Improved Statistical Machine Translation using Monolingually-derived Paraphrases

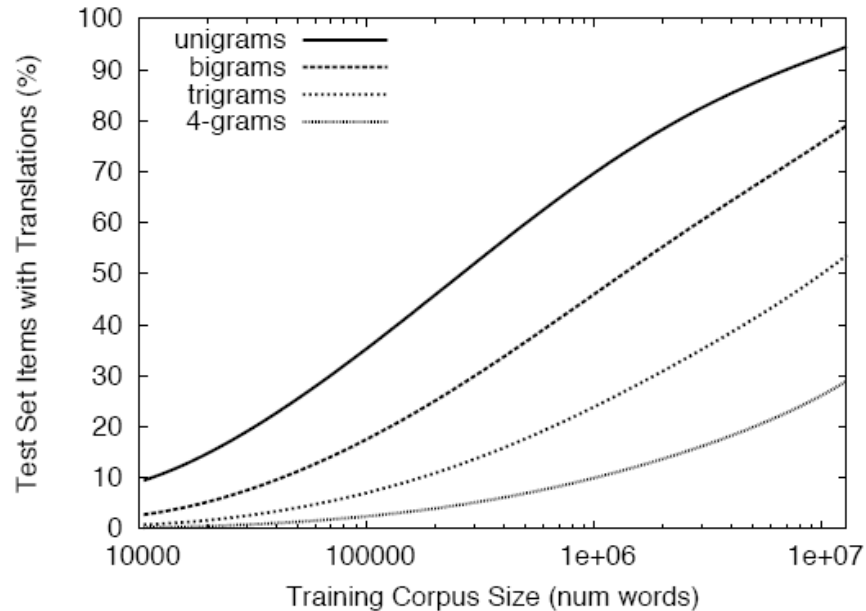Yuval Marton, Chris Callison-Burch and Phillip Resnik

EMNLP 2009

Motivation

- The problem of coverage in SMT

  Unigrams
  - 10K -> 10%
  - 100K -> 30%
  - 10M -> 90%

[Callison-Burch et al. (2006)]

- SMT systems are unable to handle OOV words

This paper

- Augment a Statistical Machine Translation system using paraphrases for OOV words

- Closely related to:

  Chris Callison-Burch, Philipp Koehn, and Miles Osborne. "Improved statistical machine translation using paraphrases". In Proceedings NAACL-2006.

- Main difference:
  – To generate paraphrases, Callison-Burch et al. (2006) uses *pivoting* through other languages; Requires parallel data with one side containing the source language.
  – This paper uses only monolingual data in source language

Approach

- For each OOV phrase phr
  - Build distributional profile (context vector) DP_{phr}
  - For each occurrence of phr in the corpus, find its left and right context L_R
  - For each such context, find paraphrase candidates with the same context

    i.e. all X such that LXR appears in the corpus
  - Build distributional profiles for DP_x
  - Rank all X by measuring profile similarity between DP_x and DP_{phr}
  - Pick top k candidates

Approach (Cont.)

- Building DPs
  - Uses a sliding window of 6 words
  - Log-likelihood ratio as the concurrence measure
- Finding context
  - Very short or very frequent contexts are non-informative
  - Start with one word and grow context until it reaches a stop word
- Ranking candidates
  - Uses cosine similarity between DPs

Approach (Cont.)

- Adding paraphrases to the phrase table

  - For each paraphrase f' that has a translation e, add additional entry (e, f) to the phrase table

  - Additional phrasal feature

$$h(e, f) = \begin{cases} psim(DP_{f'}, DP_f) & \text{If phrase table entry } (e,f) \text{ is generated from } (e,f') \\ & \text{using monolingually-derived paraphrases.} \\ \\ 1 & \text{Otherwise} \end{cases}$$

Experiments

- English-to-Chinese (E2C) and Spanish-to-English (S2E)
- Standard SMT system (GIZA++/Moses/MERT)
- Training Data

Set	# Tokens Source+Target
E2C 29K	0.8 + 0.6
E2C Full	6.4 + 5.1
bnc+apw	187
S2E 10K	0.3 + 0.3
S2E 20K	0.6 + 0.6
S2E 80K	2.3 + 2.3
wmt09	84
wmt09+acquis	139
wmt09+acquis+afp	402

- Dev/Test
  - E2C (MT05/MT08); S2E (Europarl dev06/test06)

E2C Results

- Character-based Bleu and TER

dataset	E2C model	BLEU	TER
29k	baseline	15.21	90.354
29k	1grams	16.87***	90.370
29k	1-6grams	16.54***	90.376
29k	1 + 2-6grams	**16.88***	**90.349**
Full	baseline	**22.17**	**90.398**
Full	1grams	21.64***	90.459
Full	1-6grams	21.75	90.421
Full	1 + 2-6grams	21.39***	90.433

S2E Results

bitext	mono.corp.	features	minScore	BLEU	TER
10k	(baseline)	–	–	23.78	62.382
10k	wmt09	1-4grams	.6	23.81	
10k	wmt09	1-2+3-4gr	.6	23.92	62.202
10k	wmt09+aquis	1-4grams	.6	**24.13*****	**61.739**
10k	wmt09+aquis	1grams	.6	24.11	61.979
20k	(baseline)	–	–	24.68	62.333
20k	wmt09+aquis	1-4grams	.6	**24.75**	**61.528**
80k	(baseline)	–	–	**27.89**	57.977
80k	wmt09+aquis	1-4grams	.6	27.82	**57.906**
10k	wmt09+aquis	1grams	.3	**24.11**	61.979
10k	wmt09+aquis+afp	1grams	.3	23.97	61.974
20k	wmt09+aquis+afp	1grams	.3	**24.77**	61.276
80k	wmt09+aquis+afp	1grams	.3	27.84***	**57.781**

Paraphrase Examples

Paraphrase	Score		Paraphrase	Score
Source: deal			*Source: fall*	
agreement	0.56		rise	0.87
accord	0.53		slip	0.82
talks	0.45		tumbled today	0.68
contract	0.42		fell today	0.67
peace deal	0.33		tumble	0.65
merger	0.32		fall tokyo ap stock prices fell	0.56
agreement is	0.30		are mixed	0.54
Source: to provide any other			*Source: we have a situation that*	
to give any	0.74		uncontroversial question about our	0.66
to give further	0.70		obviously with the developments this morning	0.65
to provide any	0.68		community staffing of community centres	0.64
to give any other	0.62		perhaps we are getting rather impatient	0.63
to provide further	0.61		er around the inner edge	0.60
to provide other	0.53		interested in going to the topics	0.60
to reveal any	0.52		and that is the day that	0.60
to provide any further	0.48		as a as a final point	0.59
to disclose any	0.47		left which it may still have	0.56
to publicly discuss the	0.43			

Discussion

- Monolingually derived paraphrases help improve performance of smaller systems, but not for larger systems
  - May help low-density languages or special domains
- Larger monolingual corpora yields better paraphrases

- Is Bleu, a good metric here?
  - Callison-Burch (2006) shows that Blue is insensitive to their improvements between 60-75% of the time.