Bilingual Pivotingによる言い換え獲得の相互情報量に基づく一般化

梶原 智之　　小町 守　　持橋 大地
首都大学東京　　統計数理研究所

大規模な言い換え知識PPDBが多くのNLP応用タスクで活躍

PPDB: Paraphrase Database

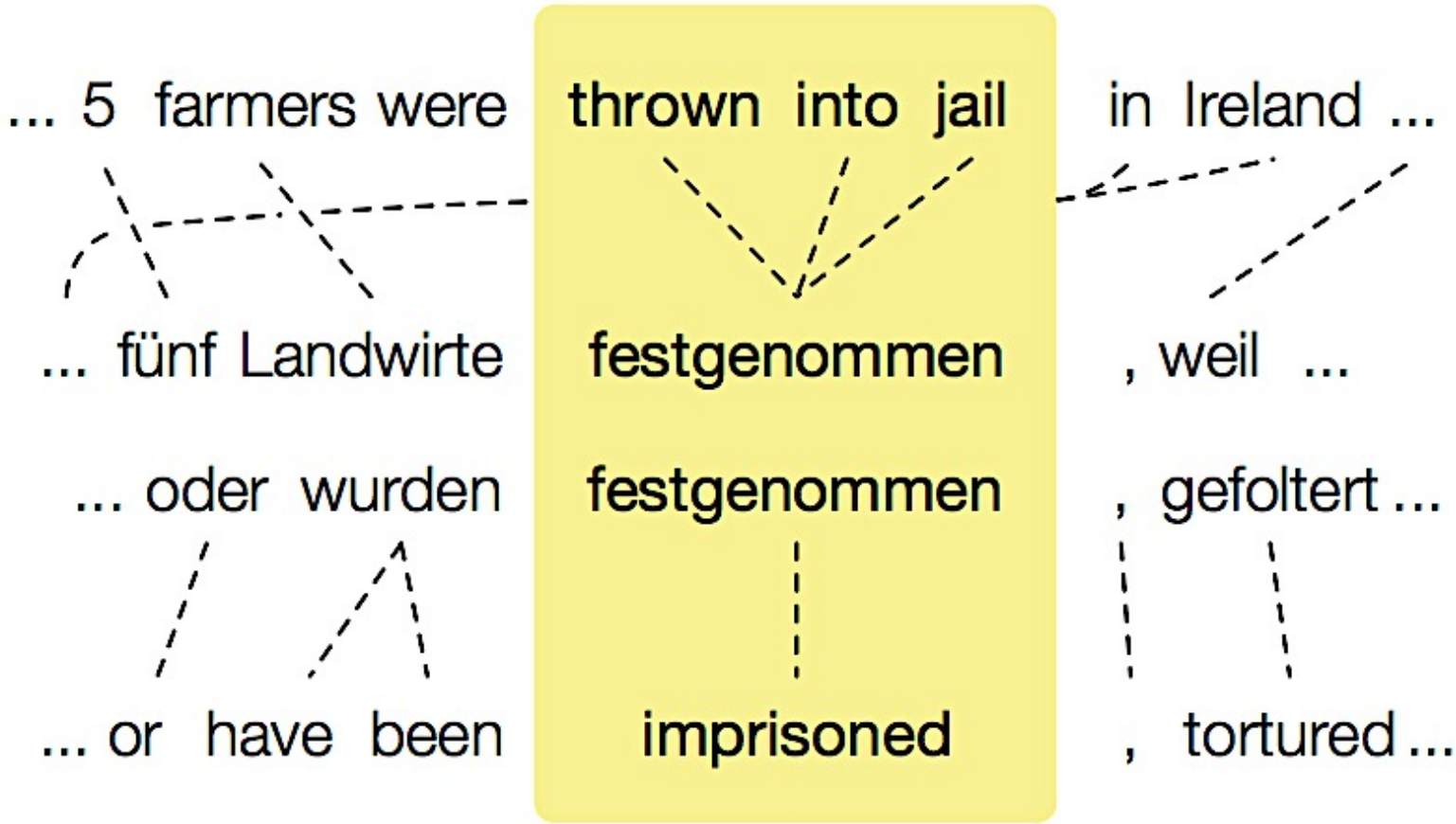
- 英語　100M フレーズ対 [Ganitkevitch+ 13, Pavlick+ 15]
- 日本語　15M フレーズ対 [Mizukami+ 14]

タスク	目的
機械翻訳	未知語を削減
意味的文間類似度	単語アライメントを改善
単語分散表現の学習	同義語の教師データを列挙
テキスト平易化	平易な言い換え候補を列挙

しかし、ノイズが多いので、言い換え知識PPDBを改善したい

- hardware: 18 / 192 語しか正しい言い換えがない
  hw, equipment, material, materiel, computer, apparatus, hardcore, appliance, physical, team, accessory, …
- infringed: 17 / 129 語しか正しい言い換えがない
  broken, flouted, injured, encroached, undermined, prejudiced, trodden, impaired, transgressed, abused, …

背景：Bilingual Pivoting [Bannard+ 05]



パラレルコーパス上での
2次の単語アライメント確率 → 言い換え確率

背景：Bilingual Pivoting → PPDB

Bilingual Pivoting

$$p(e_2|e_1) = \sum_f p(e_2|f, e_1)\, p(f|e_1)$$

$$\approx \sum_f p(e_2|f)\, p(f|e_1)$$

e_1とe_2の条件付き
独立性を仮定した近似

PPDB

$$s_{bp}(e_1, e_2) = -\lambda_1 \log p(e_2|e_1) - \lambda_2 \log p(e_1|e_2)$$

$$= \log p(e_2|e_1) + \log p(e_1|e_2)$$

両方向の言い換え確率を考慮する対数線形モデル
本研究では、$\lambda_1 = \lambda_2 = -1$（先行研究では+1）

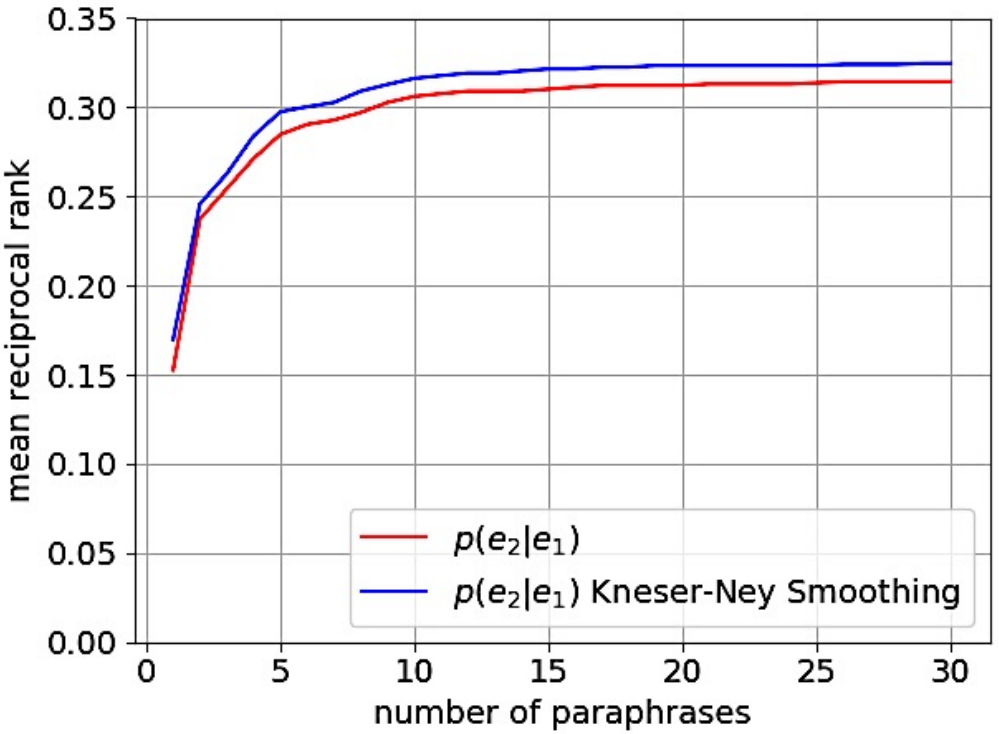
Bilingual Pivotingの課題

$$p(e_2|e_1) \approx \sum_f p(e_2|f)\, p(f|e_1)$$

$$s_{bp}(e_1, e_2) = \log p(e_2|e_1) + \log p(e_1|e_2)$$

- 仮説1：低頻度語の確率を過剰に大きく推定してしまう
- 仮説2：高頻度語はアライメント誤りの影響を受けやすく
  不当に多くの単語の言い換えになってしまう
- 仮説3：分布類似度とは異なる観点から単語間の
  同義性を捉えている（e.g. hardware ↔ team）

課題1：低頻度語の確率を過推定
対策1：Kneser-Ney Smoothing

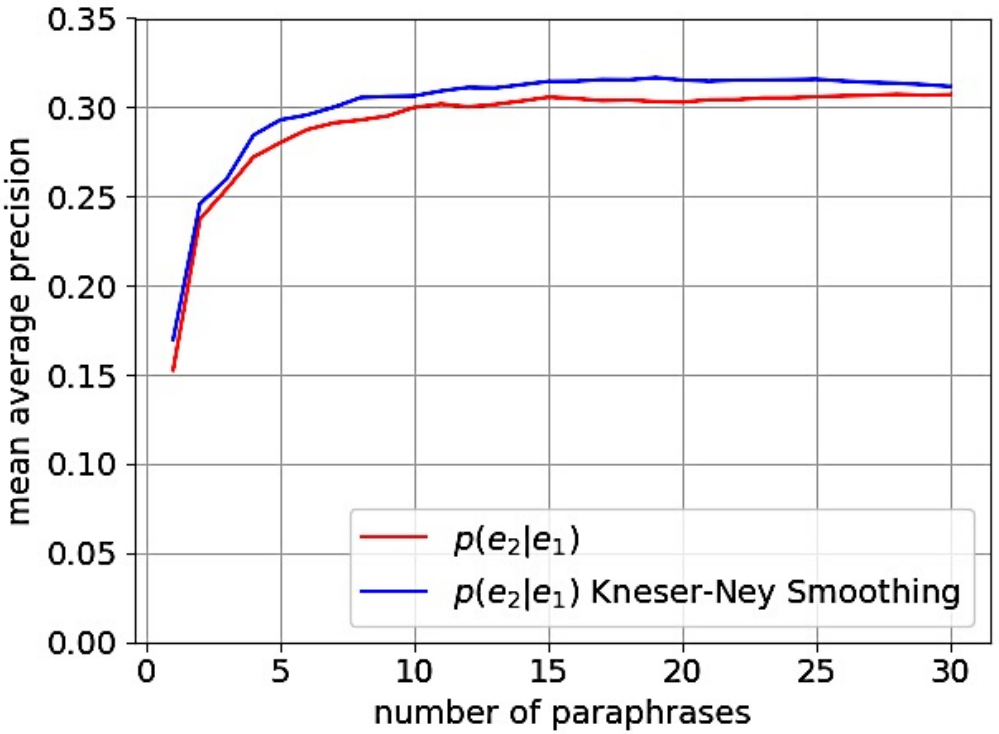


平均逆順位（MRR）
初めて正解が出現する順位の逆数の平均値

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}$$

Kneser-Ney Smoothingを用いて
パラレルコーパスのデータスパースネス問題に対処する

課題1：低頻度語の確率を過推定
対策1：Kneser-Ney Smoothing



平均適合率（MAP）
適合率の平均値

$$\mathrm{AP} = \frac{1}{|A|} \sum_{k=1}^{|A|} P(\mathrm{rank}_{\mathrm{A_k}})$$

$$\mathrm{MAP} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathrm{AP}_i$$

以降の実験では、$p(e_2|e_1)$に常にKneser-Ney Smoothingを適用

課題2：高頻度語が言い換えノイズ
対策2：相互情報量に基づく一般化

PPDB

$$s_{bp}(e_1, e_2) = \log p(e_2|e_1) + \log p(e_1|e_2)$$

PMI

$$s_{pmi}(e_1, e_2) = \log p(e_2|e_1) + \log p(e_1|e_2) - \log p(e_1) - \log p(e_2)$$

$$= \log \frac{p(e_2|e_1)}{p(e_2)} + \log \frac{p(e_1|e_2)}{p(e_1)} = 2\mathrm{PMI}(e_1, e_2)$$

$$\because \mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(y|x)}{p(y)} = \log \frac{p(x|y)}{p(x)}$$

課題3：分布類似度の長短と傾向が違う
対策3：分布類似度を組み込む

Local PMI

$$\mathrm{LPMI}(x, y) = n(x, y) \cdot \log \frac{p(x, y)}{p(x)p(y)}$$

低頻度な単語対において、偶然の共起によりPMIが
不当に大きくなる低頻度バイアスの問題を緩和する

提案手法

$$s_{lpmi}(e_1, e_2) = cos(\vec{e}_1, \vec{e}_2) \cdot s_{pmi}(e_1, e_2)$$
$$= cos(\vec{e}_1, \vec{e}_2) \cdot 2\mathrm{PMI}(e_1, e_2)$$

我々は共起の強さではなく、言い換えらしさを求めたい

提案手法の特徴

パラレルコーパスと単言語コーパスの
両方から得られる情報を組み合わせる頑健性

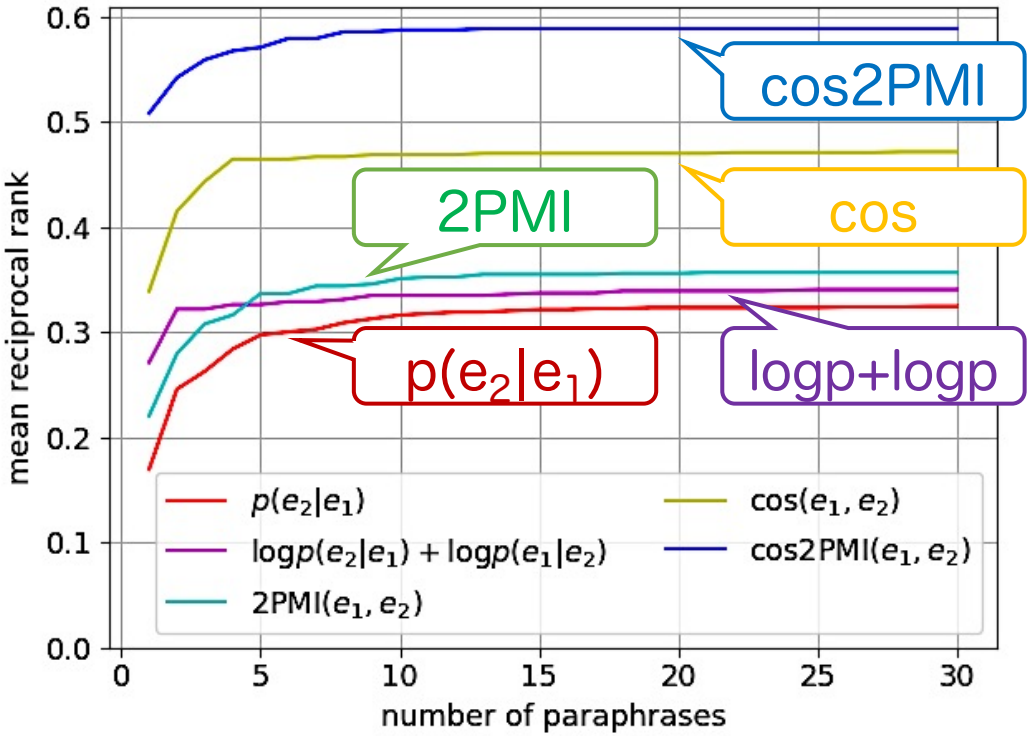
$$\cos(e_1, e_2)2\mathrm{PMI}(e_1, e_2) = \boldsymbol{\cos(\overrightarrow{e_1}, \overrightarrow{e_2})} \left\{ \log \frac{\boldsymbol{p(e_2|e_1)}}{\boldsymbol{p(e_2)}} + \log \frac{\boldsymbol{p(e_1|e_2)}}{\boldsymbol{p(e_1)}} \right\}$$

- $\boldsymbol{p(e_2 \mid e_1)}$：類義語や反義語によるノイズが少ない
- $\boldsymbol{\cos(\overrightarrow{e_1}, \overrightarrow{e_2})}$：無関係な単語によるノイズが少ない

実験設定：英語の言い換えランキング

- $p(e_2|e_1)$
  - Europarl-v7：パラレルコーパス（英語 ↔ 仏語）
  - Giza++：単語アライメント
  - 170,682,871 単語対の言い換え候補（e_1とe_2の組）を獲得　※ e_1=e_2の組は除いた
- $p(e_1)$ および $\cos(\overrightarrow{e_1}, \overrightarrow{e_2})$
  - English Gigaword 5th Edition：単言語コーパス
  - Kenlm：1-gram言語モデル
  - word2vec：分布類似度（CBOWモデル）
- 評価用データ
  - Human Paraphrase Judgments [Pavlick+ 15]
  - 26,456語対に5人が5段階の同義性スコアを付与

提案手法 cos2PMI が言い換えランキングを大幅に改善 (1/2)

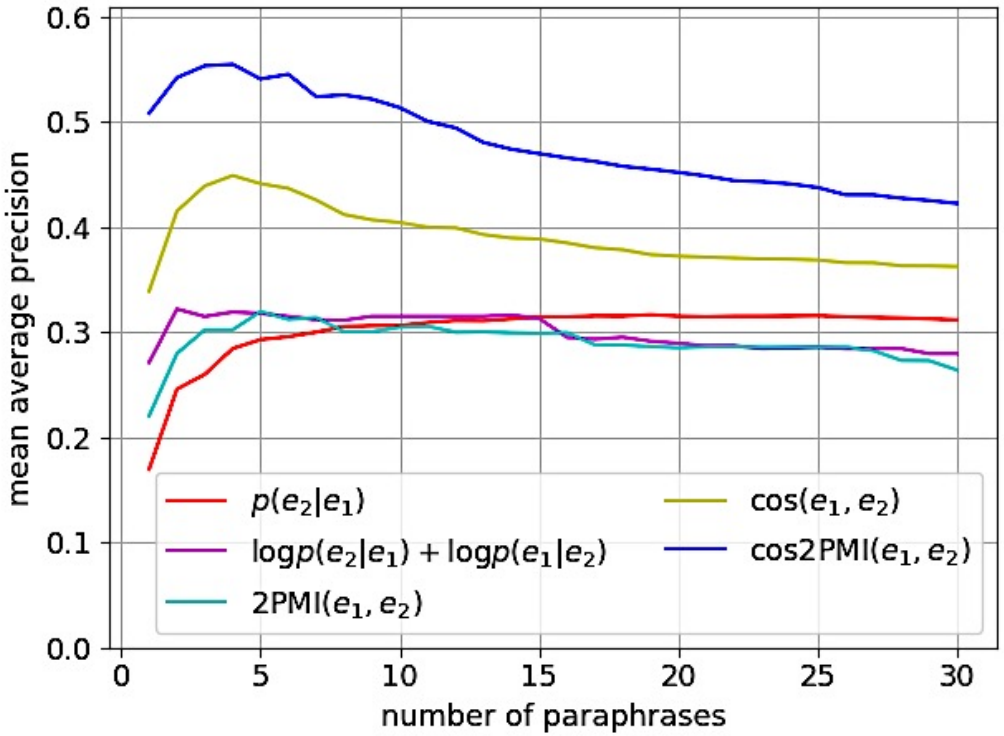


平均逆順位（MRR）
初めて正解が出現する順位の逆数の平均値

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}$$

- PMIはTop-5以降の言い換えのランキングに対して有効
  ※ 低頻度バイアスの影響でトップには頻度1の語が出現
- COSとの組み合わせによって大幅に性能を改善できた

提案手法 cos2PMI が言い換えランキングを大幅に改善 (2/2)



平均適合率（MAP）
適合率の平均値

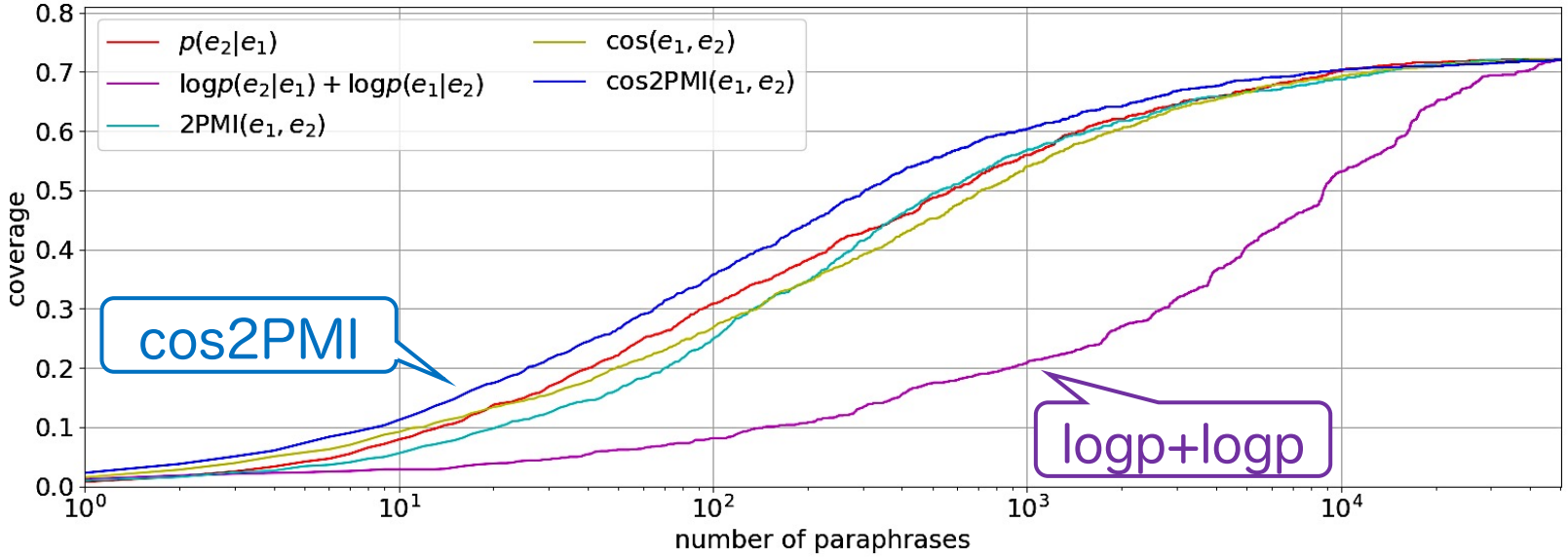
$$\mathrm{AP} = \frac{1}{|A|} \sum_{k=1}^{|A|} P(\mathrm{rank}_{\mathrm{A_k}})$$

$$\mathrm{MAP} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \mathrm{AP}_i$$

- PMIはTop-5以降の言い換えのランキングに対して有効
  ※ 低頻度バイアスの影響でトップには頻度1の語が出現
- COSとの組み合わせによって大幅に性能を改善できた

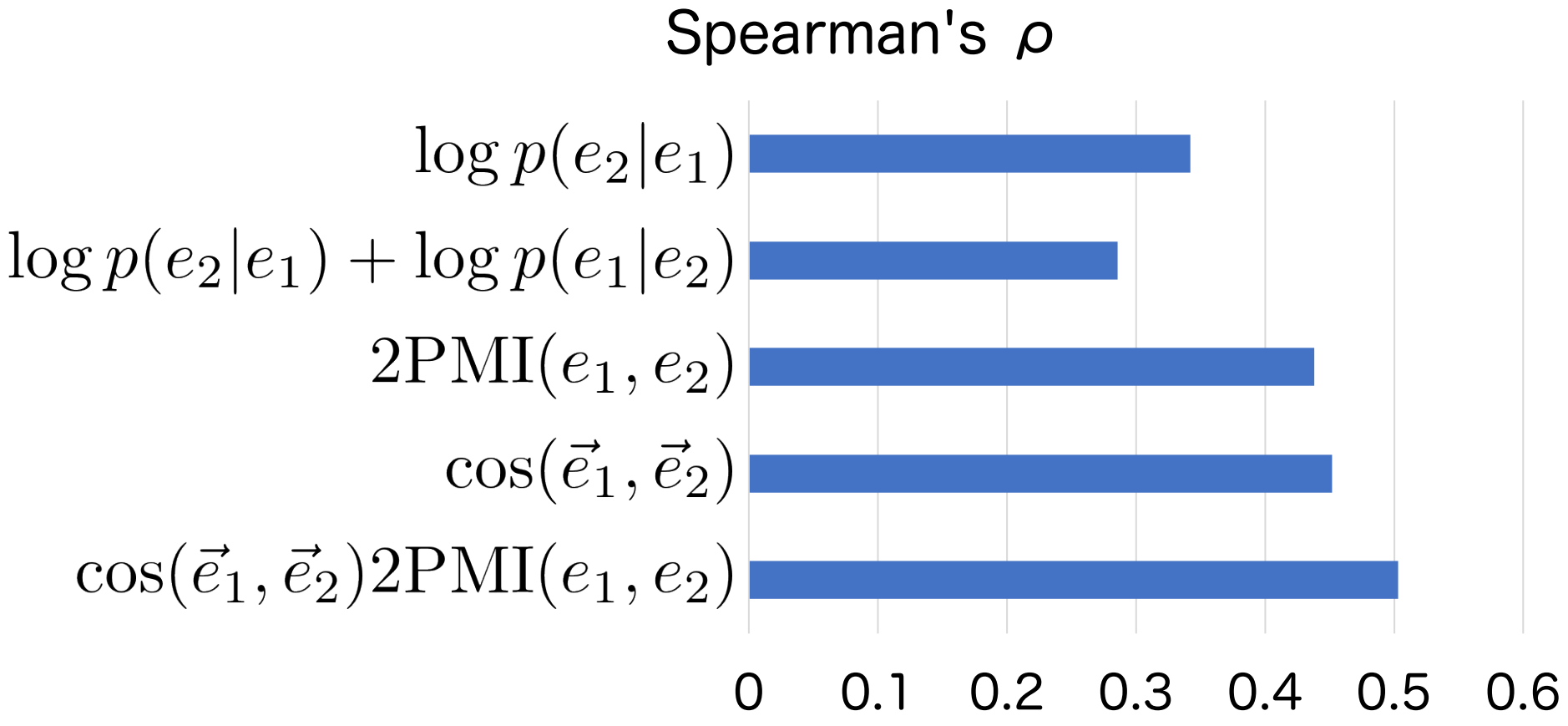
上位の言い換えランキングだけではなく提案手法は常に高性能

Coverage: 正解のカバー率



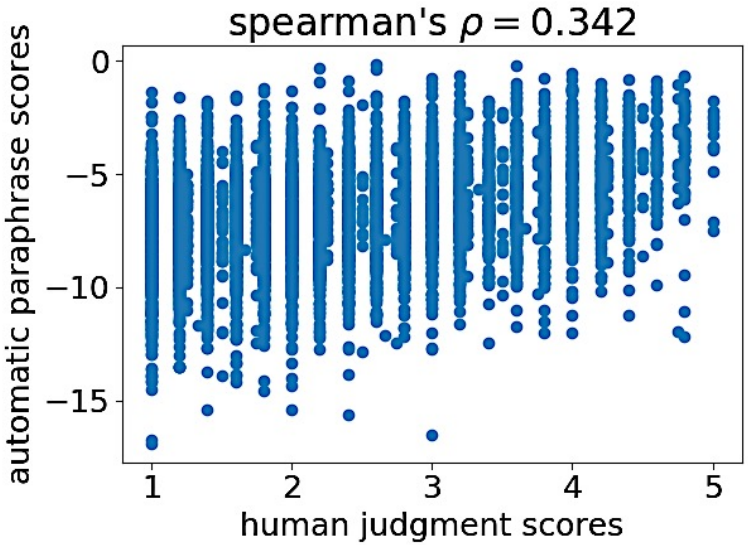
- asなどの高頻度な機能語が特に単語アライメント誤りの影響を受けやすく、最大5万種類を超える言い換え候補を持つ
- logp+logp (PPDB)はTop-5以降の信頼性が著しく低い

相関係数も同様の結果を示した
cos2PMI > cos > PMI > BP

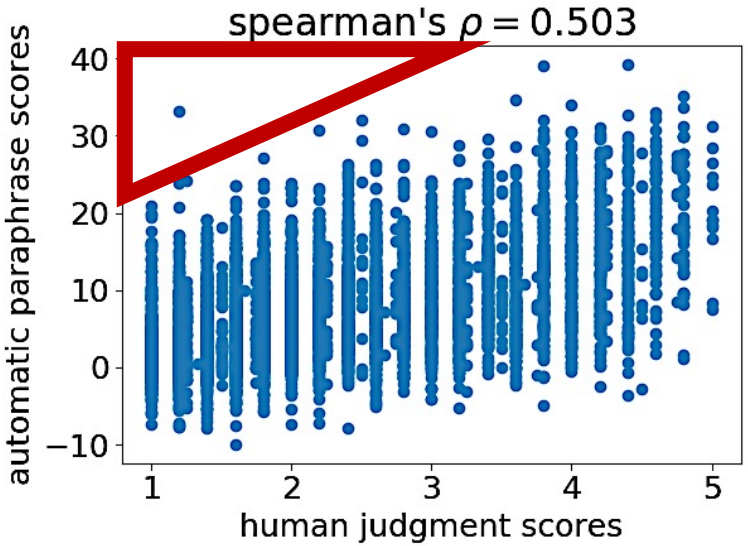


False Positive を削減できた

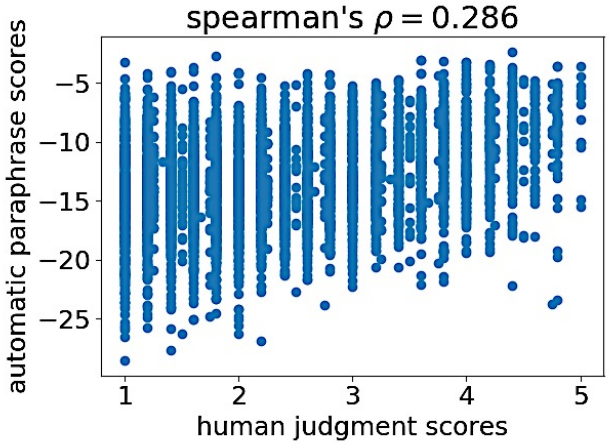
$\log p(e_2|e_1)$



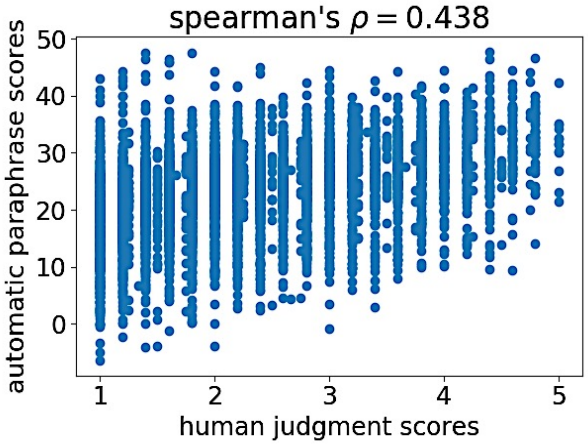
$\cos(\vec{e}_1, \vec{e}_2)2\mathrm{PMI}(e_1, e_2)$



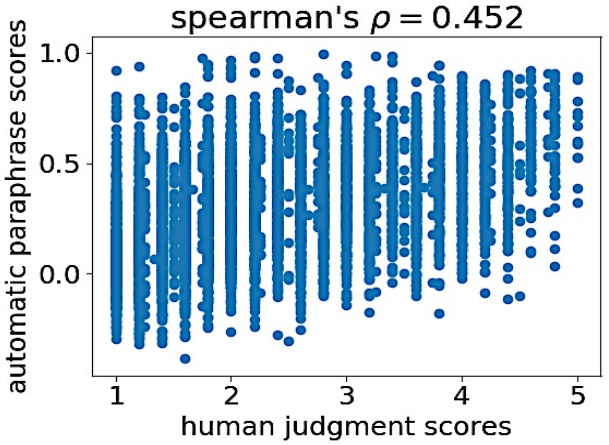
$\log p(e_2|e_1) + \log p(e_1|e_2)$



$2\mathrm{PMI}(e_1, e_2)$



$\cos(\vec{e}_1, \vec{e}_2)$



Cultural の言い換えランキング Top-10

	BP	2BP	PMI	COS	COS2PMI
1	diverse	culturally	culturally-based	historical	socio-cultural
2	harvests	culture	culturaldevelopment	culture	culture
3	firstly	151	cultural-social	educational	multicultural
4	understand	charter	economic-cultural	linguistic	intercultural
5	flowering	monuments	culture-	multicultural	educational
6	trying	art	cultural-educational	cross-cultural	intellectual
7	structure	casal	kulturkampf	diversity	culturally
8	january	kahn	cultural-political	technological	sociocultural
9	culture	13	multiculture	intellectual	heritage
10	culturally	caning	culturally	preservation	architectural

提案手法は、ノイズ・低頻度語・類義語などに頑健

Labourersの正しい言い換え

BP	2BP	PMI	COS	COS2PMI
1. workers	9. gardeners	10. workmen	2. workers	2. workers
2. employees	42. harvesters	11. wage-earners	8. people	4. workmen
9. farmers	62. workers	16. earners	10. persons	5. craftsmen
13. labour	71. seafarers	19. workers	11. farmers	6. wage-earners
16. gardeners	73. unions	21. craftsmen	15. craftsmen	9. persons
17. people	99. homeworkers	22. workforces	26. wage-earners	12. employees
28. workmen	283. works	26. employed	27. workmen	13. earners
30. employed	394. workmen	27. employees	29. harvesters	15. farmers
33. craftsmen	395. employees	50. labour	31. seafarers	18. people
59. harvesters	412. wage-earners	55. persons	32. employees	19. workforces
80. work	415. craftsmen	75. farmers	42. gardeners	37. harvesters
88. earners	417. earners	103. homeworkers	47. earners	42. individuals
90. wage-earners	419. labour	105. individuals	55. workforces	53. labour
106. persons	420. employed	112. work	57. individuals	55. seafarers
109. individuals	431. people	135. people	79. unions	65. gardeners
114. seafarers	433. farmers	187. harvesters	103. labour	88. employed
115. unions	446. workforces	273. gardeners	140. homeworkers	100. homeworkers
131. workforces	451. work	317. seafarers	144. work	105. work
166. homeworkers	453. persons	456. unions	170. employed	149. unions
401. works	474. individuals	469. works	222. works	254. works

外的評価：意味的文間類似度タスク

- 与えられた2文間の類似度 [0.0, 1.0] を推定する
- 人手で付与された類似度とシステムが出力した類似度のPearson相関係数の高さを評価する

類似度	文
1.0	The bird is bathing in the sink. Birdie is washing itself in the water basin.
0.2	The woman is playing the violin. The young lady enjoys listening to the guitar.

手法：Sultan et al., 2015

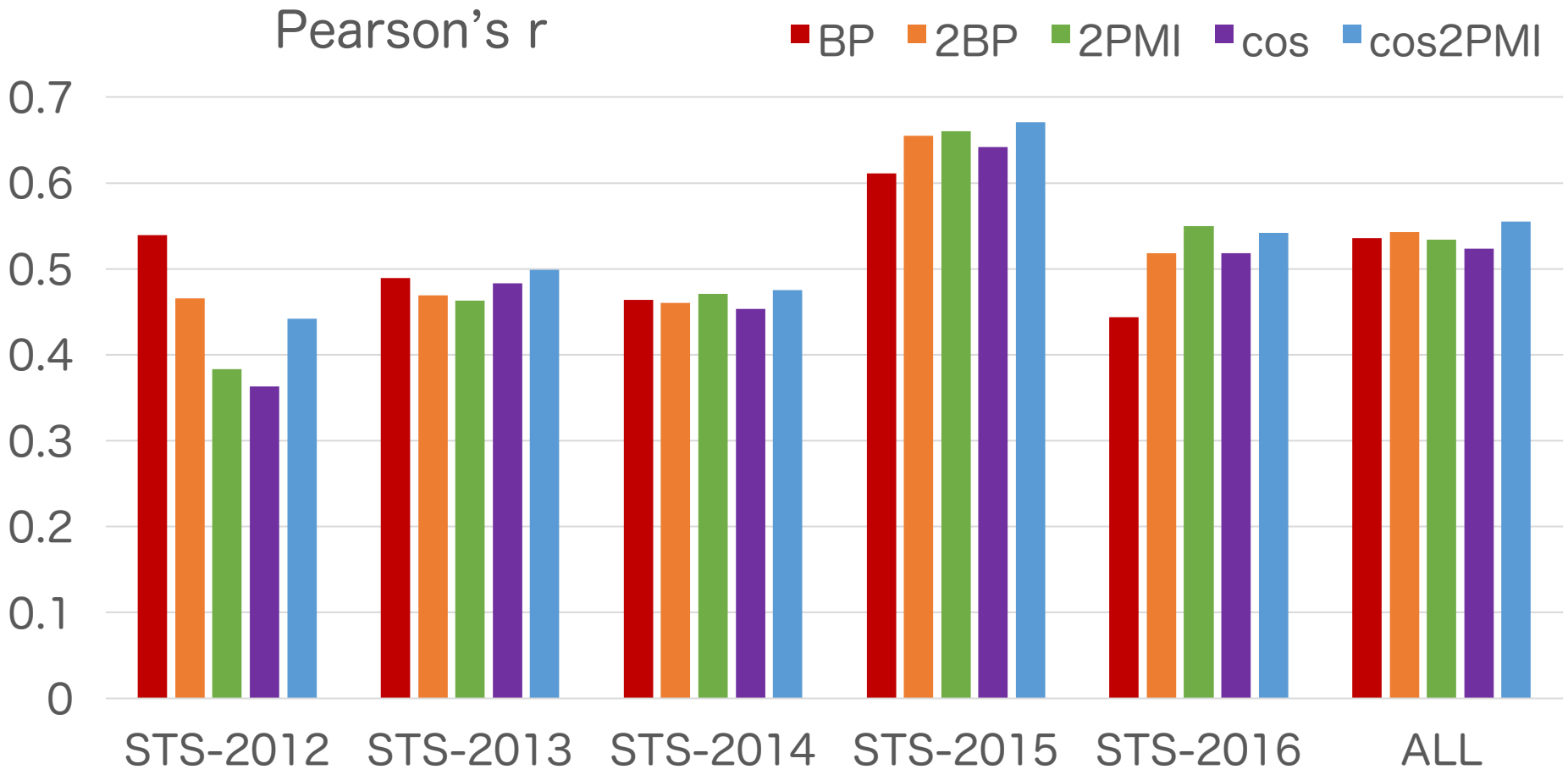
The bird is bathing in the sink .

Birdie is washing itself in the water basin .

$$\mathrm{STS}(s_1, s_2) = \frac{n_a(s_1) + n_a(s_2)}{n(s_1) + n(s_2)}$$

1.PPDBの言い換えを用いて単語アライメント
2.アラインされた単語の割合を文間類似度とする

Top-10の言い換えを用いた評価



関連研究

- Levy and Goldberg, NIPS-2014

  → 単語分散表現の学習手法 Skip-gram with Negative-sampling を Shifted Positive PMI で一般化

- 本研究

  → 言い換え獲得手法 Bilingual Pivoting を（重み付き）PMI で一般化

Paraphrase Database

- Bilingual Pivotingを用いて24言語で開発されている
- MT、QA、STSなど多くの応用タスクで活躍している
- 大規模な単言語コーパスは各言語で容易に利用できるため提案手法で多言語の多くのタスクを改善できる

Bilingual Pivotingによる言い換え獲得の相互情報量による一般化

- 単語間の言い換え獲得をPMIを用いて一般化
- パラレルコーパスの情報（Bilingual Pivoting）と単言語コーパスの情報（単語出現確率・分布類似度）の相補的な組み合わせによって頑健な言い換え獲得を実現
- 獲得した英語と日本語の言い換え対を以下で公開

https://github.com/tmu-nlp/pmi-ppdb