Improving SMT by Paraphrasing the Training Data

Francis Bond (NICT)
Eric Nichols (NAIST), Darren Scott Appling (Georgia Tech),
Michael Paul (NICT)
<bond@ieee.org>

October 20, 2008

Motivation

- SMT quality improves with more (in-domain) data
  - +2.5% BLEU if you double the bitext
  - +0.5% BLEU for target)
- But bitext is rare (and expensive to make)
  which makes SMT/EBMT hard to deploy
- How can we cheaply make more bitext?
  - Translate more — expensive
  - Find more — great if there is some
  - Extend existing data — our approach

In other words — Paraphrasing

(1) この ことから、会社には事故の責任が無いことになる。

It follows from this that the company is not responsible for the accident.

It follows that the company isn't responsible for the accident from this.

It follows that the company is not responsible for the accident from this.

That the company isn't responsible for the accident follows from this.

Paraphrasing

- There are often multiple ways of saying the same thing
  - Almost always they differ in some nuance
  - But sometimes these differences are negligible
- Two main kinds of paraphrase
  - Structural — different word order, different function words
    We can do this by parsing and generating (using HPSG)
  - Lexical — different (open class) word choice
    We can do this with WSD and then sense substitution
- Need to do together for full generality
  I like pears. $\leftrightarrow$ *Pears please me.*

Previous Work

- Paraphrase by finding translation equivalents (Callison-Burch et al., 2006)
  - SMT with language A and C
  - find places where multiple A_1, A_2 links to one C
  - SMT between A and B
    - If A_1-B is not in phrase table, try to replace by A_2-B
- Example: *used, use, spent, utilize. to use* (Es: *usado*)
- Raises BLEU 1% for for small training data, saturates at 160K; good for unknown words/phrases

Previous Work (2)

- Noun phrase rewriting (Nakov, 2008)

  (2) *of members of the Irish parliament*

  of irish parliament members

  of irish parliament's members

- Explicit Paraphrasing to make translation easier
  - Source rewriting in RBMT (Shirai et al., 1993)
    - Dependency tree reordering using rules
  - Re-ordering for SMT (Komachi et al., 2006)
    - Chunk reordering using rules

Our Approach

- Attempt to build new bitext pairs
  - replace one side with
    grammatical, semantically equivalent variants
- Parse to a structural meaning representation:
  MRS (Minimal Recursion Semantics)
  rank with stochastic model
- Generate from 1-best MRS
  rank with stochastic model
  select top n (up to 10)

English Paraphrases

- Using the English Resource Grammar (Flickinger, 2000)
- Parse with PET; generate with LKB
- Various kinds of variation:
  - Phrase order
  - Closed class words: *everyone, everybody*
  - Contractions: *going to* vs *gonna*
  - Numbers: *three* vs *3*
  - Correction: *I read the the book* vs *I read the book*
  - Punctuation

Why is this useful?

- Consider: *going to* (main verb) vs *gonna* (auxilliary)
  - *I am going to the store* (200)
  - *I am going to cry* (600) vs *I am gonna cry*
  - *I am gonna cry* (9) vs *I am going to cry*
- Paraphrasing disambiguates
- Paraphrasing helps with scarcity
- More constrained than phrase table paraphrasing

Experiment

- Paraphrasing (English)
  ERG grammar, pet parser, lkb generator
  DELPH-in: Deep Linguistic Processing with HPSG Initiative

- SMT system: Moses (Koehn et al., 2007)
  replacing giza++ with mgiza

- Corpora
  - Tanaka Corpus (EJ) (2005 version)
    147,190 training, 4,500 dev, 4,500 test
  - IWSLT corpus (EJ) (2005 version)
    42,699 training, 2,108 dev, 500 test (different set)

Semantic Representation

Everybody often goes to the the movies.

$$
\left\langle\, h_1,\ \left|\begin{array}{l}
h_3{:}\text{person}\,(\text{ARG0}\ x_4\{\text{PERS}\ 3, \text{NUM}\ sg\}),\\
h_5{:}\text{every\_q}\,(\text{ARG0}\ x_4,\ \text{RSTR}\ h_6,\ \text{BODY}\ h_7),\\
h_8{:}\text{\_often\_a\_1}\,(\text{ARG0}\ e_9\{\text{TENSE}\ untensed\},\ \text{ARG1}\ e_2\{\text{TENSE}\ pres\}),\\
h_8{:}\text{\_go\_v\_1}\,(\text{ARG0}\ e_2,\ \text{ARG1}\ x_4),\\
h_8{:}\text{\_to\_p}\,(\text{ARG0}\ e_{10}\{\text{TENSE}\ untensed\},\ \text{ARG1}\ e_2,\ \text{ARG2}\ x_{11})\\
h_{12}{:}\text{\_the\_q}\,(\text{ARG0}\ x_{11},\ \text{RSTR}\ h_{14},\ \text{BODY}\ h_{13}),\\
h_{15}{:}\text{\_movie\_n\_of}\,(\text{ARG0}\ x_{11},\ \text{ARG1}\ i_{16}\{\text{SF}\ prop\})
\end{array}\right|\\
\{\, h_6 =_q h_3,\ h_{14} =_q h_{15} \,\}\,\right\rangle
$$

Paraphrases

Everybody often goes to the the movies.

Everyone often goes to the movies.	7.7
Everybody often goes to the movies.	7.7
Everyone goes often to the movies.	0.5
Everybody goes often to the movies.	0.5
Everyone goes to the movies often.	-0.3
Everybody goes to the movies often.	-0.3

For the Tanaka Corpus, 83.4% could be paraphrased.

# distinct paraphrases	1	2	3	. . .	10
% of sentences	53.4	31.2	21.2		1.1

Corpus Expansion

- (d)istributed: rotate between the original sentence and each paraphrase until the data has been padded out
- (f)irst: after all paraphrases have been used, the first (original) sentence is repeated to pad out the data
- (v)arying: add just the paraphrases (always does worse)

d	e_0	e_1	e_2	e_0	e_1
f	e_0	e_1	e_2	e_0	e_0
v	e_0	e_1	e_2		

Table 1: Paraphrase distributions ($n = 4, m = 2$)

Results (TC-EJ)

Lang Pair	Corpus	Paraphrases Added	Bleu	Variance	Delta
EJ	Tanaka Corpus	0	25.96	±0.71	-
EJ	Tanaka Corpus	d.2	26.10	±0.74	+0.14
EJ	Tanaka Corpus	**d.4**	**26.25**	±0.71	**+0.29**
EJ	Tanaka Corpus	**d.6**	**26.63**	±0.72	**+0.67**
EJ	Tanaka Corpus	d.8	26.16	±0.71	+0.20
EJ	Tanaka Corpus	**f.4**	**26.28**	±0.73	**+0.32**
EJ	Tanaka Corpus	f.6	26.13	±0.68	+0.17
EJ	Tanaka Corpus	f.8	25.83	±0.65	-0.13

Results (TC-JE)

Lang Pair	Corpus	Paraphrases Added	Bleu	Variance	Delta
JE	Tanaka Corpus	0	18.75	±0.82	-
JE	Tanaka Corpus	**d.2**	**19.09**	±0.74	**+0.34**
JE	Tanaka Corpus	d.4	18.42	±0.79	-0.33
JE	Tanaka Corpus	d.6	18.71	±0.83	-0.04
JE	Tanaka Corpus	d.8	18.90	±0.77	+0.15
JE	Tanaka Corpus	f.4	18.92	±0.81	+0.17
JE	Tanaka Corpus	f.6	19.02	±0.80	+0.27
JE	Tanaka Corpus	**f.8**	**19.19**	±0.82	**+0.44**

Tanaka Corpus

- EJ: Significant improvements overall (up to +.67)
- JE: Some improvement, some degradation
- Hard to decide how many paraphrases to use
- Should probably try to select good paraphrases
- More consistent, but worse absolute, results with older Moses

IWSLT

- EJ: less improvement over all (up to +.61)
- JE: non-significant improvement
- Probably because of more paraphrases in the original corpus
- More consistent, but worse absolute, results with older Moses

Why we think it works

- We can improve precision for most tasks by adding another knowledge source
- Paraphrasing with the HPSG grammar allows the system to generalize over variants
- May also make alignment easier for some word orders
- But we do not generate all variants
- And we do not rank so well (lm from tourism coprus)

Future Work (1)

- Explicitly weight the equivalents
  - n paraphrases with weight $1/n$ or $p(s)$
- Also do Japanese (using Jacy)
- Retrain parse/generation ranking models (need treebanks)
- Only take paraphrases with score above a threshold

Future Work (2)

- Add an En-EN transfer step
  - NP rewriting
  - Idiom $\leftrightarrow$ literal
  - Active $\leftrightarrow$ passive
  - Lexical paraphrases
- Combine with a Ja-Ja transfer step
  - Statement $\leftrightarrow$ question
  - Positive $\leftrightarrow$ negative
- Use to normalize (All $going\ to$ to $gonna$?)

Resource Availability

We have release the paraphrased corpus. It should be possible to reproduce the results.

- Paraphrased Tanaka Corpus
  `http://www2.nict.go.jp/x/x161/en/member/bond/data/`.
  - Tanaka Corpus (2005 version)
    (2008 version $\approx 12\%$ corrected) — coming soon
  - English grammar, parser generator (ERG, pet, lkb)
    DELPH-IN: `www.delph-in.net`
- SMT system (Moses, mgiza++ and dependencies)
  - Ubuntu NLP repositories (packaged by Eric Nichols)

Conclusion

- We can improve the quality of SMT-based translations
  - by automatically creating more training data
- We create more training data by paraphrasing one side
  - Parse to semantic representation (MRS)
    select the most plausible interpretation
  - Generate all sentences with the same meaning
    select the (n-most) fluent sentences
- The resources needed to do this are publicly available

Comparison

- Resources needed
  - A different bitext (Callison-Burch et al., 2006)
  - An HPSG grammar (En, Ja, De, No, Es, Pt, Gr)
- Effectiveness
  - Good for small corpus (Callison-Burch et al., 2006)
  - Small constant improvement (source language paraphrase)
- No reason not to combine the two methods

Bibliography

References

Chris Callison-Burch, Phillip Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24, 2006.

Dan Flickinger. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28, 2000. (Special Issue on Efficient Processing with HPSG).

Philipp Koehn, Wade Shen, Marcello Federico, Nicola Bertoldi, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Ondrej Bojar, Richard Zens, Alexandra Constantin, Evan Herbst, Christine Moran, and Alexandra Birch. Moses: Open source toolkit for statistical machine

translation. In *Proceedings of the ACL 2007 Interactive Presentation Sessions*, Prague, 2007. URL `http://www.statmt.org/moses/`.

Mamoru Komachi, Yuji Matsumoto, and Masaaki Nagata. Phrase reordering for statistical machine translation based on predicate-argument structure. In *Proceedings of IWSLT 2006*, 2006.

Preslav Nakov. Improved statistical machine translation using monolingual paraphrases. In *Proceedings of the European Conference on Artificial Intelligence (ECAI'08)*, Patras, Greece, 2008.

Satoshi Shirai, Satoru Ikehara, and Tsukasa Kawaoka. Effects of automatic rewriting of source language within a Japanese to English MT system. In *Fifth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-93*, pages 226–239, Kyoto, 1993.