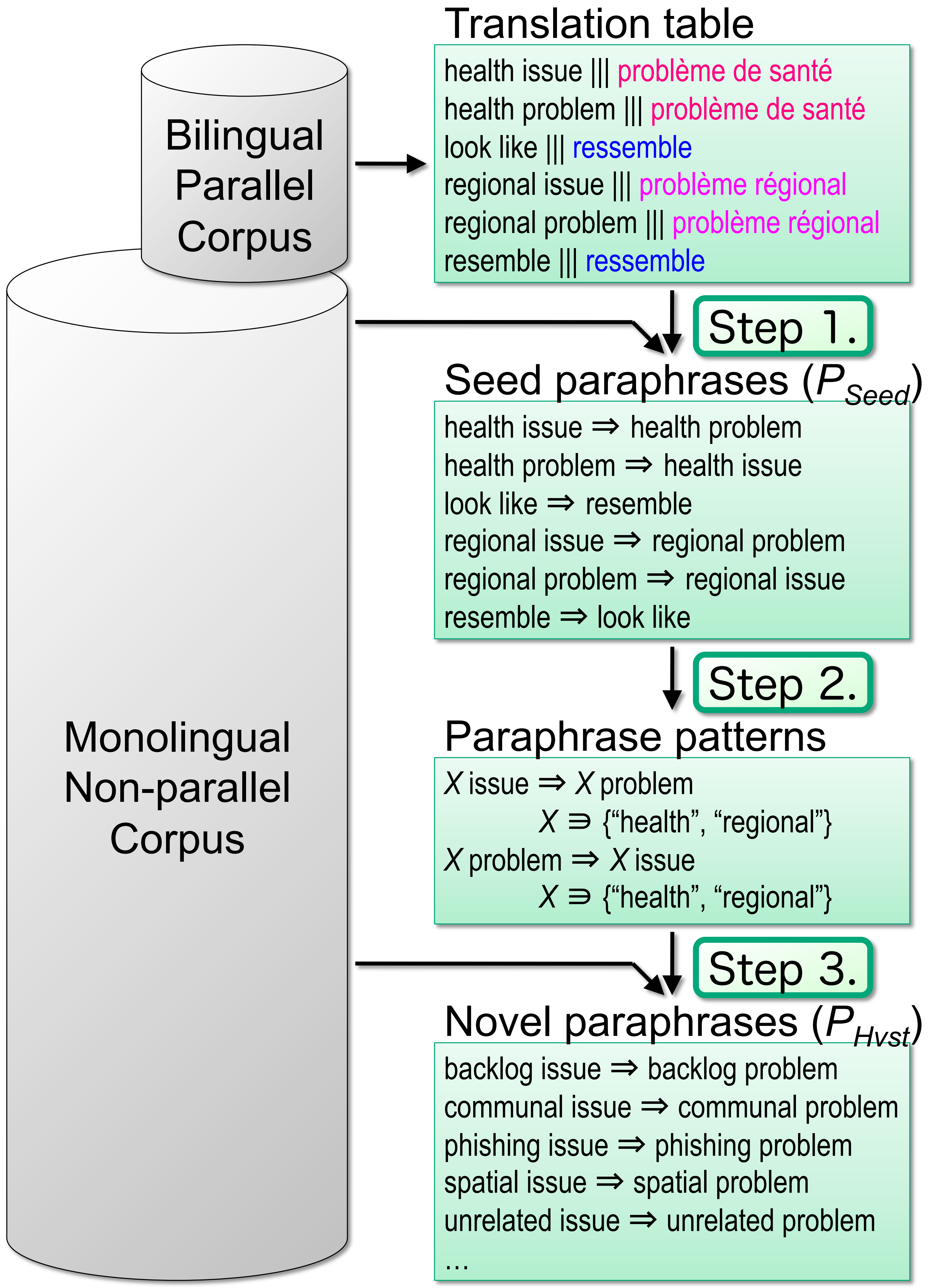
Enlarging Paraphrase Collections through Generalization and Instantiation

Atsushi Fujita (Future University Hakodate) *fujita@fun.ac.jp*
Pierre Isabelle, Roland Kuhn (National Research Council Canada)

Summary

- Paraphrase acquisition
  - Through generalization and instantiation
  - Using both bilingual and monolingual data
- Resources
  - Corpora (bilingual parallel and monolingual)
  - Tokenizer
  - SMT system
  - Lists of stop words
  - (optional) Morphological resources



Bilingual Parallel Corpus

Translation table

health issue ||| problème de santé
health problem ||| problème de santé
look like ||| ressemble
regional issue ||| problème régional
regional problem ||| problème régional
resemble ||| ressemble

Step 1.

Seed paraphrases (P_{Seed})

health issue ⇒ health problem
health problem ⇒ health issue
look like ⇒ resemble
regional issue ⇒ regional problem
regional problem ⇒ regional issue
resemble ⇒ look like

Step 2.

Paraphrase patterns

X issue ⇒ X problem
X ∋ {“health”, “regional”}
X problem ⇒ X issue
X ∋ {“health”, “regional”}

Monolingual Non-parallel Corpus

Step 3.

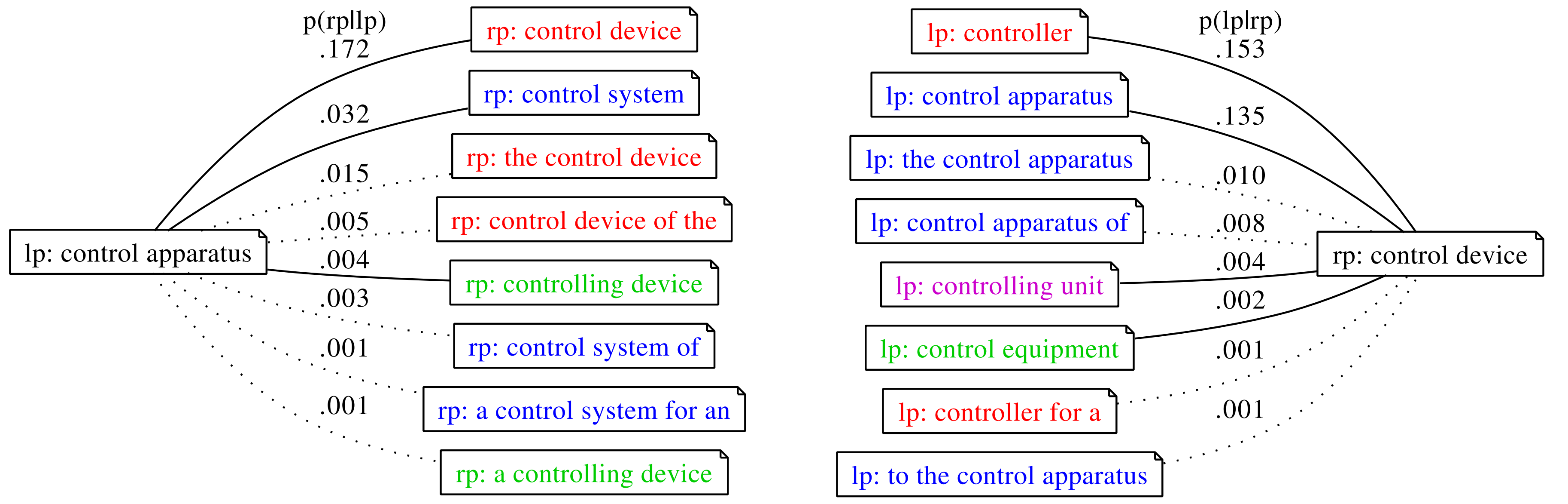
Novel paraphrases (P_{Hvst})

backlog issue ⇒ backlog problem
communal issue ⇒ communal problem
phishing issue ⇒ phishing problem
spatial issue ⇒ spatial problem
unrelated issue ⇒ unrelated problem
…

Step 1. Seed Paraphrase Acquisition

- Pivot-based PA using generic SMT systems
  - e.g., Phrase-based SMT system [Koehn, 03]

1. Clean up phrase table: sig. pruning [Johnson+, 07]
2. Pair phrases that get translated to the same phrases [Bannard and Callison-Burch, 05]
3. Filter paraphrase candidate pairs
   - **3a.** stop word differences, word super-sequences
   - **3b.** conditional probability and contextual similarity



Step 2. Paraphrase Pattern Induction

- Identical words of LHS and RHS → Variable slots
  - Ignore morphological variation
    e.g., number (sg./pl.), gender, case, person, tense
- Related work
  - Develop patterns manually [Jacquemin, 99][Fujita+, 07]
  - Add contextual constraints [Callison-Burch, 08][Zhao+, 09]

Step 3. Paraphrase Instance Acquisition

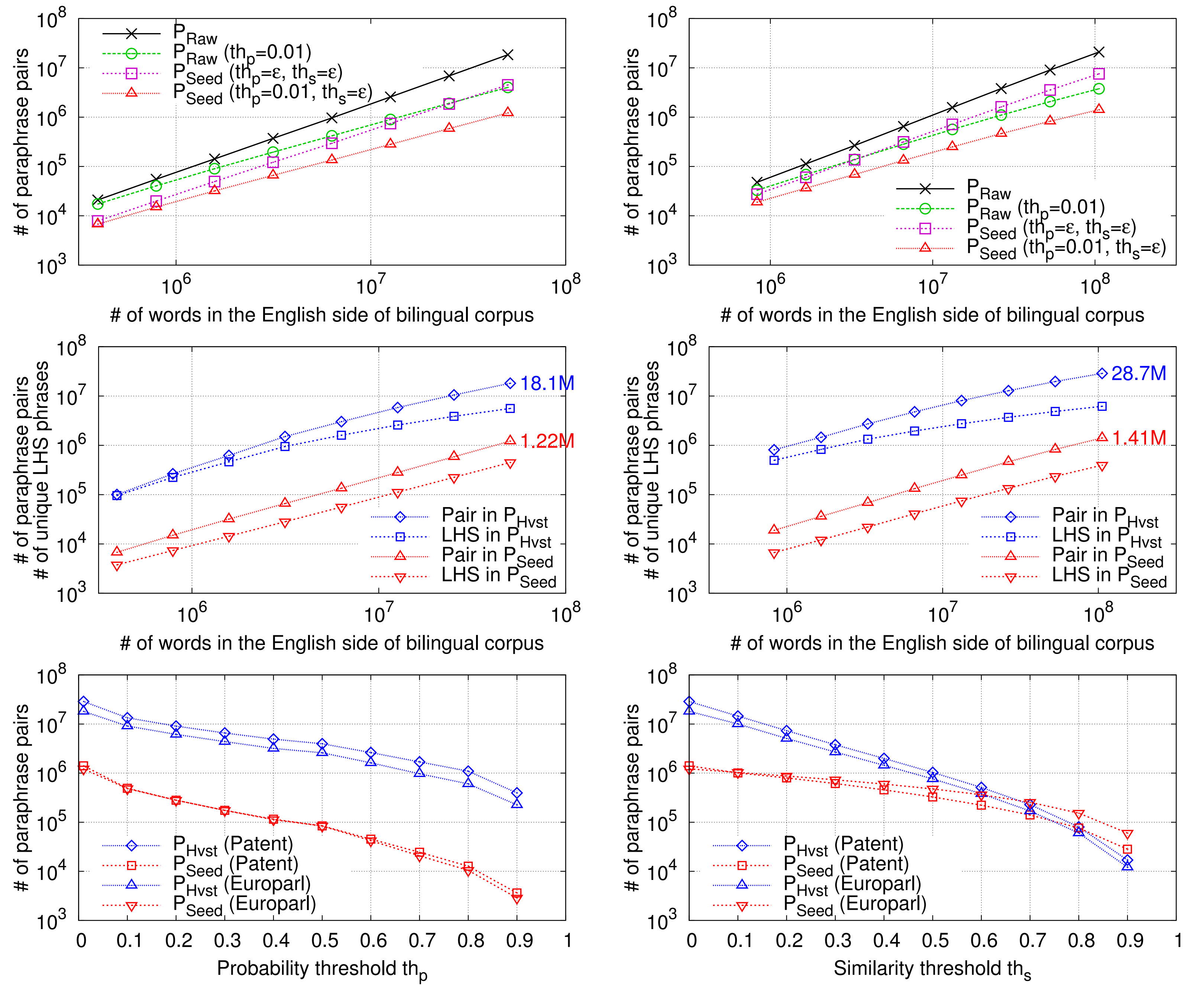
- Harvest novel instances of the patterns

1. Collect expressions that match both sides of the pattern
2. Score each instance by contextual similarity

- Related work
  - Learn class-dependent patterns [De Saeger+, 09;11]
  - (Pattern-dependent) set expansion

Large multiple of # of seeds

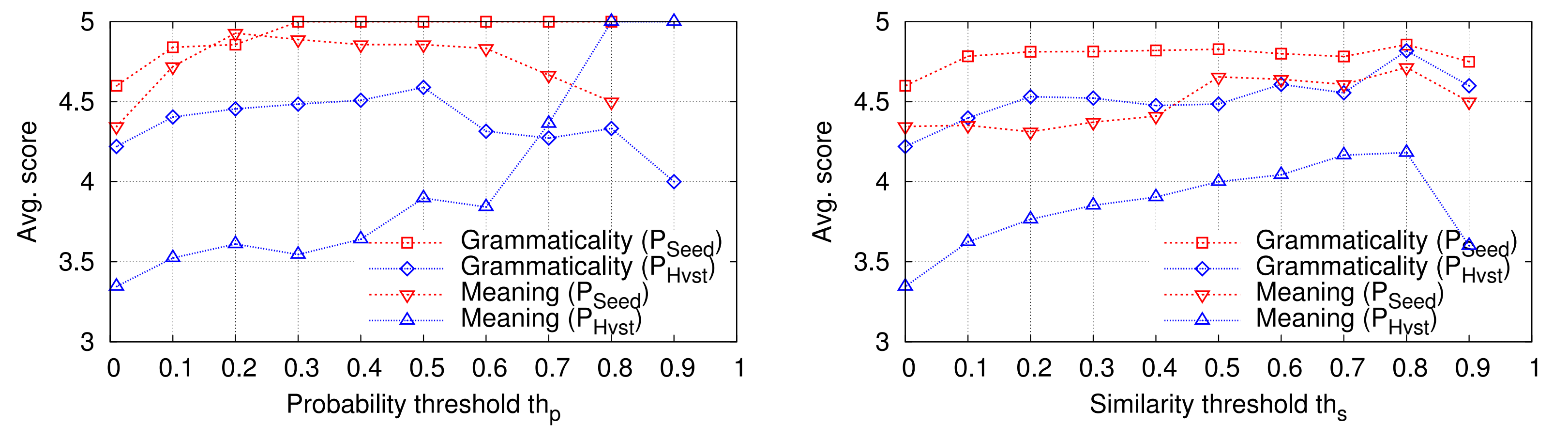
- (A) Europarl + GigaFrEn, (B) NTCIR Patent data
  - One-variable patterns and single words
  - High leverage rate (# pair): ≧1580%, ≧2140%
  - Paraphrases for many novel phrases



Good quality

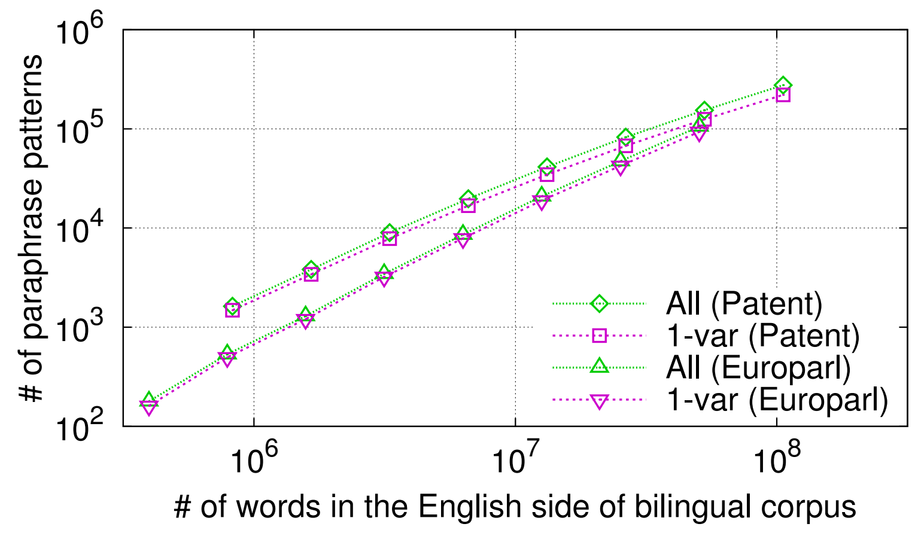
- Human evaluation of phrase substitutions
  - Europarl paraphrases on WMT “newstest” data
  - Comparable to the state-of-the-art

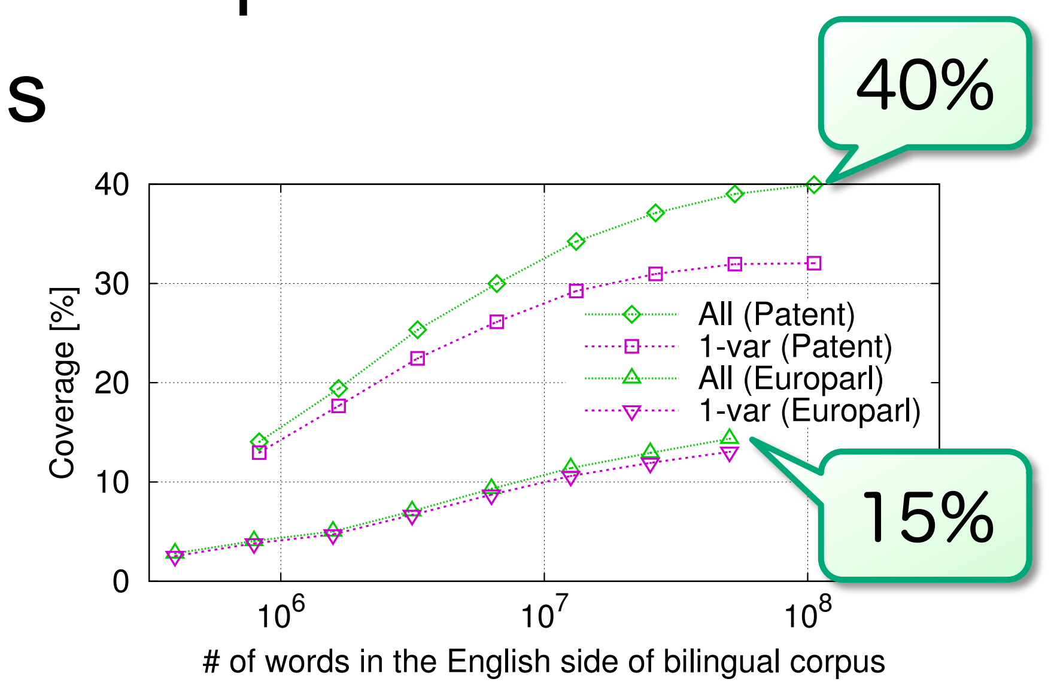
		5-pt		Binary		
	n	G	M	G≥4	M≥3	Both
P_{Seed}	55	4.60	4.35	.85	.93	.78
P_{Hvst}	295	4.22	3.35	.74	.67	.55
Total	350	4.28	3.50	.76	.71	.58



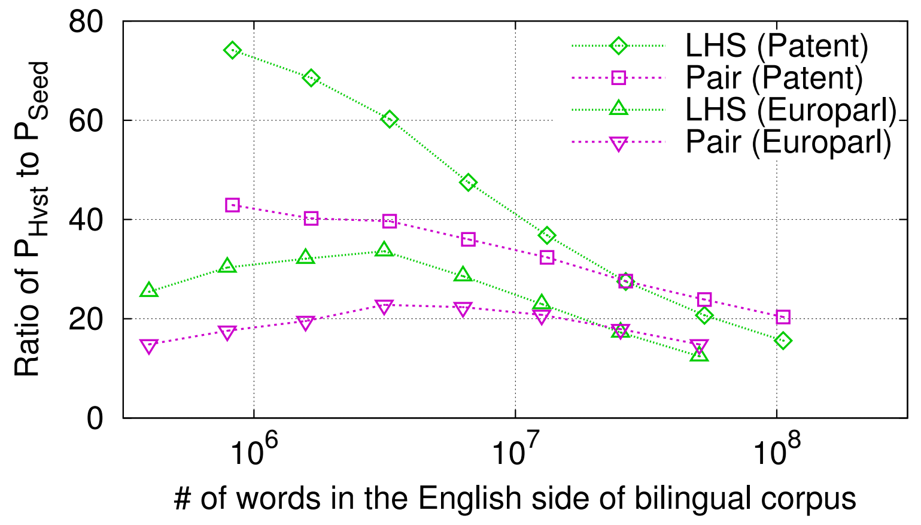
Additional statistics

- Paraphrase patterns
  - Coverage depends on corpus/domain
  - Mostly 1-var patterns

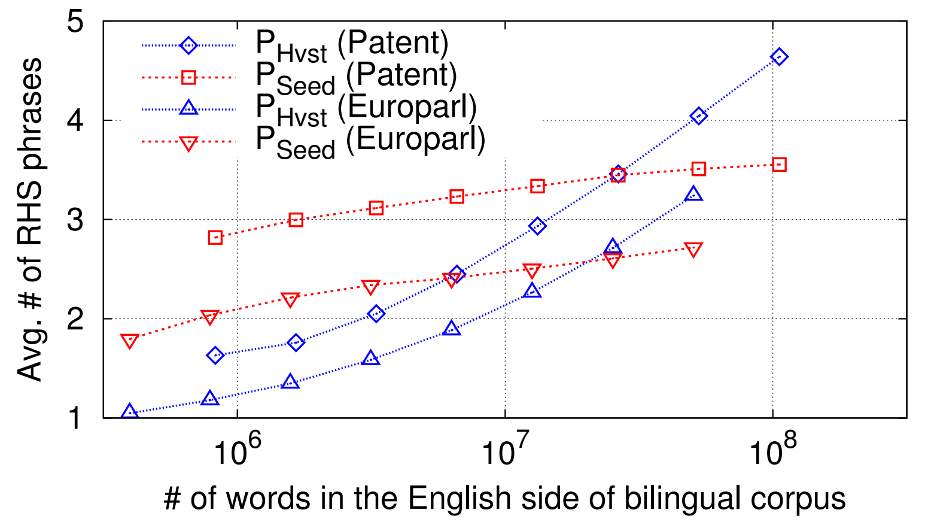




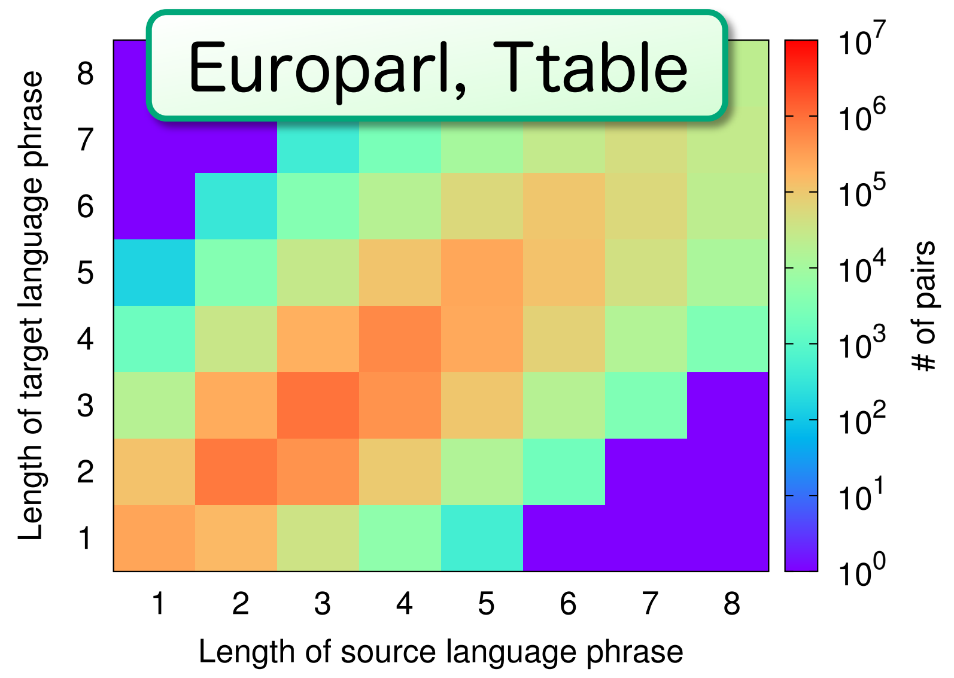
- Leverage rate
  - Small bilingual data
    → High leverage

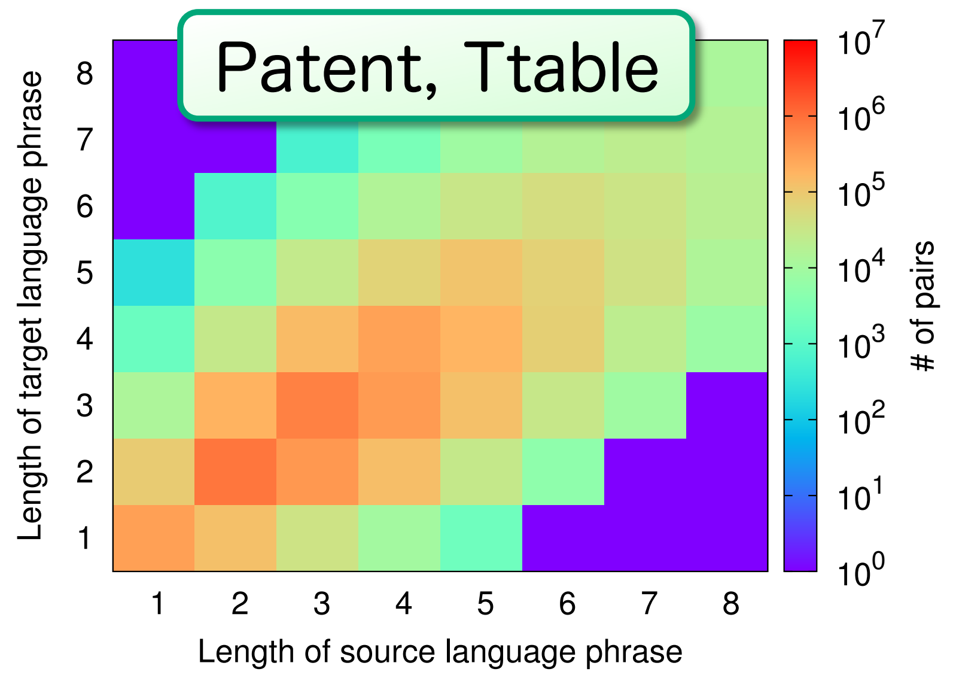


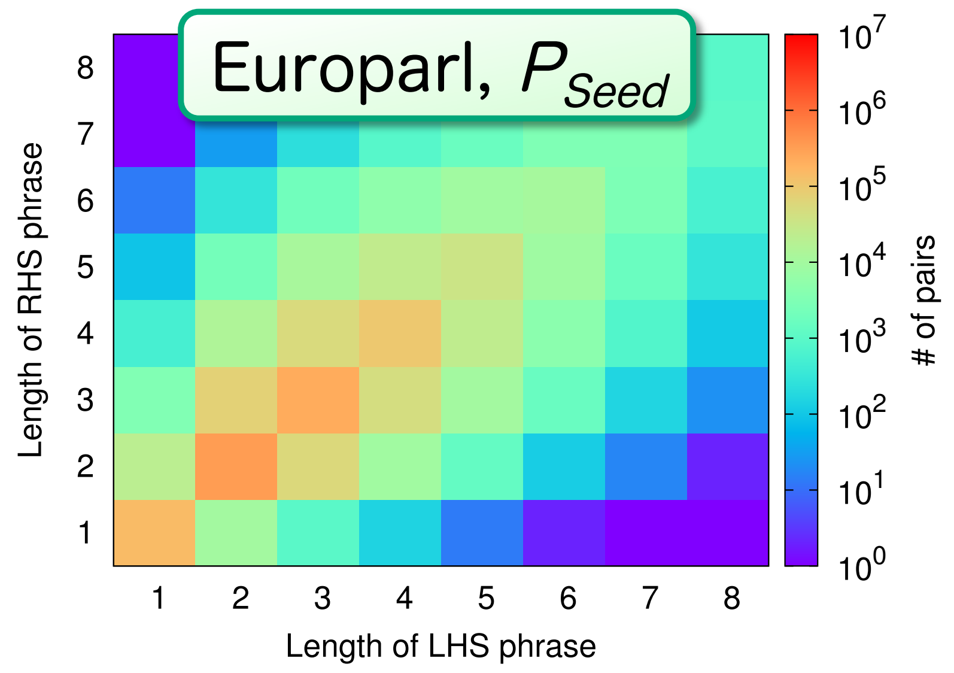
- Relative yield
  - P_{Seed} grows slowly
  - P_{Hvst} depends on patterns

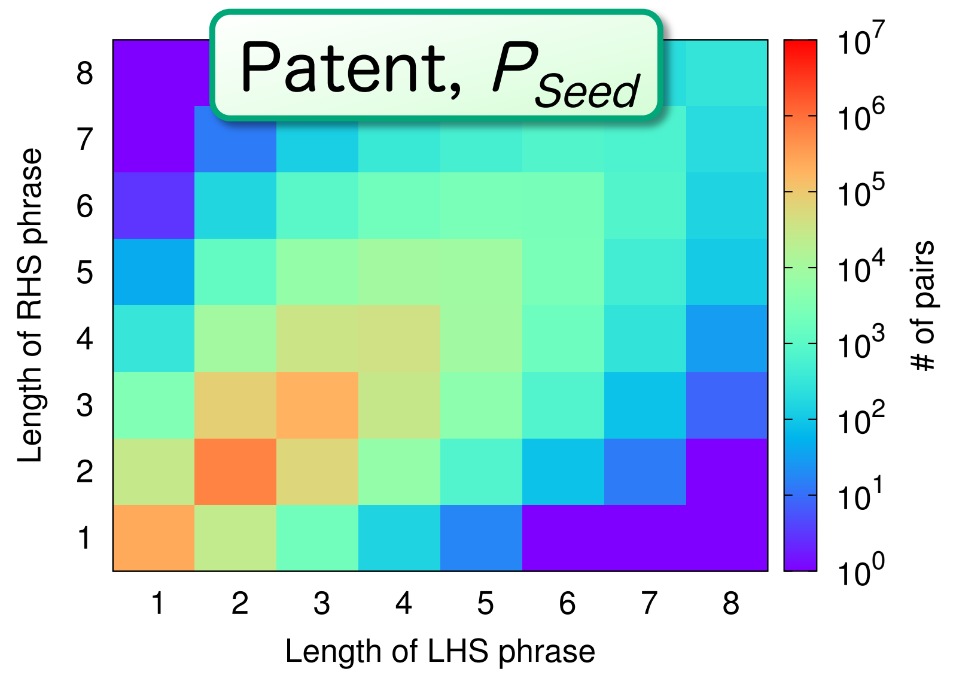


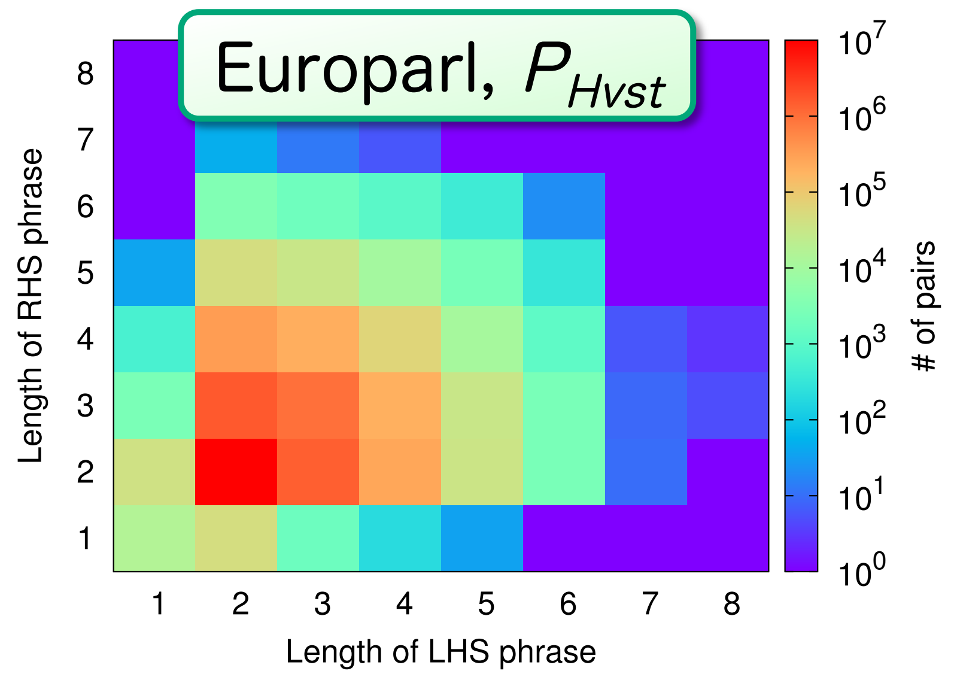
- Phrases tend to be short
  - Our filters tend to discard long phrases
  - Setting: 1-var patterns & single-word fillers

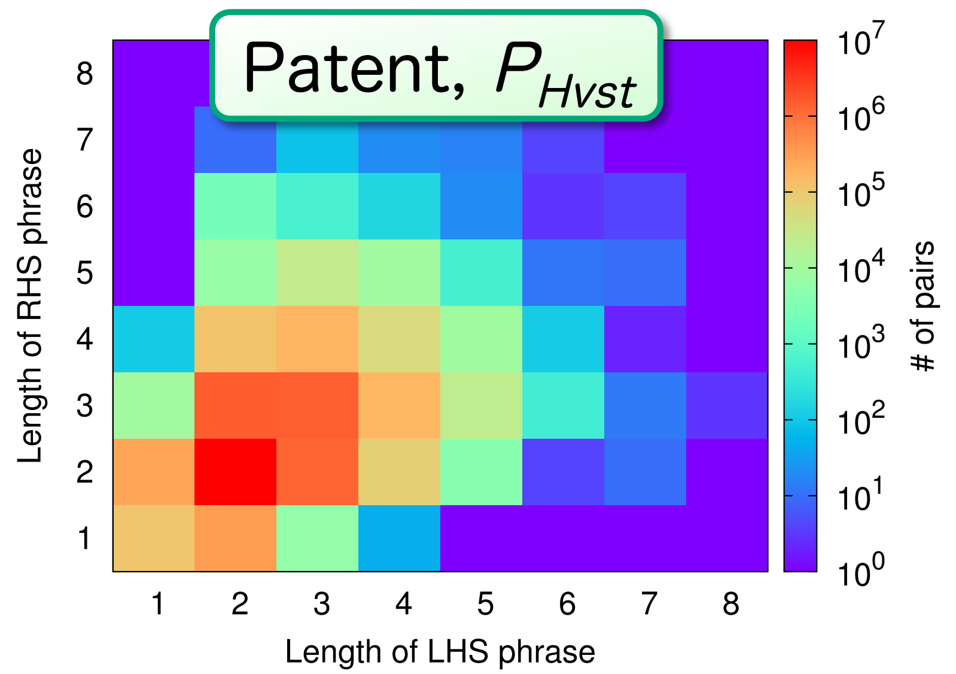






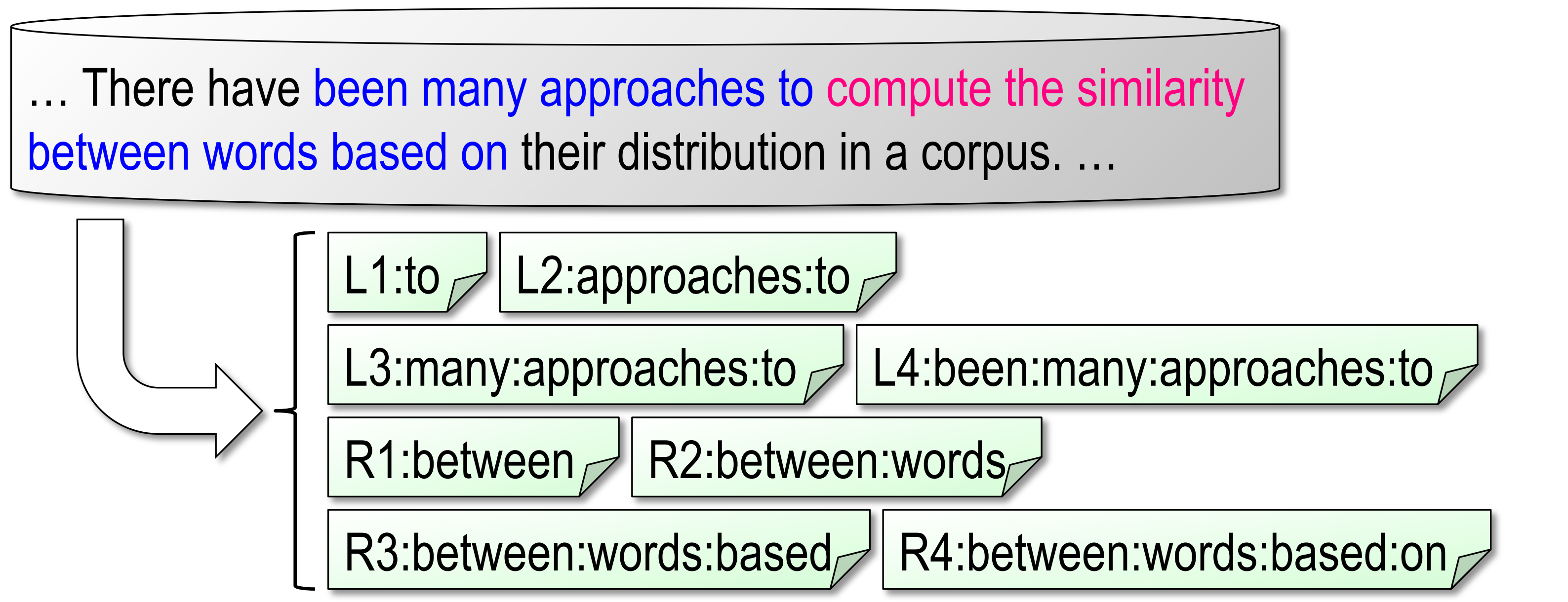






Recipe for contextual similarity

- Ingredients
  - Extract contextual features: adjacent *n*-grams
    - cf. Bag-of-words (cheap but noisy)
    - cf. Dependency trees (accurate but expensive)
  - Weight and filter features: nothing
  - Aggregate into a single value: cosine of vectors



... There have been many approaches to compute the similarity between words based on their distribution in a corpus. ...

L1:to
L2:approaches:to
L3:many:approaches:to
L4:been:many:approaches:to
R1:between
R2:between:words
R3:between:words:based
R4:between:words:based:on

Examples

Europarl, P_{Seed}

multi-lateral ⇒ multilateral
i would like to start by congratulating ⇒ let me first of all congratulate
transitional {process, year} ⇒ {process, year} of transition
in the course of the last few {months} ⇒ during recent {months}

Europarl, P_{Hvst}

transitional {task, strategy, phase, costs, ...}
⇒ {task, strategy, phase, costs, ...} of transition
in the course of the last few {weeks, years, decades}
⇒ during recent {weeks, years, decades}

Patent, P_{Seed}

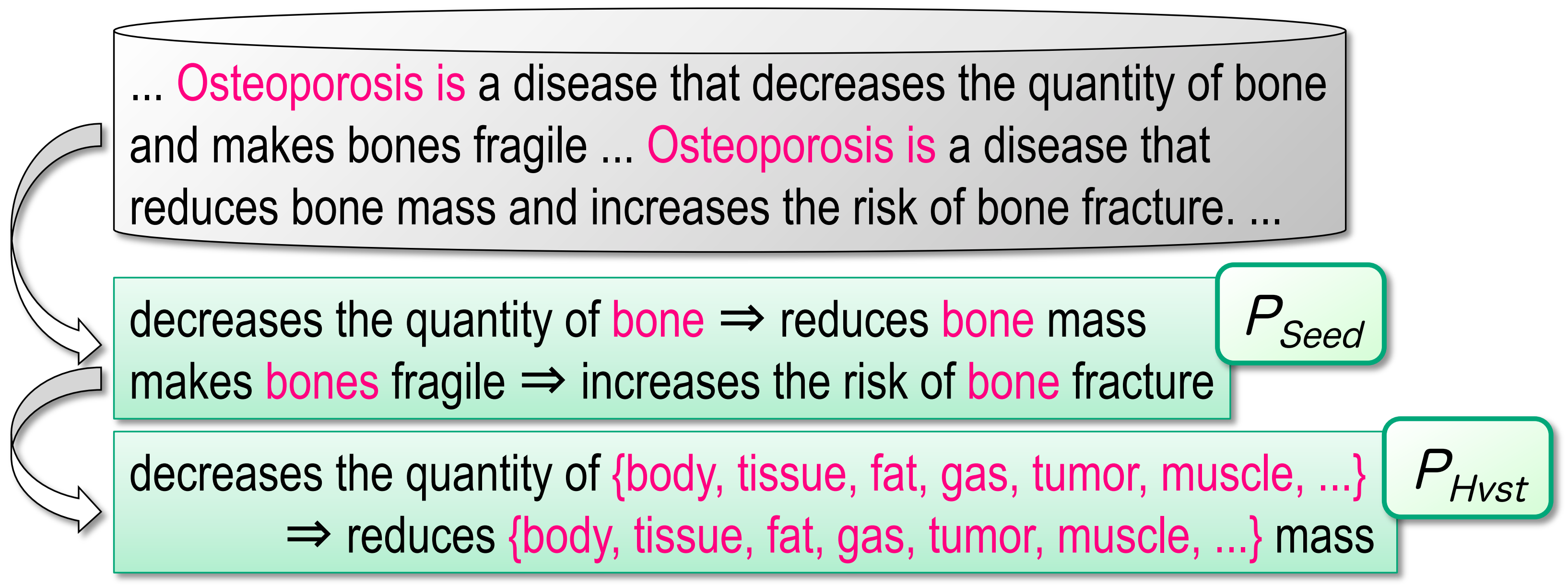
overall structure ⇒ entire configuration
in accordance with the structure mentioned above ⇒ due to such a constitution
{bypass, chip} condensers ⇒ {bypass, chip} capacitors
will be described with reference to {drawings} ⇒ is explained based on the {drawings}

Patent, P_{Hvst}

{layer, ceramic, ferroelectric, solid, ...} condensers
⇒ {layer, ceramic, ferroelectric, solid, ...} capacitors
will be described with reference to {embodiments}
⇒ is explained based on the {embodiments}

Alternative for seed paraphrases

- Any high-prec. set can be used as P_{Seed}
  - e.g., Multiple definition sentences [Hashimoto+, 11]



... Osteoporosis is a disease that decreases the quantity of bone and makes bones fragile ... Osteoporosis is a disease that reduces bone mass and increases the risk of bone fracture. ...

P_{Seed}

decreases the quantity of bone ⇒ reduces bone mass
makes bones fragile ⇒ increases the risk of bone fracture

P_{Hvst}

decreases the quantity of {body, tissue, fat, gas, tumor, muscle, ...}
⇒ reduces {body, tissue, fat, gas, tumor, muscle, ...} mass

Human evaluation: details

- Show 5 alternatives at the same time
  - To make results more consistent
  - To reduce the human labor
- “Grammaticality” and “Meaning equiv.”
  - 5-pt scales and binary prec. [Callison-Burch, 08]
  - [Callison-Burch, 08]
    - Europarl (10 langs-En) + CCG + LM
    - WMT 2007 “newstest” data
    - Binary prec. : .68 for G, .62 for M, .55 for both
    - Ours (total): .76 for G, .71 for M, .58 for both
  - [Chan+, 11]
    - Europarl (10 langs-En) + CCG + Google N-gram
    - Europarl (i.e., closed)
    - Score for 1-best: 4.2 pts for G and 3.7 pts for M
    - Ours (1-best): 4.57 pts for G and 3.96 pts for M

Limitations

- Our method (current version)
  - Does not cover totally different expressions
- Type-based approaches
  - Do not properly deal with polysemy
  - Tend to miss rare expressions
- Corpus-based approaches
  - Do not acquire expressions that do not appear

Future work

- In-depth analyses of the proposed method
  - Similarity metrics
  - Paraphrase patterns with more than one variable
  - Size & type of monolingual corpora
- Sophisticated paraphrase patterns
  - Hierarchical pattern induction
  - Deeper level of lexical correspondences
- Use for NLP applications