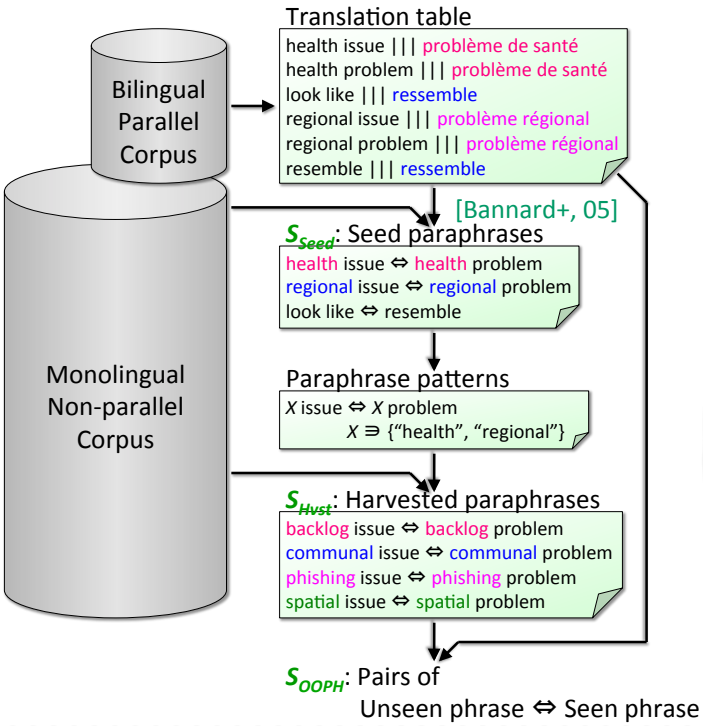
Compilation of Large-scale Paraphrase Lexicon and Its Application to Phrase-based SMT Systems

Atsushi Fujita (*fujita@paraphrasing.org*)
National Institute of Information and Communications Technology, NICT, Japan

Summary

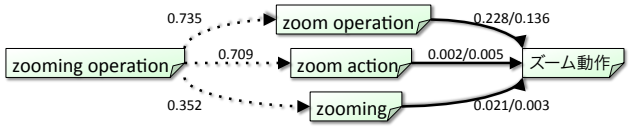
- Standard phrase-based SMT systems + paraphrases
  - Phrase table augmentation (translation pair fabrication)
- Better performance over a vanilla phrase-based SMT

Paraphrase Collections

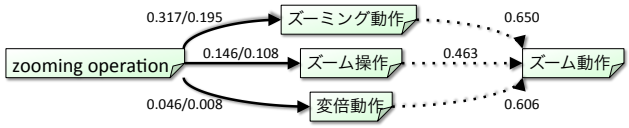


Aggregation of Multiple Paths

- Source-side augmentation



- Target-side augmentation



Features for Decoding

- Scores of individual translation pair
  - Kneser-Ney estimates (original pairs) [Chen+, 11]
  - Translation score (fabricated pairs)
    - e.g., Backward score of source-side augmentation

$$p(s'|t) = \frac{\sum_{s \in S}\left(p(s|t)Para(s \Rightarrow s')\right)}{\sum_{s \in S} Para(s \Rightarrow s')}$$

  - Zens-Ney lexical estimates (all pairs) [Zens+, 04]
- Translation pair indicators
  - Alignment indicators (IBM2, HMM, IBM4)
  - Paraphrase collection indicators (S_{Seed}, S_{Hvst}/S_{OOPH})
  - OOPH indicators (phrase-table level, corpus-level)
- Paraphrase score: an avg. score of each pair
  - PivProb [Bannard+, 05] vs. Cosine similarity [Marton+, 09]

Exercise in Ja/En patent translation

- NTCIR-10 PatentMT [Goto+, 13]

Bilingual Parallel Corpus

	En	Ja
# of snts.	3.2M	3.2M
# of words	106M	116M

Monolingual Non-Parallel Corpus

	En	Ja
# of snts.	413M	594M
# of words	13.4B	27.3B

extraction → # of translation pairs

	Ja → En	En → Ja
IBM2	9.1M	9.4M
HMM	230.6M	234.4M
IBM4	80.6M	81.8M
Union	260.4M	264.8M

filtering / expansion / filtering → # of paraphrase pairs

	th_p	th_s	En	Ja
S_{Seed}	ε	ε	7.2M	5.1M
S_{Seed}	0.01	0.1	1.1M	0.8M
S_{Hvst}	0.01	ε	272M	143M
S_{Hvst}	0.01	0.1	?????	?????

Computed only for tune+dev+test

Model selection using held-out data (ntc7 & ntc8)

(*) minimal set covering tune+dev+test datasets

System	Para score	Ja → En # of trans. pairs (*)	Ja → En BLEU		En → Ja # of trans. pairs (*)	En → Ja BLEU	
vanilla PBSMT	-	18.0M	33.30		15.5M	37.64	
Saug-S_{Seed}	PivProb	27.3M	33.65	+0.35	24.6M	37.98	+0.34
Saug-S_{Seed}	Cosine	27.3M	33.27	-0.03	24.6M	37.73	+0.09
Saug-S_{Hvst}	Cosine	23.6M	33.22	-0.08	22.0M	37.89	+0.25
Saug-S_{OOPH}	Cosine	18.1M	33.72	+0.42	15.6M	38.16	+0.52
Saug-S_{Seed}+S_{Hvst}	Cosine	32.8M	32.91	-0.39	30.9M	37.76	+0.12
Taug-S_{Seed}	PivProb	22.9M	33.34	+0.04	19.6M	37.64	+0.00
Taug-S_{Seed}	Cosine	22.9M	33.56	+0.26	19.6M	38.19	+0.55
Taug-S_{Hvst}	Cosine	29.1M	33.43	+0.13	26.8M	37.98	+0.34
Taug-S_{OOPH}	Cosine	23.4M	33.21	-0.09	21.5M	38.08	+0.44
Taug-S_{Seed}+S_{Hvst}	Cosine	33.9M	32.99	-0.31	30.8M	37.53	-0.11

Open test (ntc10)

System	Ja → En BLEU	Ja → En NIST	Ja → En RIBES	En → Ja BLEU	En → Ja NIST	En → Ja RIBES
vanilla PBSMT	31.25	8.1945	0.6892	33.88	8.2005	0.7021
Saug-S_{OOPH}	31.56	8.2507	0.6955	34.22	8.2345	0.7096
Taug-S_{Seed}	31.65	8.2198	0.6929	34.05	8.2116	0.7089

Conclusion: paraphrase-base augmentation works!

- The improvement is small, though
- Consistent improvement on various distortion limit values
- F/W: integration of further gigantic paraphrase resources

Publication

- Atsushi Fujita, Pierre Isabelle, and Roland Kuhn. Enlarging Paraphrase Collections through Generalization and Instantiation. In Proc. of EMNLP-CoNLL, 2012.
- Atsushi Fujita and Marine Carpuat. FUN-NRC: Paraphrase-augmented Phrase-based SMT Systems for NTCIR-10 PatentMT. In Proc. of NTCIR-10, 2013.