The University of Edinburgh System Description for IWSLT 2007

Josh Schroeder and Philipp Koehn

October 15, 2007

School of informatics

Introduction

- Focus on Italian - English Challenge Task
- Domain Adaptation
  - SITAL data is distinct domain from BTEC corpus
  - Cross-domain adaptation with multiple translation models
- Speech Input Experiments

System Summary - Tools

- Moses phrase-based decoder (`http://www.statmt.org/moses`)
- GIZA++ for phrase extraction (through Moses training scripts)
- SRILM for language modelling
- MERT for tuning

System Summary - Approach

- Punctuation: unpunctuated source to punctuated target
- Max sentence length 80, grow-diag-final-and phrase extraction
- 5-gram language model
- Casing: recaser trained on cased target language data
- Two separate corpora (BTEC and Europarl) for cross-domain adaptation
- Experimented with Moses' lattice input for confusion network decoding

System Summary - Data

- IT-EN only
- Europarl training data from v2 release (v3 released 28 Sept)
  `http://www.statmt.org/europarl`
- BTEC training data
- Used ACL WMT07 test data to extract matching 2000 sentences for Italian in Europarl domain
- Split SITAL development randomly in half (tuning and devtest) during translation experiments
- Used devset4 and devset5a from BTEC domain since they have lattice input format

Training Corpora

BTEC	Italian	English
Sentences	19,972	
Words	147,564	188,961
Phrase table entries	314,874	

Europarl	Italian	English
Sentences	868,047	
Words	22,586,316	25,267,363
Phrase table entries	49,018,026	

Domain Adaptation

- Concerned with cross-domain adaptation, not dynamic adaptation.

- Our previous work (ACL WMT07) focused on using separate training corpora:
  - Small in-domain set (News Commentary), with large out-of-domain supplement (Europarl).
  - Training separate models and using both in the decoder was more effective than using only one of the corpora or combining both corpora in one model.

- Is this a similar situation?

Domain Problem

The SITAL test data is not in the same domain as either the BTEC or Europarl training corpora.

We can examine this on development test sets in each domain:

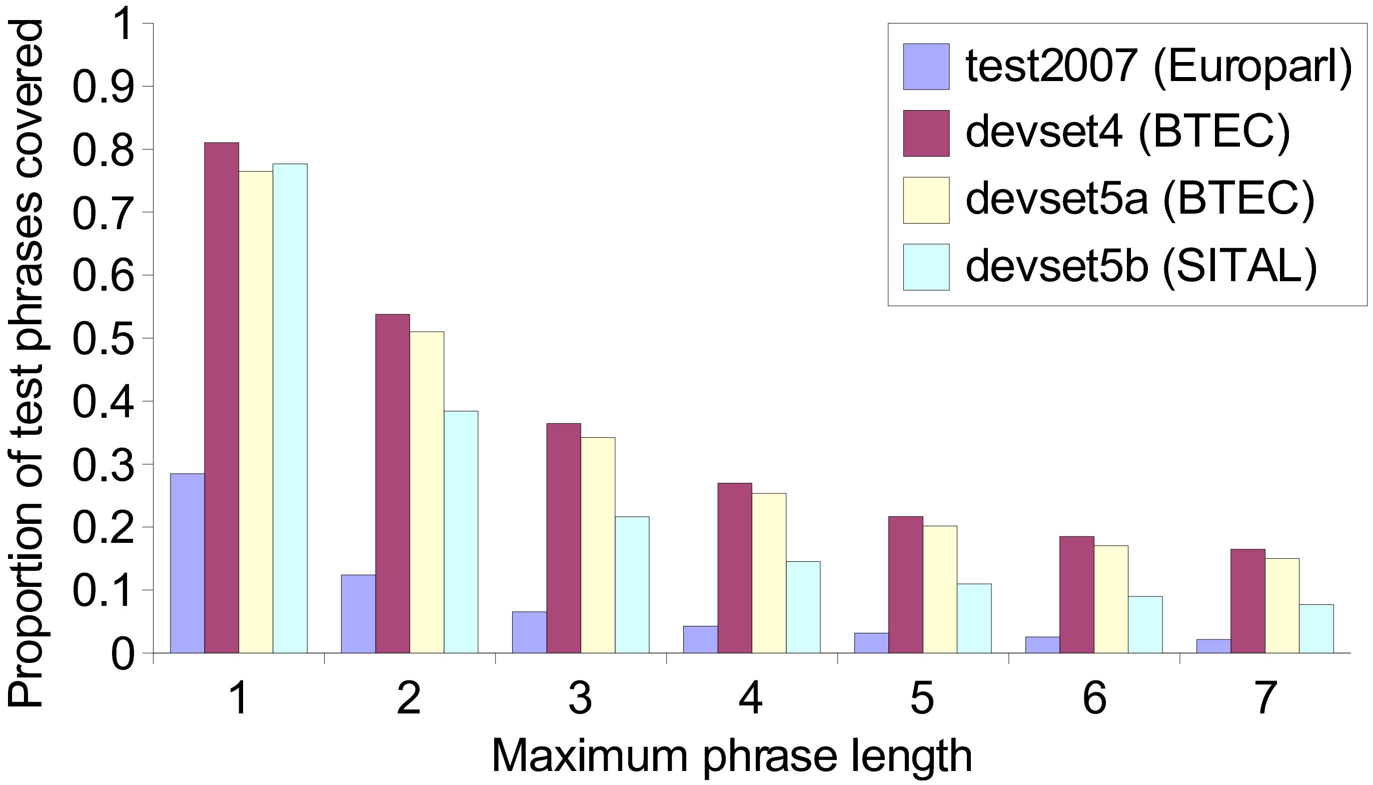
- Look at source-side LM perplexity

- Explore phrase table coverage

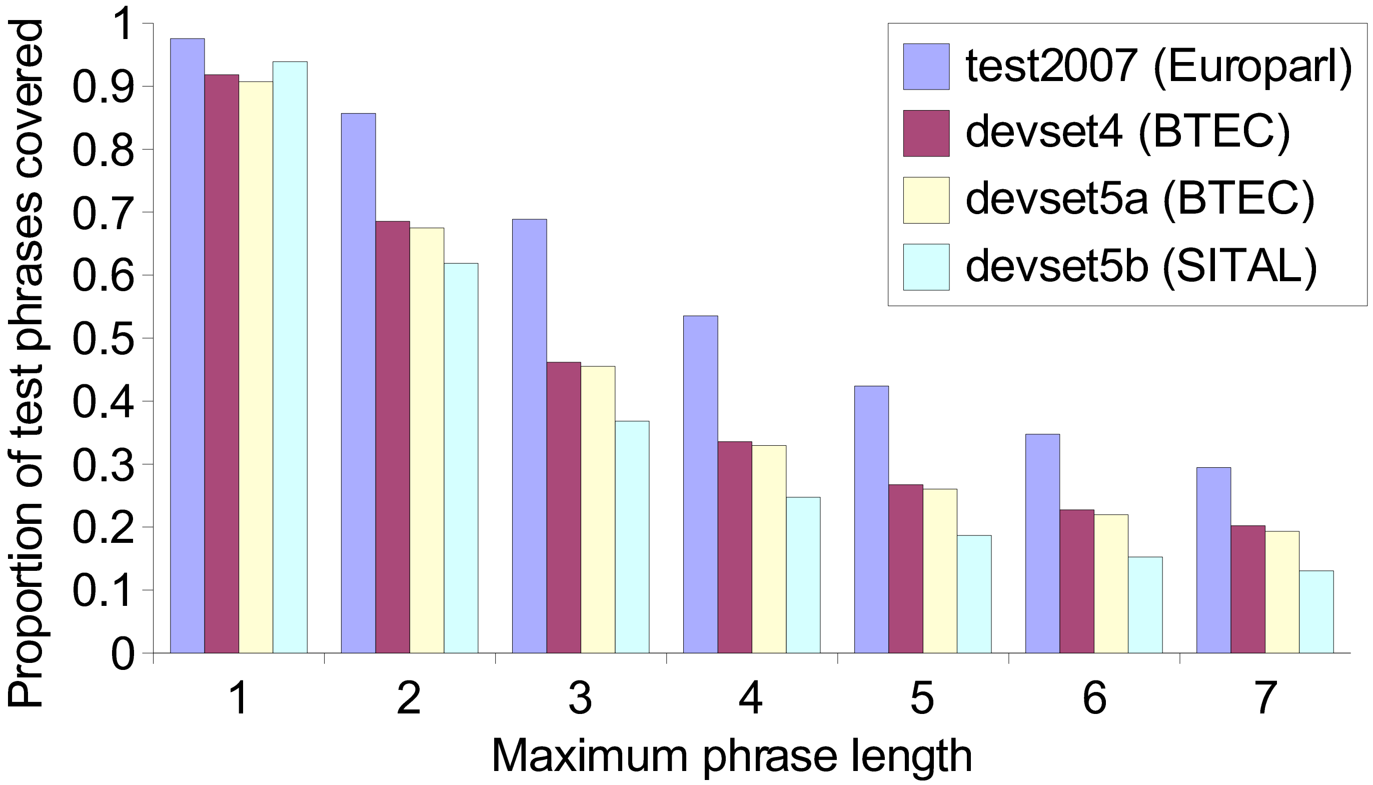
Domain Problem - LM Perplexity

	Test Set	LM	
Test Set	**Domain**	**Corpus**	**Perplexity**
test2007	Europarl	BTEC	982.9
devset4	BTEC	BTEC	171.7
devset5a	BTEC	BTEC	184.2
devset5b	SITAL	BTEC	311.8
test2007	Europarl	Europarl	94.2
devset4	BTEC	Europarl	1294.4
devset5a	BTEC	Europarl	1139.3
devset5b	SITAL	Europarl	1868.9

Domain Problem - BTEC Phrase Table Coverage



Domain Problem - Europarl Phrase Table Coverage



Domain Problem
Unigram vs. Bigram Coverage

Unigram Coverage (first 4,976 words of test set)				
Test Set	Unique Unigrams	BTEC Coverage	Europarl Coverage	Combined Coverage
test2007	1737	788 (45.4%)	1721 (99.1%)	1721 (99.1%)
devset4	1234	1000 (81.0%)	1133 (91.8%)	1160 (94.0%)
devset5a	1331	1040 (78.1%)	1212 (91.1%)	1249 (93.8%)
devset5b	600	497 (82.8%)	564 (94.0%)	570 (95.0%)

Domain Problem
Unigram vs. Bigram Coverage

Bigram Coverage (first 4,976 words of test set)				
	Unique	BTEC	Europarl	Combined
Test Set	Bigrams	Coverage	Coverage	Coverage
test2007	4010	573 (14.3%)	3505 (87.4%)	3506 (87.4%)
devset4	3303	1441 (43.6%)	1977 (59.9%)	2237 (67.7%)
devset5a	3458	1458 (42.2%)	2091 (60.5%)	2318 (67.0%)
devset5b	2384	795 (33.3%)	1336 (56.0%)	1458 (61.2%)

Domain Problem - SITAL Differences

- Oddly, the SITAL data (devset5b) has better unigram coverage but worse n-gram coverage (n $>$ 1) than the BTEC test sets (devset4 and devset5b).

- Spontaneous speech uses different word patterns?

- More repetition of a smaller set of vocabulary (half as many unique words)?

Cross-Domain Adaptation - Approaches

How do we best utilize two parallel corpora for translation in a third domain?

- Choose one corpus, build a model.
- Combine the corpora together, build one model.
- Keep corpora separate, build complex model.

Cross-Domain Adaptation - devset5b Results

Method	Table and LM Source(s)	%BLEU for	
		TEXT	1-BEST
Single corpus	Europarl	16.0	14.5
	BTEC	19.6	18.5
Corpus combination	Combined (6x BTEC + Europarl)	21.5	20.4
Separate corpora	BTEC, Europarl	**23.0**	**21.1**

Speech Input Experiments

Moses supports confusion network input. Previous work has shown that confusion network input provides better translations than 1-best input.

New input format specifies confusion network data (and more complex lattice data) in one-line format:

```
((('i',0.9,1),('eye',0.1,1),), \\
(('like',0.95,1),('lichen',0.05,2),), \\
(('them',1.0,1),),)
```

Can also be really simple:

```
((('grazie',1),),(('buongiorno',1),),)
```

Speech Input Experiments

Or complex:

```
(((’hotel’,1),),((’san’,1),),((’marco’,1),),((’*EPS*’,0.997997), \\
(’’,0.00200254),),((’un’,0.746887),(’*EPS*’,0.129127), \\
(’no’,0.113639),(’a’,0.00700725),(’ma’,0.00187423), \\
(’in’,0.00146577),),((’un’,0.878394),(’no’,0.106656), \\
(’non’,0.0149496),),((’e’,1),),((’diceva’,1),),((’che’,1),), \\
((’si’,1),),((’trova’,1),),((’in’,1),),((’una’,1),), \\
((’zona’,1),),((’centrale’,1),),)
```

Speech Input Experiments

We ran a series of experiments to test confusion network effectiveness for BTEC and SITAL data.

As shown in previous work, we expected to see confusion network inputs produce better translations than 1-best inputs.

- BTEC corpus and BTEC tune/test data performed as expected.
- Confusion network input did **not** help for SITAL data.

Speech Input Experiments - BTEC-BTEC

BTEC CORPUS		**BTEC test set devset4**		
Tuning set		**TEXT**	**1-BEST**	**CN**
BTEC	*devset5a TEXT*	40.1	34.1	34.6
	devset5a 1-BEST	40.4	34.3	34.4
	devset5a CN	**41.0**	35.6	**36.1**
SITAL	*devset5b-tune TEXT*	38.0	32.4	32.6
	devset5b-tune 1-BEST	38.0	32.4	**32.7**
	devset5b-tune CN	**38.2**	32.4	**32.7**

Speech Input Experiments - BTEC-BTEC

BTEC Corpus		**BTEC test set devset5a**		
Tuning set		**TEXT**	**1-BEST**	**CN**
BTEC	*devset4 TEXT*	**37.5**	30.8	31.1
	devset4 1-BEST	37.0	31.0	31.1
	devset4 CN	37.1	31.0	**31.2**
SITAL	*devset5b-tune TEXT*	34.5	29.1	29.1
	devset5b-tune 1-BEST	34.9	29.5	29.6
	devset5b-tune CN	**35.2**	29.5	**29.6**

Speech Input Experiments - BTEC-SITAL

BTEC Corpus		SITAL test set devset5b-test		
Tuning set		**TEXT**	**1-BEST**	**CN**
BTEC	*devset4 TEXT*	**19.3**	**17.8**	17.4
	devset4 1-BEST	17.7	16.0	15.8
	devset4 CN	18.3	16.4	16.6
BTEC	*devset5a TEXT*	15.7	14.6	13.8
	devset5a 1-BEST	**19.6**	**17.9**	17.1
	devset5a CN	18.0	16.5	16.1
SITAL	*devset5b-tune TEXT*	19.6	**18.5**	18.4
	devset5b-tune 1-BEST	19.5	18.4	18.1
	devset5b-tune CN	**19.7**	18.2	17.9

Speech Input Experiments - Domain Adaptation

Separate corpora		SITAL test set devset5b-test		
Tuning set		**TEXT**	**1-BEST**	**CN**
SITAL	*devset5b-tune TEXT*	**23.0**	**21.1**	18.6
	devset5b-tune 1-BEST	22.8	20.6	18.2
	devset5b-tune CN	—	—	—

Conclusion - Shared Task Submission

- Experiments in cross-domain adaptation showed usefulness of separate corpora approach.
- Confusion network input wasn't helping SITAL data, couldn't be fully tuned under current system.
- Better results for SITAL data by tuning with corrected text input and re-using those weights for 1-best translation.

Final submission for Italian-English was two corpus system (BTEC & Europarl), tuned with devset5b SITAL text data, used to translated text and 1-best inputs.

Conclusion - Future Work

- Investigate why SITAL confusion network data didn't help translation (Domain issue? Quality of SLF? User error?)

- Improve Moses' caching/filtering system for Lattice/Confusion Network input

- How many corpora can we use and still effectively tune?

Conclusion

Mille grazie!

Thank you!