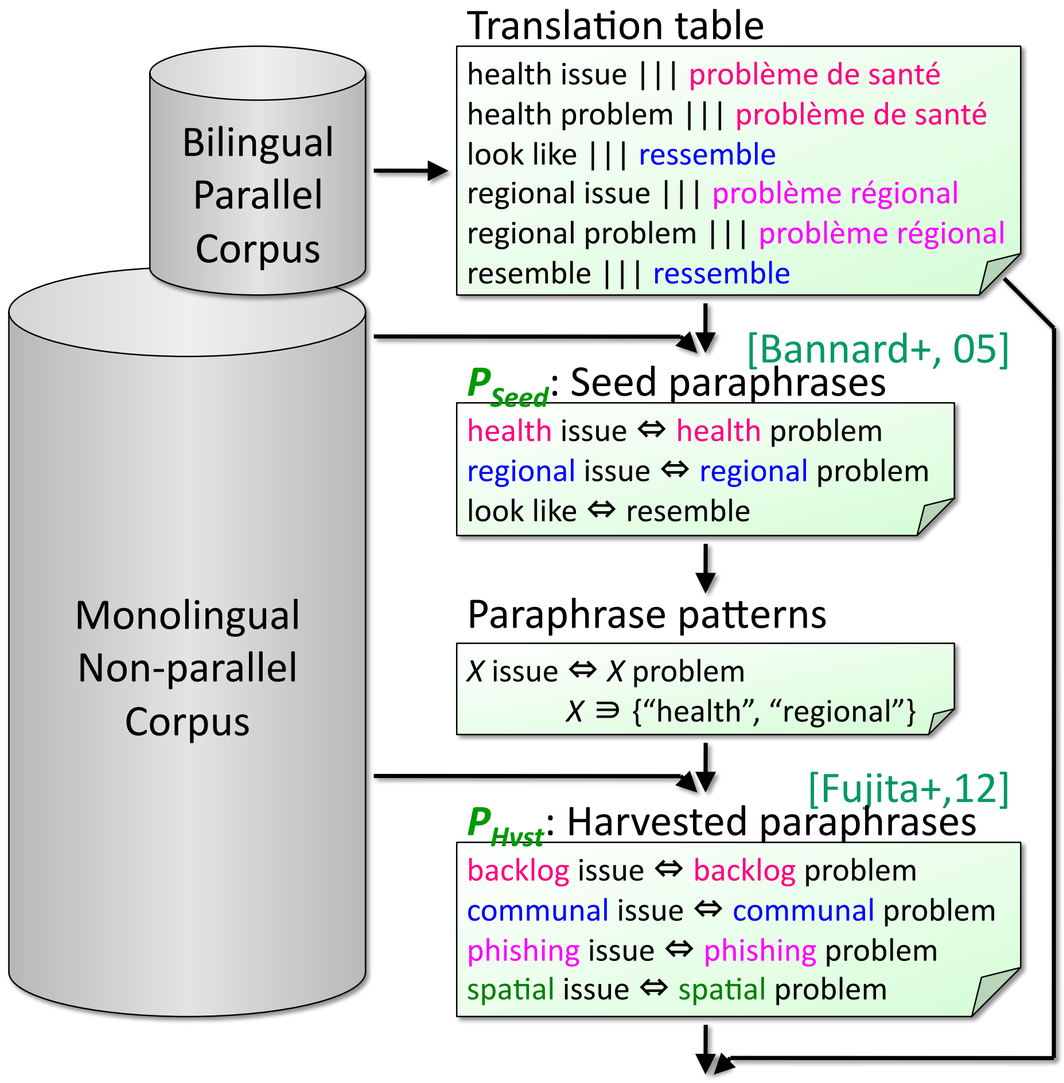
FUN-NRC: Paraphrase-Augmented Phrase-Based SMT Systems for NTCIR-10 PatentMT

Atsushi Fujita (Future University Hakodate) *fujita@fun.ac.jp*
Marine Carpuat (National Research Council Canada) *marine.carpuat@nrc.gc.ca*

Summary

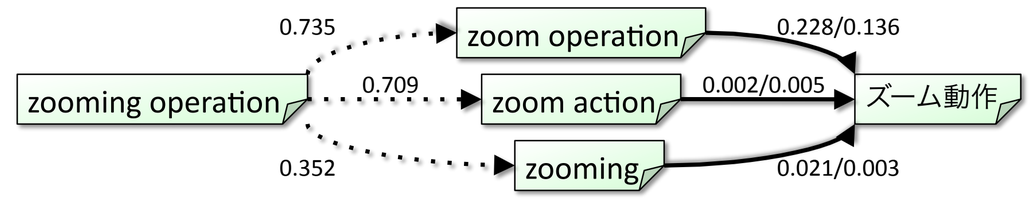
- Standard phrase-based SMT systems + paraphrases
  - Phrase table augmentation (translation pair fabrication)
- Better performance over a vanilla phrase-based SMT

Study 1. Paraphrase Collections

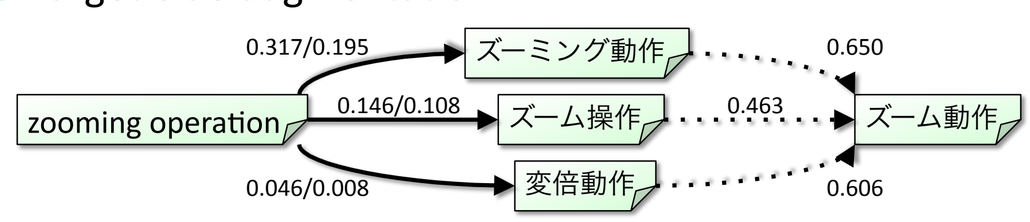


Study 2. Aggregation of Multiple Paths

- Source-side augmentation



- Target-side augmentation



Study 3. 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 (P_{Seed}, P_{Hvst}/P_{OOPH})
  - OOPH indicators (at phrase-table level, at corpus-level)
- Paraphrase score: an avg. score of each pair involved
  - PivProb [Bannard+, 05] vs. Cosine similarity [Marton+, 09]

Development

of trans. 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

extraction → filtering → expansion → filtering

of paraphrase pairs

	th_p	th_s	En	Ja
P_{Seed}	0	0	7.2M	5.1M
P_{Seed}	0.01	0.1	1.1M	0.8M
P_{Hvst}	0.01	0	272M	143M
P_{Hvst}	0.01	0.1	?????	?????

dev&test data driven

Model selection using held-out data (ntc7 & ntc8)
(Minimal phrase table is created)

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

Official Results

System	Ja → En BLEU	Ja → En NIST	Ja → En RIBES	En → Ja BLEU	En → Ja NIST	En → Ja RIBES
Saug-P_{OOPH}	31.56	8.2507	0.6955	34.22	8.2345	0.7096
Taug-P_{Seed}	31.65	8.2198	0.6929	34.05	8.2116	0.7089
*Const-Saug-P_{Hvst}	30.58	8.1114	0.6911	32.89	8.0977	0.7048
*Const mixLM	30.65	8.1400	0.6906	~~22.59~~	~~7.1185~~	~~0.6651~~
				33.03	8.1101	0.7051

*Systems built using only bilingual data.

Results w/ Relaxed Distortion Limit