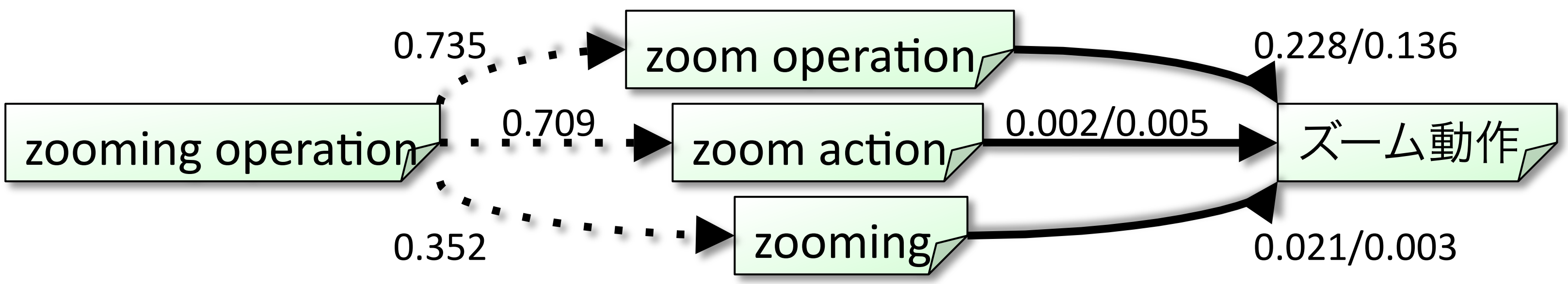
FUN-NRC: Paraphrase-Augmented Phrase-Based SMT Systems for NTCIR-10 PatentMT

Atsushi Fujita
Future **Un**iversity Hakodate
http://paraphrasing.org/~fujita/

Marine Carpuat
National **R**esearch **C**ouncil
http://marinecarpuat.weebly.com/

Summary of our systems

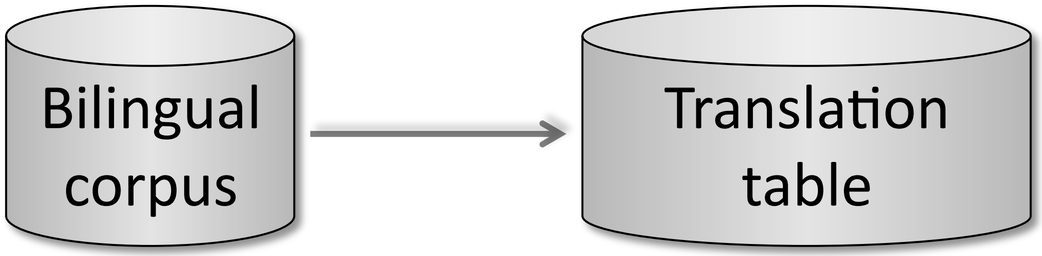
- Phrase-based SMT + paraphrases
  - State-of-the-art non-hierarchical system: PortageII @ NRC
    - Almost no language- or domain- specific knowledge
  - Phrase table augmentation
    - Paraphrases in both source & target languages (separately)
    - Comparison of paraphrase collections
    - Aggregation of multiple paths w/ feature engineering
  - Improved performance over a vanilla phrase-based SMT
    - at least BLEU, NIST, and RIBES



Motivation & proposed method

Modern SMT systems: Limitations

- Principle



- Limitations
  - At source side
    - Unseen expressions will never be translated
    - They are either dropped or retained as is
  - At target side
    - Only seen expressions can be generated as hypotheses
    - cf. Language models only ranks the given hypotheses

Expressions that convey the same meaning

- Paraphrase: monolingual

  Emma burst into tears and he tried to comfort her.

  Emma cried, and he tried to console her.

- Translation: cross-lingual

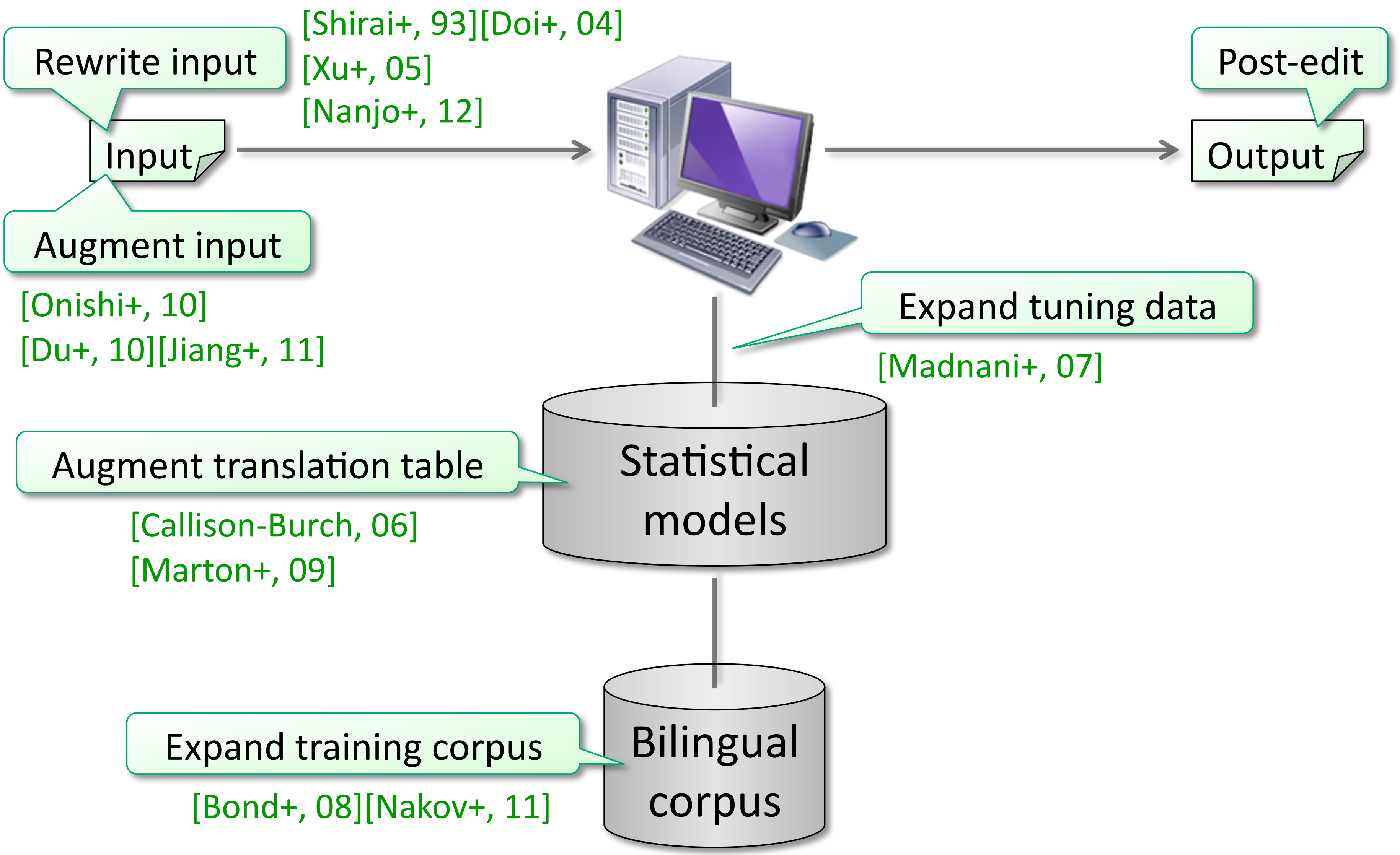
  Désirez-vous obtenir des conseils pratiques sur le déménagement?

  Are you looking for some helpful tips for moving?

Paraphrases

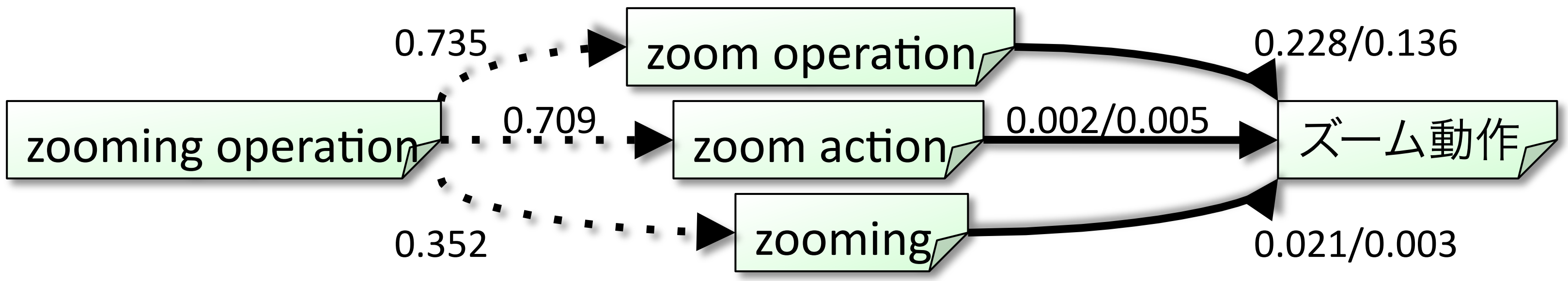
- Linguistic expressions in the same language that convey the same meaning
  - Word / word sequence
    - burst into tears ↔ cried
    - resemble ↔ look like
  - Clause (simple sentence)
    - The car collided with the bicycle ↔ The car and the bicycle collided
  - Beyond single clause
    - It was his best suit that John wore to the dance last night. ↔ John wore his best suit to the dance last night.

Prior arts in integrating paraphrases to MT

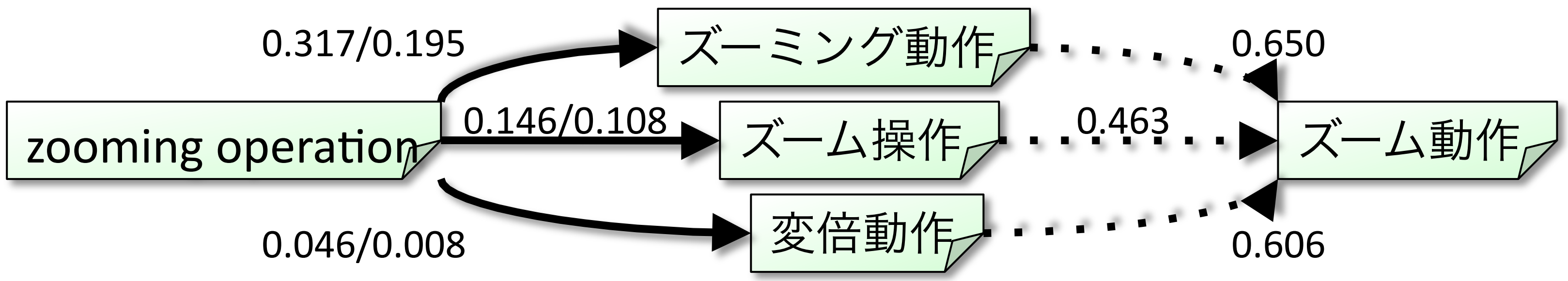


Augmentation of translation table

- Updates from [Callison-burch, 06][Marton+, 09]
  - Comparison of several paraphrase collections
  - Aggregation of multiple paths (both sides)
    - Source side (Saug): translate more phrases



    - Target side (Taug): generate more hypotheses



  - Feature engineering for decoding

Key issue: how to realize paraphrases?

- Large-scale knowledge-base is indispensable
  - ~~Handcrafting~~
  - Automatic paraphrase acquisition (PA)
- Pros. & cons. of prior arts
  - PA from Monolingual non-parallel corpora
    - Pro. Large → (potentially) high recall
    - Con. Only weak evidences → low precision
  - PA from Mono/Bi/Multi-lingual parallel corpora
    - Pro. Sentence-level equivalence → high precision
    - Con. Limited availability → low recall

PA from monolingual non-parallel corpora

- Distributional Hypothesis [Harris, 68]
  - Expressions that appear frequently in similar contexts have similar meanings
  - e.g., “Tezgüno” [Pantel+, 02]

    A bottle of tezgüno is on the table

    Everyone likes tezgüno

    Tezgüno makes you drunk

    We make tezgüno out of corn

    - Similar to wine, cognac, whiskey → alcoholic beverage
    - Con. Not necessarily equivalent: e.g., antonyms, hypernyms

resembles / looks like

tezgüno ≠ wine

decrease ≠ increase

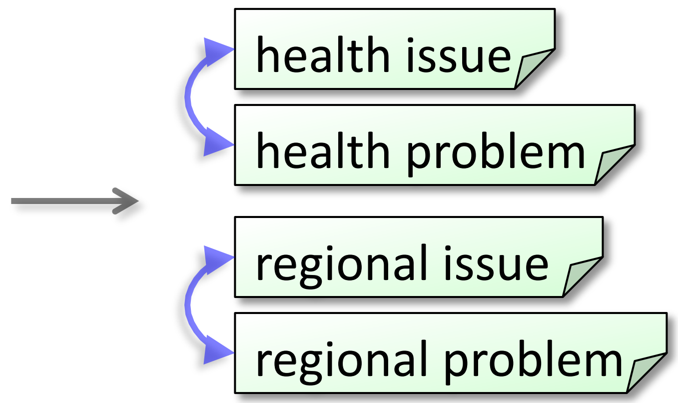
liquor ≠ wine

PA from bilingual parallel corpora

- Translations as pivot [Bannard+, 05]
  - A more reliable evidence than context
  - Obtainable from bilingual parallel corpora
    - i.e., word alignment + phrase extraction

Automatically learned translation table

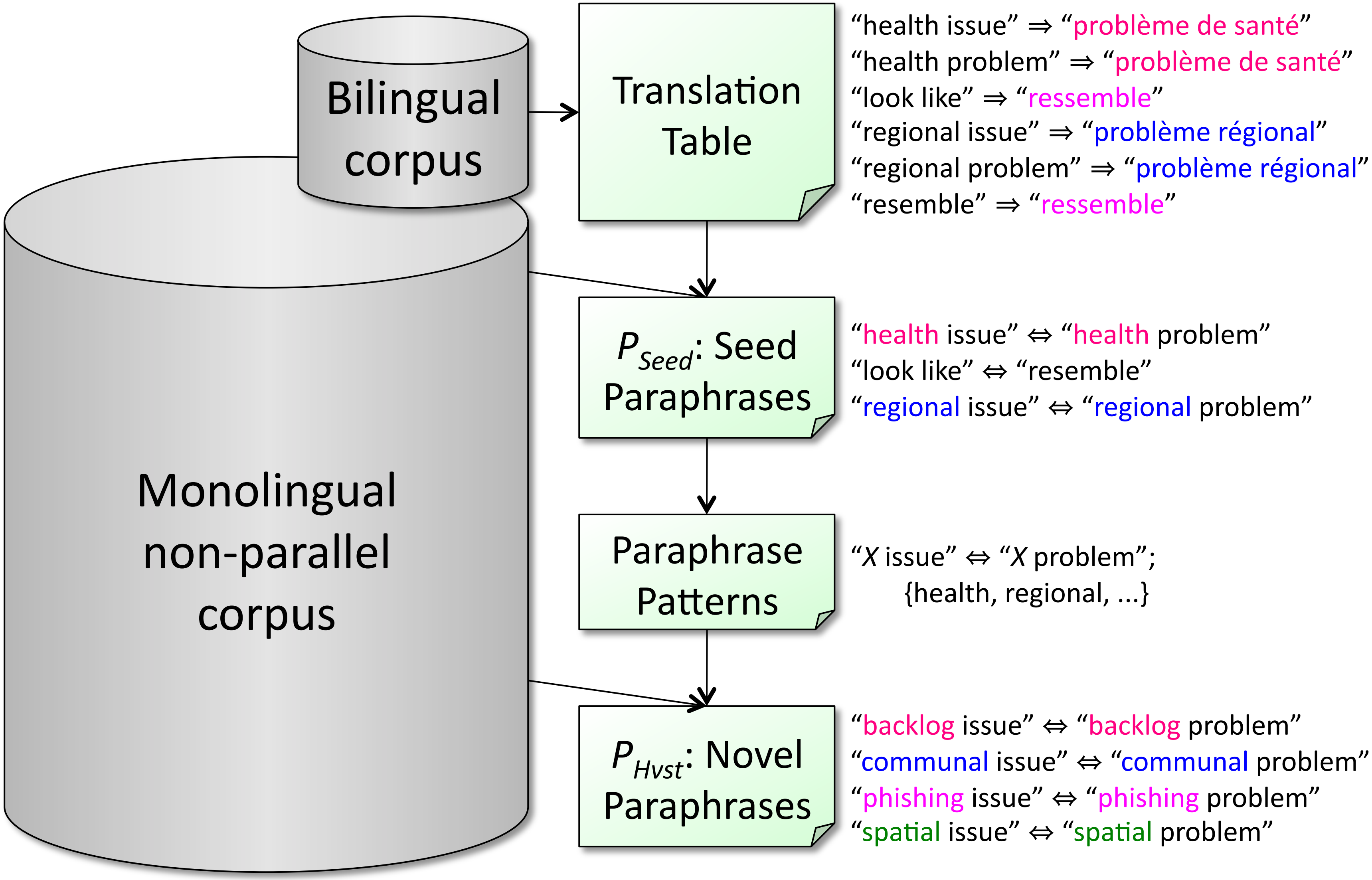
health issue	|||	problème de santé
health problem	|||	problème de santé
regional issue	|||	problème régional
regional problem	|||	problème régional



- Polysemy would generate non-paraphrases
- Con. Parallel corpora << monolingual non-parallel corpora

Paraphrase collections examined [Fujita+, 12]

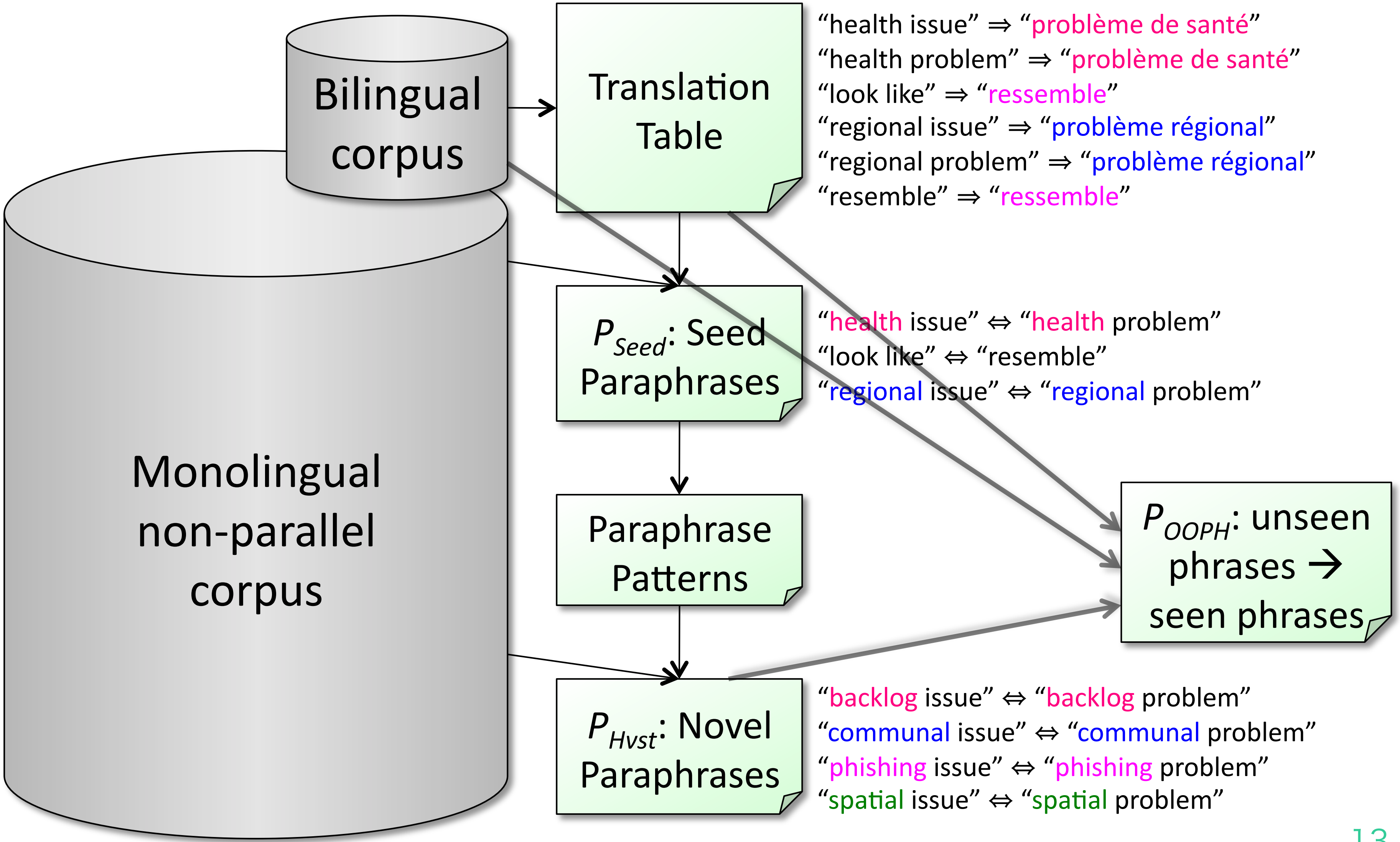
- P_{Seed}, P_{Hvst}, and P_{OOPH}



Paraphrase collections examined

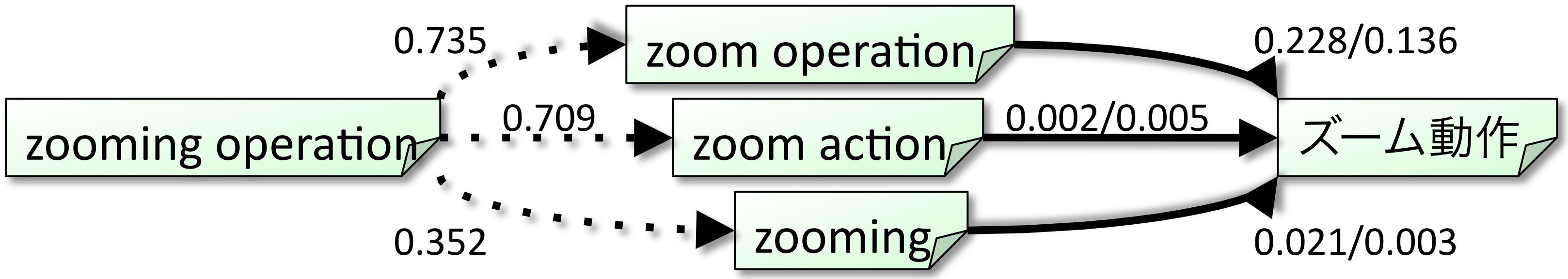
[Fujita+, 12]

- P_{Seed}, P_{Hvst}, and P_{OOPH}



Aggregation of multiple paths (1/2)

- Source-side augmentation



- Translation scores
  - Forward

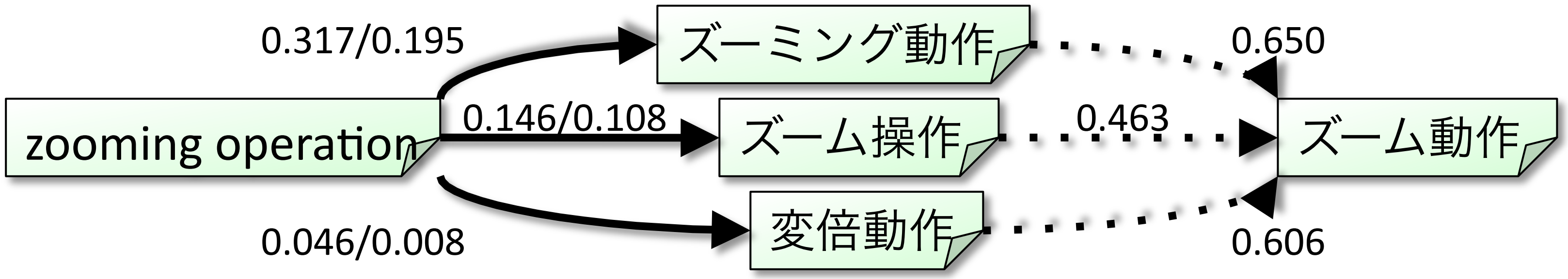
$$p(t|s') = \frac{\sum_{s\in S}\Big(p(t|s)Para(s' \Rightarrow s)\Big)}{\sum_{s\in S} Para(s' \Rightarrow s)}$$

  - Backward

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

Aggregation of multiple paths (2/2)

- Target-side augmentation



- Translation scores
  - Forward

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

  - Backward

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

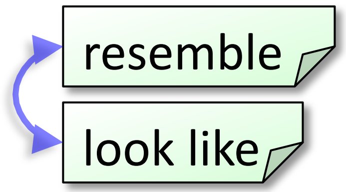
Paraphrase-related Features

Features in the translation model	Original	Source-side fabricated	Target-side fabricated
(a1) Forward translation score	Cond.Prob.	[0,1]	[0,1]
(a2) Backward translation score	Cond.Prob.	[0,1]	[0,1]
(b1) Obtained from IBM2 alignment	True/False	False	False
(b2) Obtained from HMM alignment	True/False	False	False
(b3) Obtained from IBM4 alignment	True/False	False	False
(c1) Fabricated using Seed	False	True/False	True/False
(c2) Fabricated using Hvst/OOPH	False	True/False	True/False
(d1) Unseen in the phrase table	False	True/False	True/False
(d2) Unseen in the bilingual data	False	True/False	True/False
(e1) Paraphrase score (Saug/fwd)	1	[0,1]	1
(e2) Paraphrase score (Saug/bwd)	1	[0,1]	1
(e3) Paraphrase score (Taug/fwd)	1	1	[0,1]
(e4) Paraphrase score (Taug/bwd)	1	1	[0,1]

Score of each paraphrase pair (1/2)

- ***PivProb***: Pivot-based paraphrase probability [Bannard+, 05]
  - For P_{Seed} only

s		*t*		*p(s|t)*	*p(t|s)*
look like	|||	ressemble	|||	0.0177	0.0061
resemble	|||	ressemble	|||	0.0074	0.0181

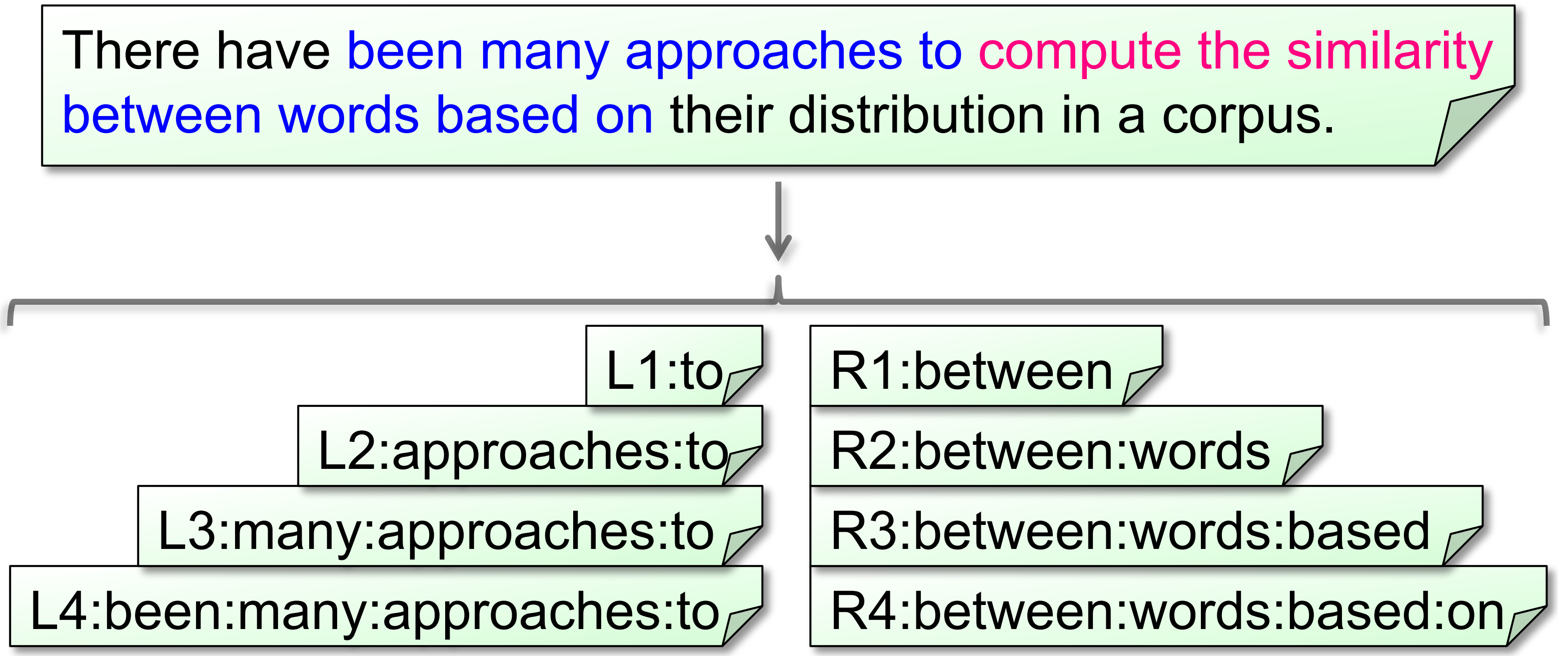


  - Asymmetric score

$$
\begin{aligned}
Para(s_1 \Rightarrow s_2) &= p(s_2|s_1) \\
&= \sum_{t \in tr(s_1) \cap tr(s_2)} p(s_2|t)p(t|s_1)
\end{aligned}
$$

Score of each paraphrase pair (2/2)

- ***CosSim***: cosine similarity of "contexts"
  - For all of P_{Seed}, P_{Hvst}, and P_{OOPH}
  - Contextual similarity in a monolingual corpus
  - Adjacent 1- to 4-grams of each token → feature vector
    - cf. cheap but noisy features, e.g., bag-of-words
    - cf. accurate but expensive features, e.g., dependency trees



Dev & Test

Our base system

- PortageII 1.0 [National Research Council, 12]
  - A state-of-the-art phrase-based SMT system
    - Reasonably good results at NIST OpenMT 2012 [Foster, 12]
  - Advanced features (cf. Moses)
    - Kneser-Ney translation probability smoothing [Chen+, 11]
    - Hierarchical lexicalized reordering [Cherry+, 12]
    - Lattice-batch-MIRA optimization [Cherry & Foster, 12]
    - etc.
  - User-friendly features
    - Highly tuned libraries for using gigantic models [Germann+, 09]
    - High stability (cf. GIZA++)
    - Fits well to cluster computing environment

Training component models

- Provided data
  - Training bi-text
    - 3.2M sentence pairs
  - Monolingual text
    - Ja: 594M sentences (27.3B words)
    - En: 413M sentences (13.4B words)
  - Data for tuning
    - 2000 sentence pairs

- Component models
  - Language models
    - SRILM-5g
  - Translation models
    - IBM2
    - HMM
    - IBM4
  - Reordering models
    - Lexical model
    - Hierarchical lexical model
  - Paraphrase tables

- Parameter tuning

of learned phrasal equivalent pairs

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

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

	th_p	th_s	En	Ja
P_{Seed}	0.01	0.1	0.7M	0.5M
P_{Seed}	0.01	0.1	3.8M	2.7M
P_{Hvst}	0.01	0.1	1.8M	1.5M

dev&test data driven filtering

Avg. BLEU score over held-out data

- On two 2006-2007 dev data (v7, v8)

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

- Human evaluation (Saug-P_{OOPH})

	Ja → En		En → Ja	
	Score	Ranking	Score	Ranking
Adequacy	2.89/5.00	10th/18	2.67/5.00	10th/14
Acceptability	0.43/1.00	8th/9	0.38/1.00	8th/9

- Automatic evaluation

	Ja → En			En → Ja		
System	BLEU	NIST	RIBES	BLEU	NIST	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~~

*Systems built using only bilingual data.

33.03 8.1101 0.7051

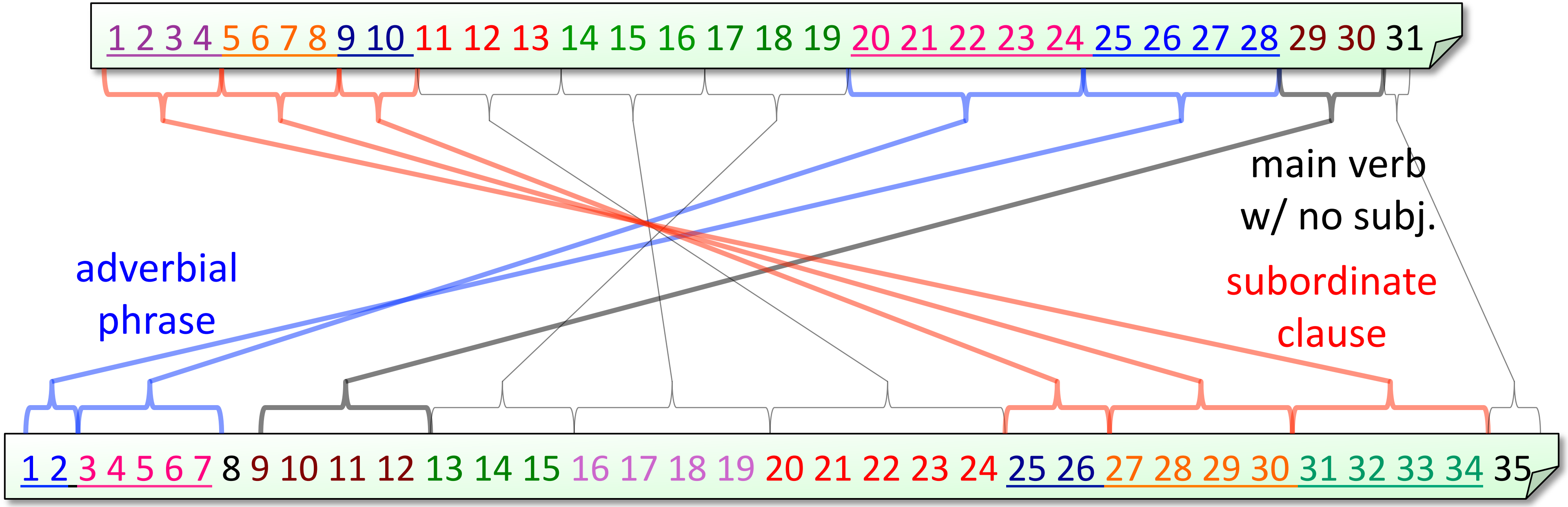
Implications

- Relatively high BLEU and NIST scores
  - Useful n-grams (~ phrases) were generated and selected

- Low RIBES score and human evaluation score
  - Reordering ability was poor
  - Features of superior systems
    - Structure-aware SMT
    - RBMT adapted to the patent domain

We've used 7 for the distortion limit ...

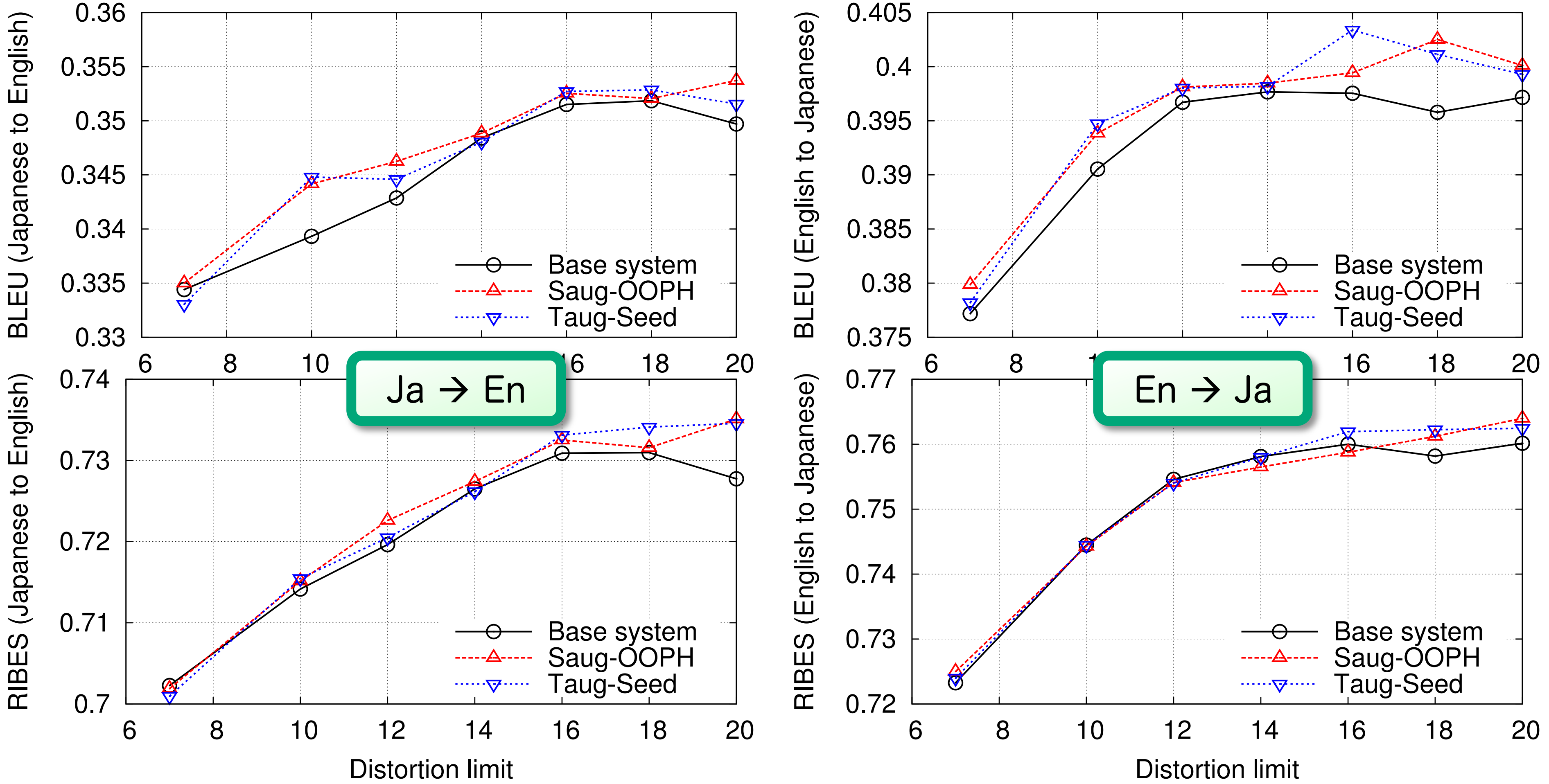
本/実施/形態/の/トレンチ/型/キャパシタ/１２０/を/含む/半導体/装置/の/製造/工程/の/一例/を/図/２/から/図/８/を/参照/し/て/説明/する/。



Referring to FIGS. 2 to 8, description will be given to an example of a manufacturing process of the semiconductor storage device which comprises the trench capacitor 120 according to the embodiment.

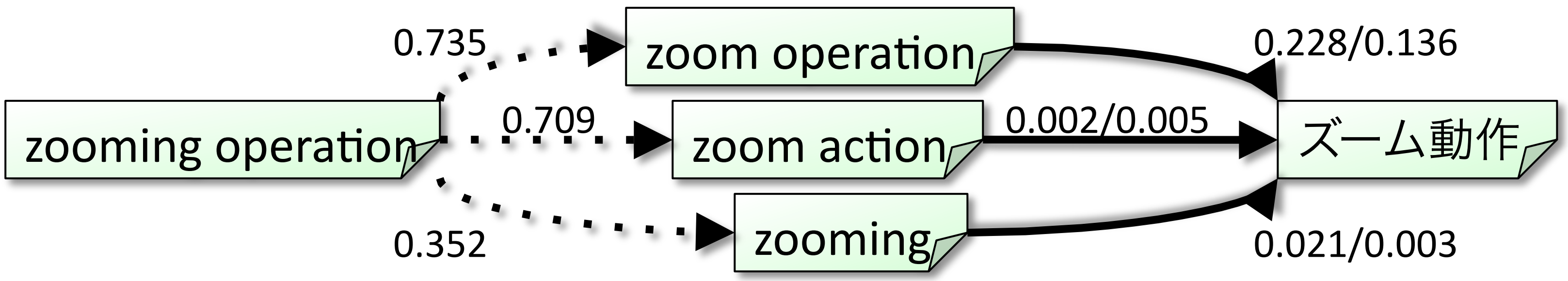
Relaxation of distortion limit

- Held-out data same as development
  - Obtained significantly higher score
  - Positive impact led by paraphrases was retained



Conclusion

- Phrase-based SMT + paraphrases
  - State-of-the-art non-hierarchical system: PortageII @ NRC
    - Almost no language- or domain- specific knowledge
  - Phrase table augmentation
    - Paraphrases in both source & target languages (separately)
    - Comparison of paraphrase collections
    - Aggregation of multiple paths w/ feature engineering
  - Improved performance over a vanilla phrase-based SMT
    - at least BLEU, NIST, and RIBES



Greatest thanks go to

- Supporters of the research program
  - NRC: National Research Council Canada
    - esp. All members in the Portage team
  - FUN: Future University Hakodate
  - JSPS: Japan Society for the Promotion of Science
- PatentMT task organizers