.2899	0.4002	0.4097	
	after	0.5304	0.3208	0.4301	0.4271	+1.74
CTB	before	0.5279	0.3012	0.4138	0.4143	
	after	0.5319	0.3053	0.4550	0.4307	+1.64
MSR	before	0.5337	0.2921	0.4214	0.4157	
	after	0.5338	0.3217	0.4406	0.4320	+1.63
PKU	before	0.5317	0.2977	0.4070	0.4120	
	after	0.5367	0.3164	0.4501	0.4344	+2.24

Effect of System Combination

- System combination and best individual systems on cross-validation

Trained on	Tested on	Combination BLEU	Best Individual Segmenter	BLEU
IWSLT07 + IWSLT08	IWSLT05	0.5457	ICTCLAS	0.5394
IWSLT05 + IWSLT08	IWSLT07	0.3268	Default	0.3264
IWSLT05 + IWSLT07	IWSLT08	0.4656	CTB	0.4550
<i>Average</i>		0.4460	–	0.4403

Main Sources of Improvement

- Pre-processing
 - +1.6: number translation
 - +0.6: English re-tokenization

- Moses tuning
 - +1.7: dropping unknown words
 - +1.4: lexicalized reordering
 - +0.9: Berkeley aligner

- Re-training
 - +2.0: as measured for Phase II (even more for Phase III)

- System combination (and segmentation)
 - +1.4: over the default segmentation
 - +1.0: over the single best segmentation
 - +0.6: over picking the best segmentation for each testset (i.e., IWSLT05, IWSLT07, IWSLT08)

Official Evaluation



Automatic Evaluation: Sign. Test

<i>“case+punc” evaluation</i>								CRR
bleu	meteor	wer	per	ter	gtm	nist	z-avg	
49.70	72.67	41.02	35.54	33.65	72.53	7.363	2.178	nlpr
44.77	68.09	44.03	38.96	35.85	69.66	6.494	1.344	nus
45.94	67.24	43.83	39.39	35.71	69.55	6.110	1.250	i2r
40.58	66.20	50.05	42.42	42.01	69.46	6.774	0.786	uw
42.38	64.48	45.66	41.73	36.25	66.83	4.858	0.545	dcu
39.53	64.18	48.45	42.81	39.38	66.87	5.853	0.489	bmrc
40.15	60.78	49.18	43.75	41.46	67.68	5.890	0.323	lium
35.33	62.70	51.93	44.82	41.81	65.96	5.821	0.068	upv
35.41	62.71	49.96	44.65	40.58	63.47	5.647	0.022	tokyo
35.65	62.27	50.78	45.06	41.57	64.60	5.610	-0.011	ict
31.49	61.71	55.88	47.60	48.06	64.79	6.154	-0.405	tottori
27.92	55.35	59.22	53.25	51.61	59.62	5.457	-1.444	greyc

Automatic Evaluation: Full Testset

<i>“case+punc” evaluation</i>								CRR
bleu	meteor	wer	per	ter	gtm	nist	z-avg	
49.69	72.66	41.04	35.55	33.67	72.52	7.696	2.239	nlpr
44.81	68.08	44.04	38.97	35.86	69.66	6.780	1.462	nus
45.95	67.25	43.83	39.38	35.70	69.56	6.384	1.370	i2r
40.61	66.21	50.04	42.39	41.99	69.47	7.048	0.944	uw
42.37	64.47	45.68	41.75	36.26	66.83	5.063	0.704	dcu
39.55	64.19	48.46	42.80	39.37	66.86	6.096	0.667	bmrc
40.14	60.76	49.21	43.78	41.48	67.68	6.119	0.497	lium
35.38	62.69	49.97	44.66	40.59	63.44	5.862	0.237	tokyo
35.29	62.66	51.99	44.86	41.86	65.93	6.047	0.268	upv
35.63	62.26	50.80	45.07	41.58	64.59	5.841	0.204	ict
31.51	61.69	55.90	47.60	48.07	64.78	6.383	-0.160	tottori
27.95	55.37	59.23	53.24	51.61	59.64	5.657	-1.117	greyc

Human Evaluation

MT	Ranking
nlpr	0.4985
nus	0.3891
i2r	0.3781
ict	0.3737
uw	0.3219
tottori	0.3174
upv	0.3125
bmrc	0.3066
lium	0.2976
tokyo	0.2956
dcu	0.2900
greyc	0.2697

MT	NormRank
nlpr	3.55
nus	3.24
i2r	3.17
ict	3.12
uw	3.01
upv	2.99
bmrc	2.95
dcu	2.91
tokyo	2.87
tottori	2.84
lium	2.78
greyc	2.63

Conclusion and Future Work



We Also Tried...

- Hierarchical phrase-based SMT model
 - Performed worse than phrase-based SMT
 - Combining with phrase-based SMT did not help

- Word sense disambiguation for SMT
 - +0.5-1.0 Bleu points
 - could not be included in our final submission due to logistic issues

Y. S. Chan, H. T. Ng, and D. Chiang, "Word sense disambiguation improves statistical machine translation," in Proceedings of ACL, 2007.

Conclusion and Future Work

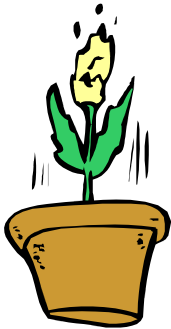
□ Main feature of our system

- Using different Chinese word segmentations
- System combination
- Retrain on the development data after tuning

□ Future work

- Better integration of different Chinese word segmentations
- Lattice-based system combination
- Incorporate word sense disambiguation

Thank You



Any questions?

This research was supported by research grants
CSIDM-200804 and POD0713875.