POSTECH Machine Translation System for IWSLT 2008 Evaluation Campaign

Jonghoon Lee and Gary Geunbae Lee {jh21983, gblee}@postech.ac.kr

Baseline system

Task : BTEC (AE, CE, and CS)

Corpus used: (provided by IWSLT 2008 only)
- •Arabic : Romanize, tokenize, and attach POS tag using Arabic analyzer[M. Diab]
- •Chinese : attach POS using Stanford parser
- •Spanish : tokenize punctuation marks
- •English : tokenize punctuation marks

Translation Modeling : Moses training script
- •Phrase translation probability(bi-direction)
- •Word translation probability(bi-direction)
- •Phrase penalty
- •Distance based distortion model

Language Modeling : SRILM
- •N-gram back-off

Weight optimizing : MERT module in Moses
Decoding: Moses decoder

Corpus statistics

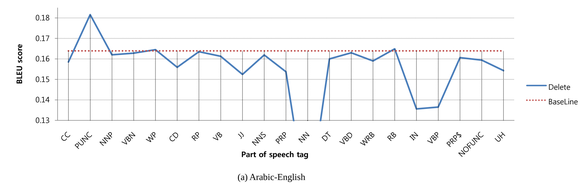
		Arabic	Chinese	English	Spanish
Train	Sent.		19972		
	Word	150303	171591	189558	185527
	Vcb.	14854	8428	8170	10995
Dev1	Sent.	506	506	506*16	
	Word	2865	3354	61176	
	Vcb.	1102	880	983	
Dev2	Sent.	500	500	500*16	
	Word	3040	3449	61615	
	Vcb.	1180	920	979	
Dev3	Sent.	506	506	506*16	506*16
	Word	2918	3767	62690	60501
	Vcb.	1174	931	997	1151
Dev4	Sent.	489	489	489*7	
	Word	4825	5715	46042	
	Vcb.	1473	1143	1157	
Dev5	Sent.	500	500	500*7	
	Word	5341	6066	51874	
	Vcb.	1797	1339	1354	
Dev6	Sent.	489	489	489*6	
	Word	2757	3169	22366	
	Vcb.	1119	881	924	
Test	Sent.	507	507		
	Word	2955	2808		
	Vcb.	1139	885		

Source word Deletion

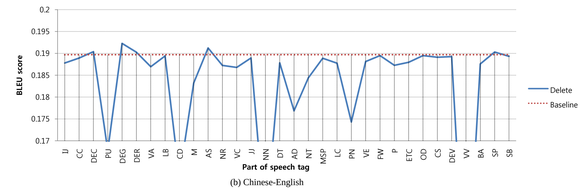
Motivation:
- •Language difference
- •Untranslatable (useless) words

Method
- •Exclude a untranslatable group of Source vocabulary from training
  →Grouping by Part Of Speech
- •Identifying untranslatable words
  → POS-wise Deletion test

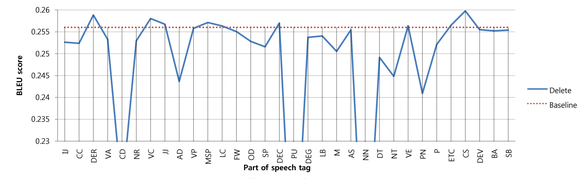
Deletion test



(a) Arabic-English



(b) Chinese-English



(c) Chinese-Spanish

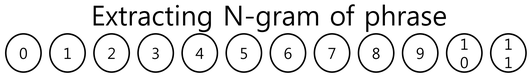
Phrase level language model (Multiword n-gram)

Motivation
- To cover longer history
- modeling Inter-phrase reordering directly

Method
- Phrase vocabulary from phrase-table
- N-gram of phrase

$$P(e|f) = p(f|e)^{\lambda_{1-1}} \times l(f|e)^{\lambda_{1-2}} \times p(e|f)^{\lambda_{1-3}} \times l(e|f)^{\lambda_{1-4}} \times P(e)^{\lambda_5} \times P_d(e,f)^{\lambda_6} \times \exp(length(e)\lambda_4 + 1) \times P_{phrase}(e)^{\lambda_5}$$

Extracting N-gram of phrase



[0] [1] [2]
[1] [2] [3]
[1] [2] [3 4]
[0 1] [2] [3]
[0 1] [2] [3 4]

[2] [3] [4]
[2] [3 4] [5]
[2] [3] [4 5]
[3] [4] [5]
[3] [4 5] [6]

[3] [4 5] [6 7]
[3 4] [5] [6]
[3 4] [5] [6 7]
[4] [5] [6]
[4] [5] [6 7]

[4 5] [6] [7]
[4 5] [6 7] [8]
[4 5] [6 7] [8 9 10]
[5] [6] [7]
[5] [6 7] [8]

[5] [6 7] [8 9 10]
[6] [7] [8]
[6] [7] [8 9 10]
[6 7] [8] [9]

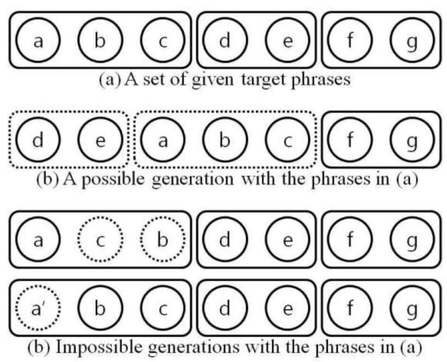
[6 7] [8] [9 10]
[6 7] [8 9 10] [11]
[7] [8] [9]
[7] [8] [9 10]

[7] [8 9 10] [11]
[8] [9] [10]
[8] [9] [10 11]
[8] [9 10] [11]

[9] [10] [11]

→ 33 phrase trigrams

Reordering in phrase SMT



Effect of phrasal LM

	Word 3gram	Word 6gram	Word 3gram Phrase 2gram
AE	0.3892	0.4025	0.3940
CE	0.3024	0.2998	0.3039
CS	0.2378	0.2485	0.2570

Evaluation

Dev. set : Merged set of all possible dev set

Contrast:
- • Moses system

Primary:
- • Moses system
  +Phrasal Language Model
  +Source word deletion

Conclusions & Further works:
- •Detailed Identification method is Required for source word deletion
- •Find out features that prefer to Model longer history

MERT result on Dev. set

		Baseline *contrast*	Deleting	PLM	Both *primary*
AE	CRR	0.2700	0.2712	0.2703	0.2718
	ASR	0.1628	0.1657	0.1627	0.1659
CE	CRR	0.1896	0.1922	0.1899	0.1920
	ASR	0.1233	0.1214	0.1239	0.1221
CS	CRR	0.2443	0.2578	0.2551	0.2580
	ASR	0.1677	0.1771	0.1772	0.1782

Changes in BLEU

	CRR	ASR
Ar-En	↓	↑
Ch-En	↓	↑
Ch-Es	↑	↓

Official evaluation Result

			BLEU	NIST	WER	PER	GTM	METEOR	TER
BTEC_AE case punc	CRR	Primary	0.3878	7.6156	0.4690	0.4198	0.6994	0.6177	41.9660
		Contrast	0.3892	7.5924	0.4662	0.4201	0.6967	0.6167	41.3530
	ASR	Primary	0.2999	6.3244	0.5441	0.4904	0.6306	0.5482	48.6370
		Contrast	0.2973	6.3502	0.5554	0.5011	0.6224	0.5441	49.8150
BTEC_AE no case no punc	CRR	Primary	0.3867	8.1558	0.4742	0.4183	0.6866	0.6172	41.0170
		Contrast	0.3895	8.1078	0.4717	0.4183	0.6843	0.6189	40.4640
	ASR	Primary	0.2929	6.5991	0.5600	0.4976	0.6059	0.5429	49.1490
		Contrast	0.2875	6.6507	0.5754	0.5080	0.6031	0.5407	50.7550
BTEC_CE case punc	CRR	Primary	0.2841	6.3012	0.6179	0.5302	0.6299	0.5104	54.1560
		Contrast	0.3024	6.4593	0.6141	0.5264	0.6308	0.5150	53.6900
	ASR	Primary	0.2624	6.2410	0.6432	0.5546	0.6048	0.4897	57.8100
		Contrast	0.2511	6.0865	0.6557	0.5591	0.5886	0.4851	59.2570
BTEC_CE no case no punc	CRR	Primary	0.3052	7.1788	0.6056	0.4924	0.6591	0.5462	52.9830
		Contrast	0.3212	7.3788	0.6026	0.4896	0.6595	0.5533	52.4600
	ASR	Primary	0.2792	6.9036	0.6415	0.5272	0.6262	0.5199	57.7760
		Contrast	0.2692	6.7507	0.6552	0.5397	0.6129	0.5168	59.2280
BTEC_CS case punc	CRR	Primary	0.2594	5.3343	0.6249	0.5494	0.5728	0.2731	54.2000
		Contrast	0.2378	5.0502	0.6433	0.5752	0.5453	0.2695	56.1500
	ASR	Primary	0.2104	5.4017	0.7335	0.6398	0.5643	0.2658	70.7500
		Contrast	0.2204	5.0648	0.6836	0.6031	0.5297	0.2535	60.6000
BTEC_CS no case no punc	CRR	Primary	0.2537	6.0118	0.6472	0.5445	0.5764	0.2823	56.2900
		Contrast	0.2340	5.9162	0.6553	0.5571	0.5621	0.2827	57.5340
	ASR	Primary	0.1908	5.3856	0.7628	0.6462	0.5650	0.2722	74.7870
		Contrast	0.2150	5.5472	0.6944	0.5941	0.5341	0.2651	61.9350