MIPA: Mutual Information Based Paraphrase Acquisition via Bilingual Pivoting

Tomoyuki Kajiwara Tokyo Metropolitan University
Mamoru Komachi Tokyo Metropolitan University
Daichi Mochihashi The Institute of Statistical Mathematics

TOKYO METROPOLITAN UNIVERSITY

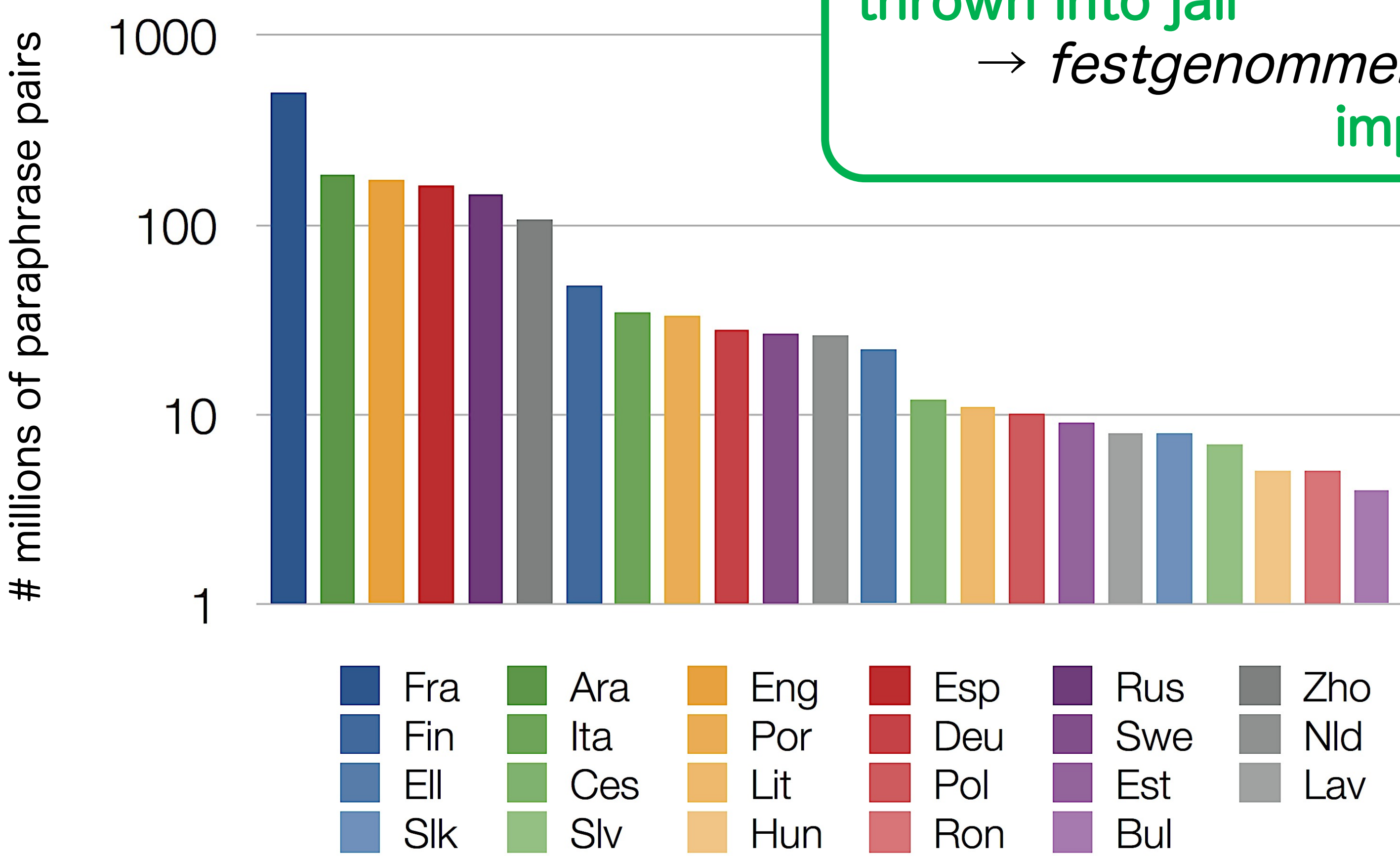
Research Organization of Information and Systems
The Institute of Statistical Mathematics

Paraphrase Lexicons are useful for many NLP applications

PPDB: Millions of paraphrase pairs in 24 languages

[Ganitkevitch+ 2013, Ganitkevitch+ 2014, Mizukami+ 2014, Pavlick+ 2015]

thrown into jail
→ *festgenommen* →
imprisoned



We reduce the noise included in PPDB

PPDB is proven useful for

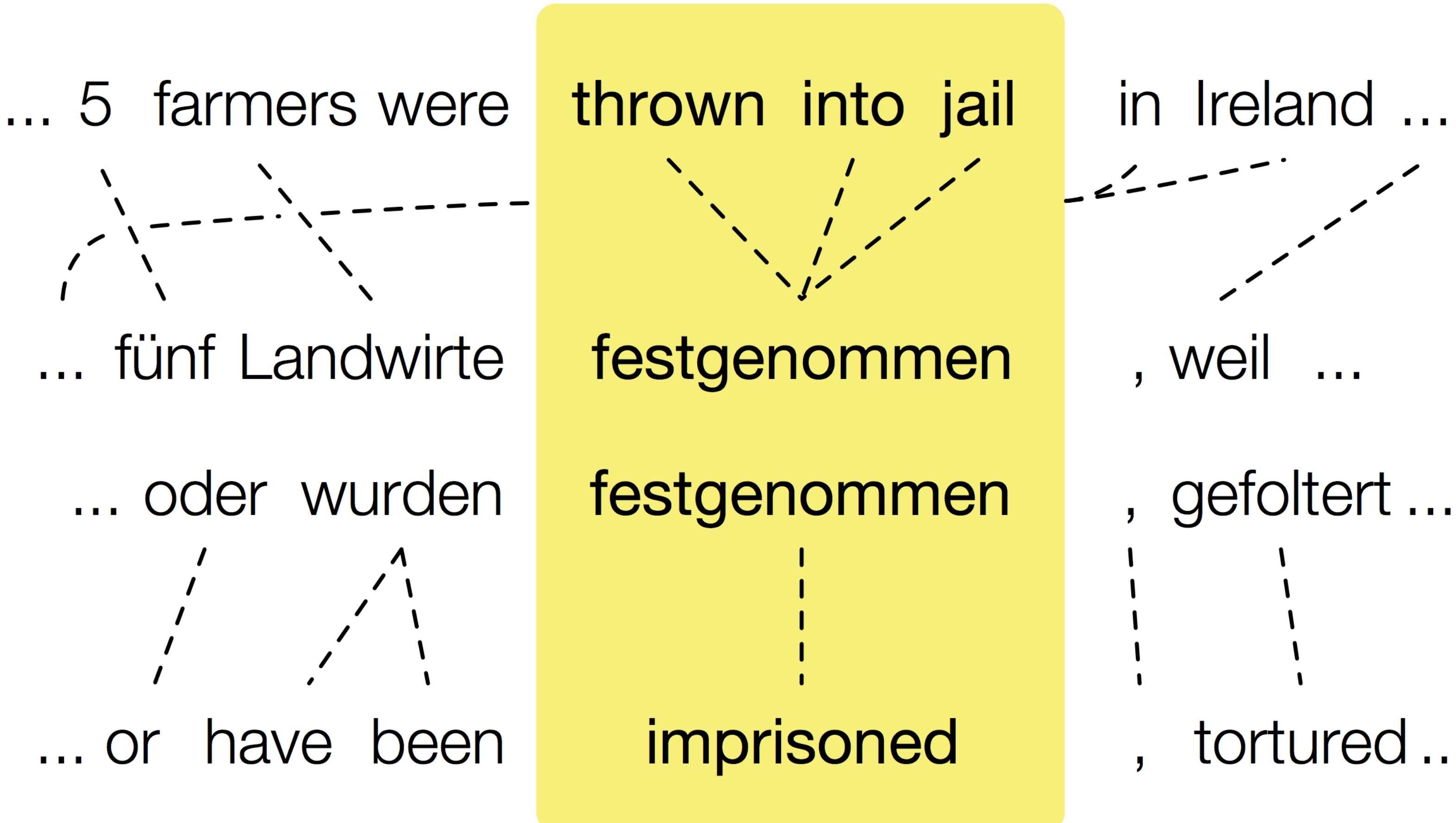
- Semantic Textual Similarity [Sultan+ 2015]
- Machine Translation [Mehdizadeh Seraj+ 2016]
- Text Simplification [Xu+ 2016]

However, PPDB includes noise caused by word alignment errors on bilingual pivoting.

hardware: only 18 / 192 words are correct paraphrases in PPDB

hw, equipment, material, materiel, computer, apparatus, hardcore, appliance, physical, team, accessory, …

Bilingual Pivoting [Bannard+ 2005]



Two-level word alignment probability on a bilingual corpus → Paraphrase probability

Bilingual Pivoting → PPDB

Bilingual Pivoting

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

PPDB

Bilingual Pivoting → PPDB

Bilingual Pivoting

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

Assumes conditional independence of e_1 and e_2

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

PPDB

Bilingual Pivoting → PPDB

Bilingual Pivoting

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

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

Assumes conditional independence of e_1 and e_2

PPDB

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

A log-linear model that considers paraphrase probability in both directions.

Bilingual Pivoting → PPDB

Bilingual Pivoting

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

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

Assumes conditional independence of e_1 and e_2

PPDB

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

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

A log-linear model that considers paraphrase probability in both directions. We set $\lambda_1 = \lambda_2 = -1$ (PPDB: $\lambda_1 = \lambda_2 = 1$).

Problems of Bilingual Pivoting

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

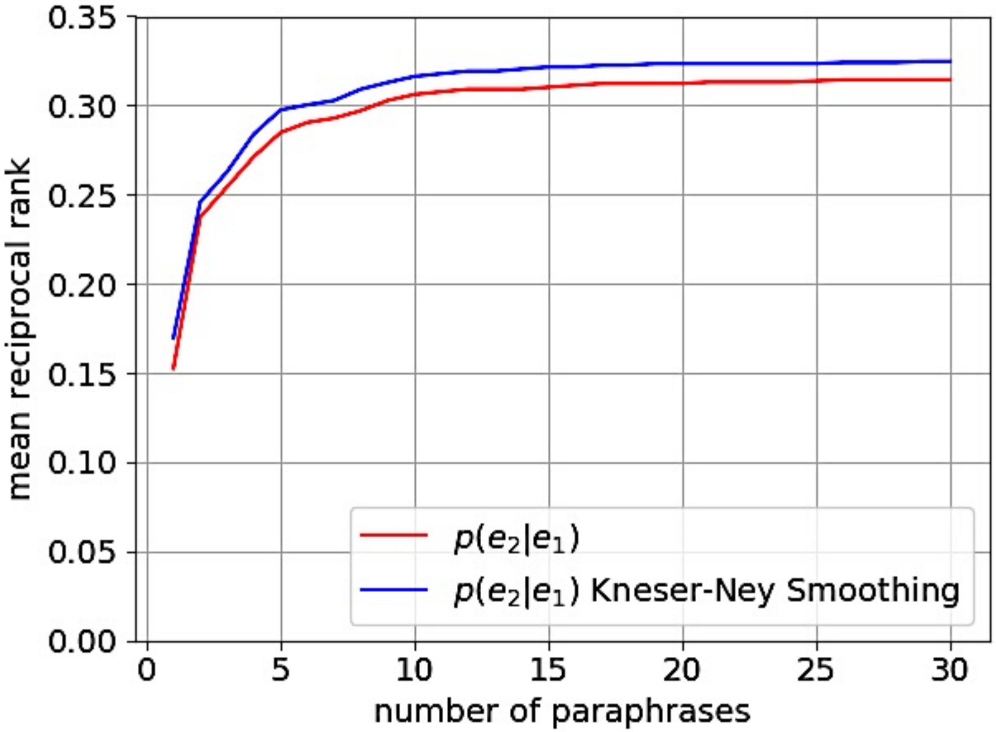
$$s_{bp}(e_1, e_2) = \log p(e_2|e_1) + \log p(e_1|e_2)$$

1. Word alignment probability may be overestimated for low-frequency word pairs.

2. High-frequency words may be assigned as a paraphrase for too many words due to misalignment.

3. Bilingual Pivoting may capture synonymity between words from a different viewpoint from Distributional Similarity.
   (e.g. Distributional Similarity does not erroneously recognize that *hardware* and *team* are synonymous.)

1. Kneser-Ney Smoothing of Bilingual Pivoting



Mean Reciprocal Rank

The average of the reciprocals of the ranking at which the correct paraphrase first appears.

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

- Word alignment probability may be overestimated for low-frequency word pairs.
- We propose using Kneser-Ney smoothing to mitigate overestimation of word alignment probability.

Problems of Bilingual Pivoting

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

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

1. Word alignment probability may be overestimated for low-frequency word pairs.

2. High-frequency words may be assigned as a paraphrase for too many words due to misalignment.

3. Bilingual Pivoting may capture synonymity between words from a different viewpoint from Distributional Similarity.
   (e.g. Distributional Similarity does not erroneously recognize that *hardware* and *team* are synonymous.)

2. Generalization of Bilingual Pivoting using PMI

PPDB

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

PMI

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

2. Generalization of Bilingual Pivoting using PMI

PPDB

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

PMI

$$\begin{aligned} s_{pmi}(e_1, e_2) &= \log p(e_2|e_1) + \log p(e_1|e_2) - \log p(e_1) - \log p(e_2) \\ &= \log \frac{p(e_2|e_1)}{p(e_2)} + \log \frac{p(e_1|e_2)}{p(e_1)} = 2\mathrm{PMI}(e_1, e_2) \end{aligned}$$

2. Generalization of Bilingual Pivoting using PMI

PPDB

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

PMI

$$\begin{aligned} s_{pmi}(e_1, e_2) &= \log p(e_2|e_1) + \log p(e_1|e_2) - \log p(e_1) - \log p(e_2) \\ &= \log \frac{p(e_2|e_1)}{p(e_2)} + \log \frac{p(e_1|e_2)}{p(e_1)} = 2\mathrm{PMI}(e_1, e_2) \end{aligned}$$

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

Problems of Bilingual Pivoting

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

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

1. Word alignment probability may be overestimated for low-frequency word pairs.

2. High-frequency words may be assigned as a paraphrase for too many words due to misalignment.

3. **Bilingual Pivoting may capture synonymity between words from a different viewpoint from Distributional Similarity.**

   (e.g. Distributional Similarity does not erroneously recognize that *hardware* and *team* are synonymous.)

3. Incorporating Distributional Similarity

Local PMI

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

In low-frequency word pairs, it is well-known that PMI becomes unreasonably large because of coincidental co-occurrence.
In order to avoid this problem, Local PMI assigns weights to PMI depending on the co-occurrence frequency of word pairs.

MIPA

3. Incorporating Distributional Similarity

Local PMI

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

In low-frequency word pairs, it is well-known that PMI becomes unreasonably large because of coincidental co-occurrence.
In order to avoid this problem, Local PMI assigns weights to PMI depending on the co-occurrence frequency of word pairs.

MIPA

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

Our aim is to estimate not the strength of co-occurrence, but the synonymity between words.

MIPA: Complementary use of Bi- and Mono-lingual corpus

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

- $\boldsymbol{p(e_2|e_1)}$
  - Synonymity estimated using **bilingual corpus**
  - There is little noise due to **antonym word pairs**

- $\boldsymbol{\cos(\overrightarrow{e_1}, \overrightarrow{e_2})}$
  - Synonymity estimated using **monolingual corpus**
  - There is little noise due to **unrelated word pairs**

MIPA can accurately estimate synonymity between words
by using both bilingual and monolingual corpus complementary.

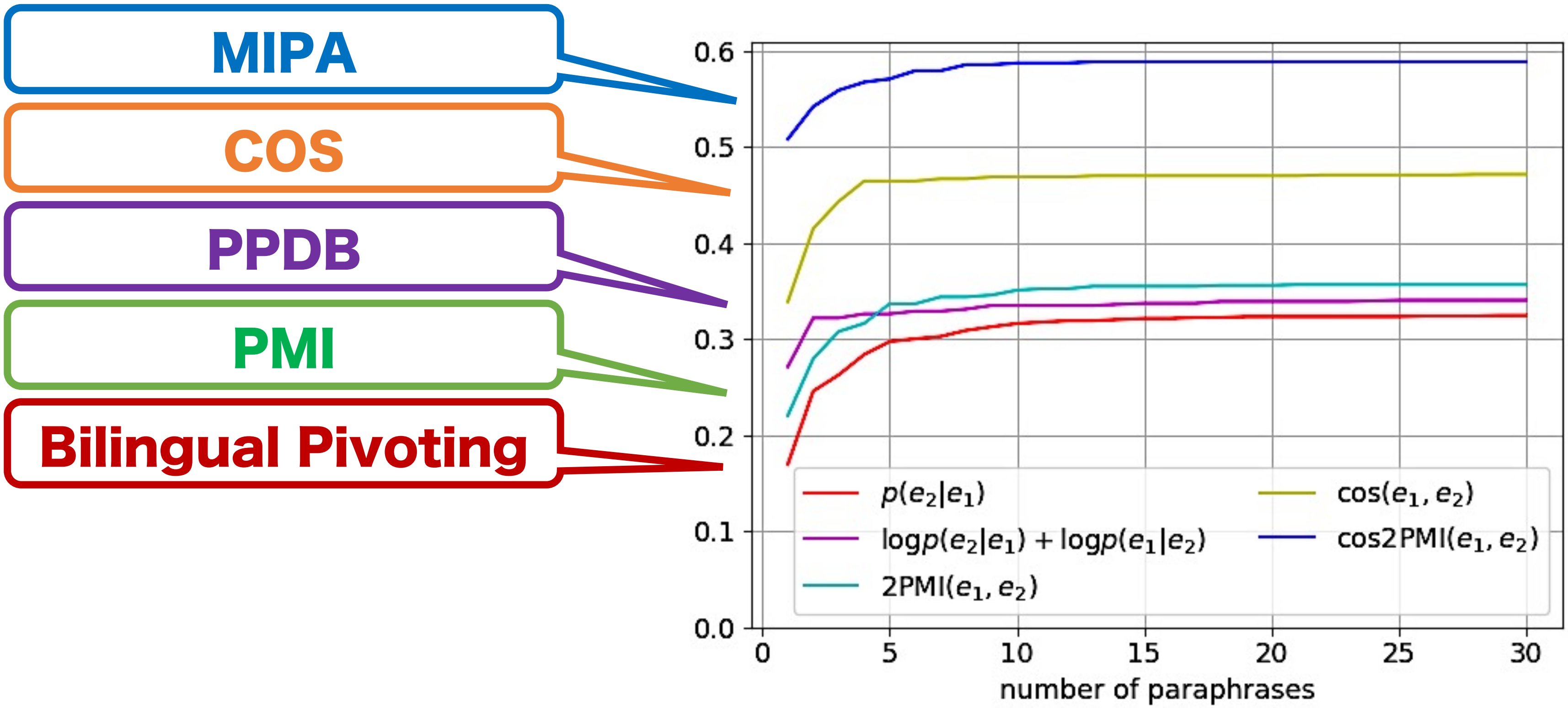
Experiments: English Lexical Paraphrase Ranking

- $p(e_2|e_1)$
  - Europarl-v7: En-Fr parallel corpus
  - Giza++: word alignment tool (IBM model 4)
  - Paraphrase Candidates: 170M word pairs, excepting the paraphrase of itself (e_1=e_2)

- $p(e_1)$ and $\cos(\overrightarrow{e_1}, \overrightarrow{e_2})$
  - English Gigaword 5th Edition: monolingual corpus
  - Kenlm: 1-gram language model
  - word2vec: word embeddings (CBOW model)

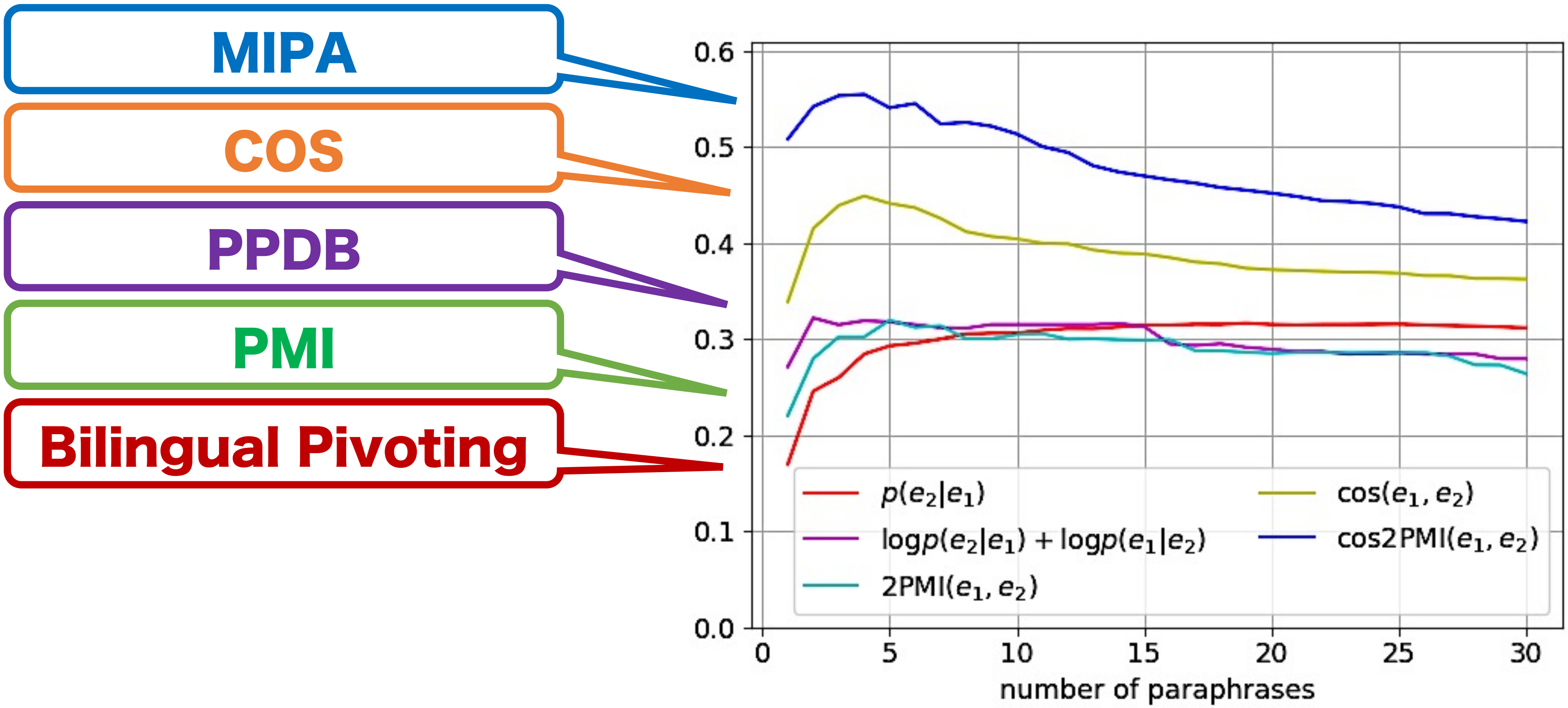
- Evaluation Dataset
  - Human Paraphrase Judgments [Pavlick+ 2015]
  - Five-step manual evaluation of 26K word pairs

Mean Reciprocal Rank



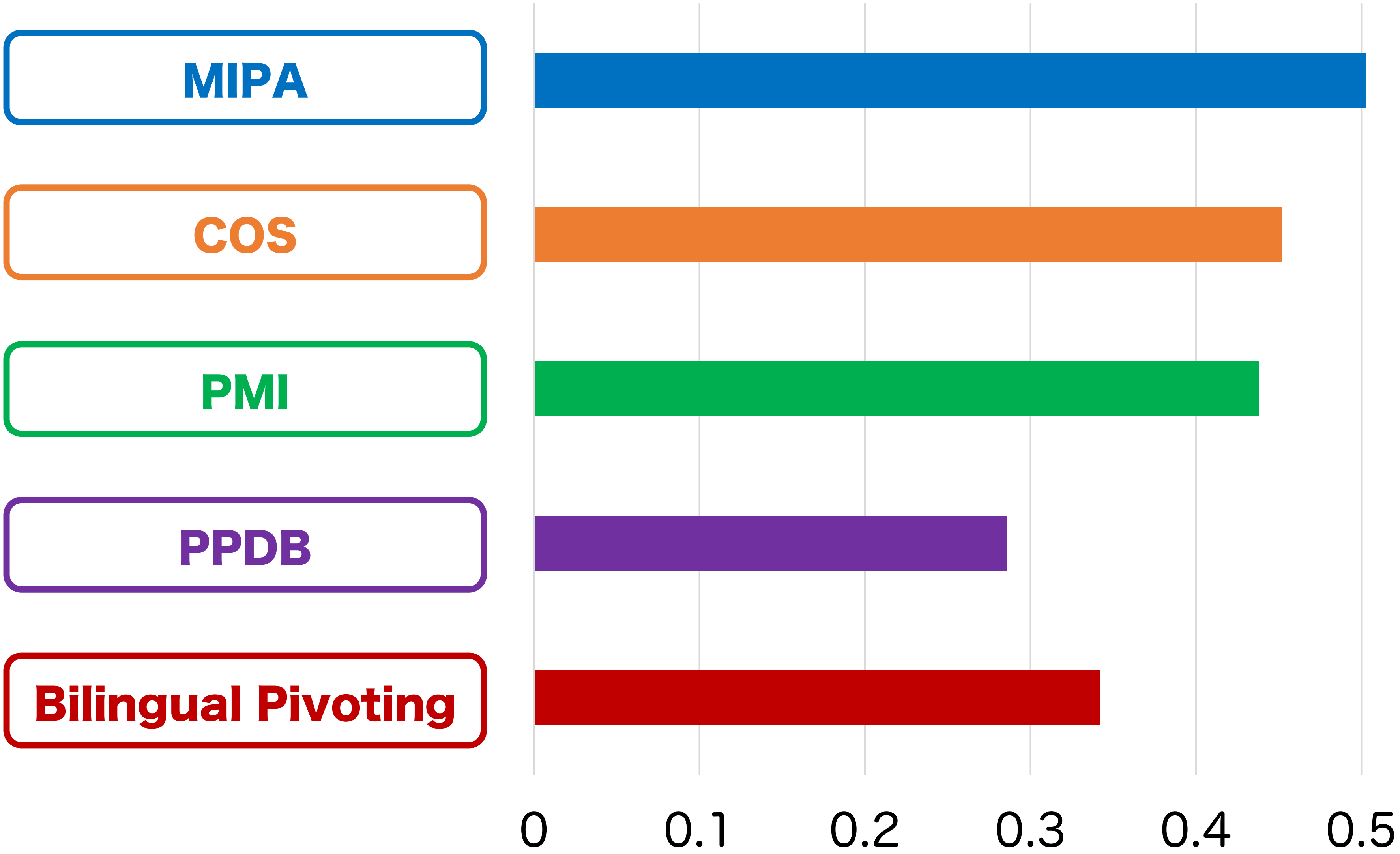
- **PMI** is inaccurate in higher-ranked paraphrases due to the low-frequency bias.
- **MIPA** greatly improved by combining with **COS**.

Mean Average Precision



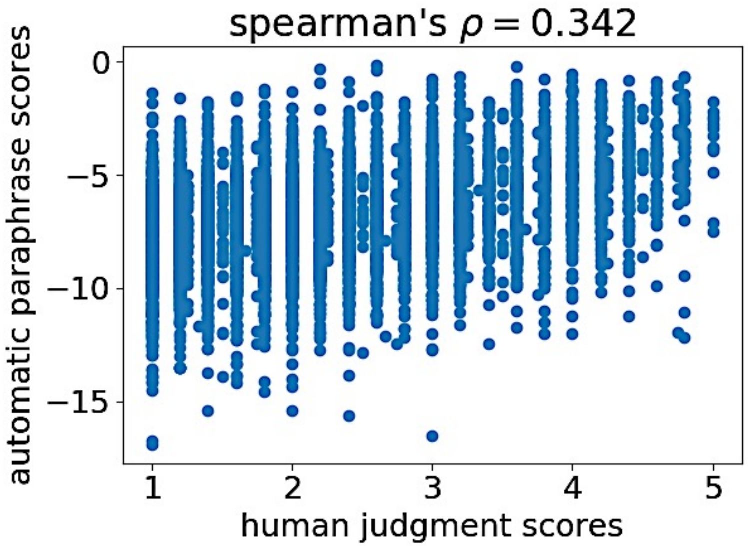
- **PMI** is inaccurate in higher-ranked paraphrases due to the low-frequency bias.
- **MIPA** greatly improved by combining with **COS**.

Spearman's Correlation Coefficient

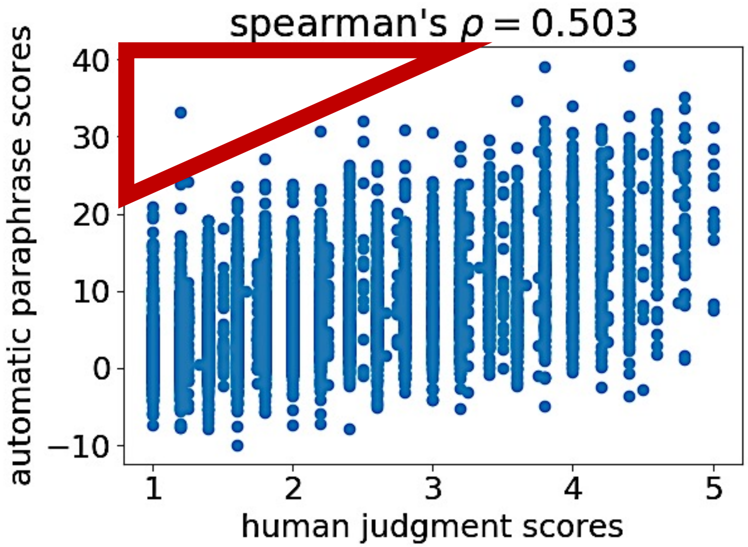


MIPA succeeded in reducing False Positives

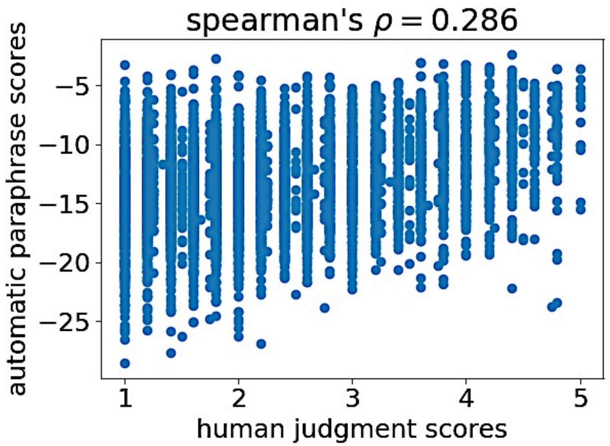
Bilingual Pivoting



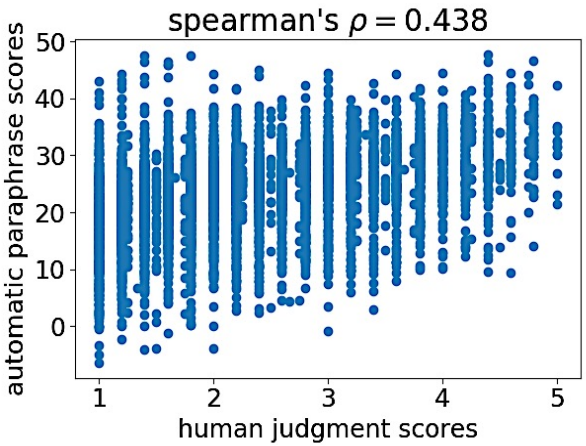
MIPA



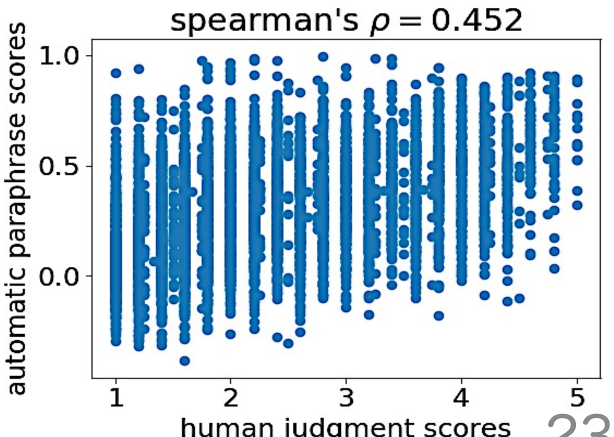
PPDB



PMI



COS



Top-10 paraphrase examples of "cultural"

	Bilingual Pivoting	PPDB	PMI	COS	MIPA
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

MIPA can exclude noise and low-frequency words.

Extrinsic Evaluation: Semantic Textual Similarity

- STS task deals with estimating the semantic similarity [0.0, 1.0] between two sentences.
- We conducted the evaluation by applying Pearson's correlation coefficient with a five-step manual evaluation using five datasets (SemEval-2012 ~ SemEval-2016).

Similarity	Sentence Pair
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.

PAS: Paraphrase Alignment Similarity [Sultan+ 2015]

- This is an unsupervised STS method computed based on PPDB
- PAS achieved excellent results in the STS task of SemEval-2015

The bird is bathing in the sink .

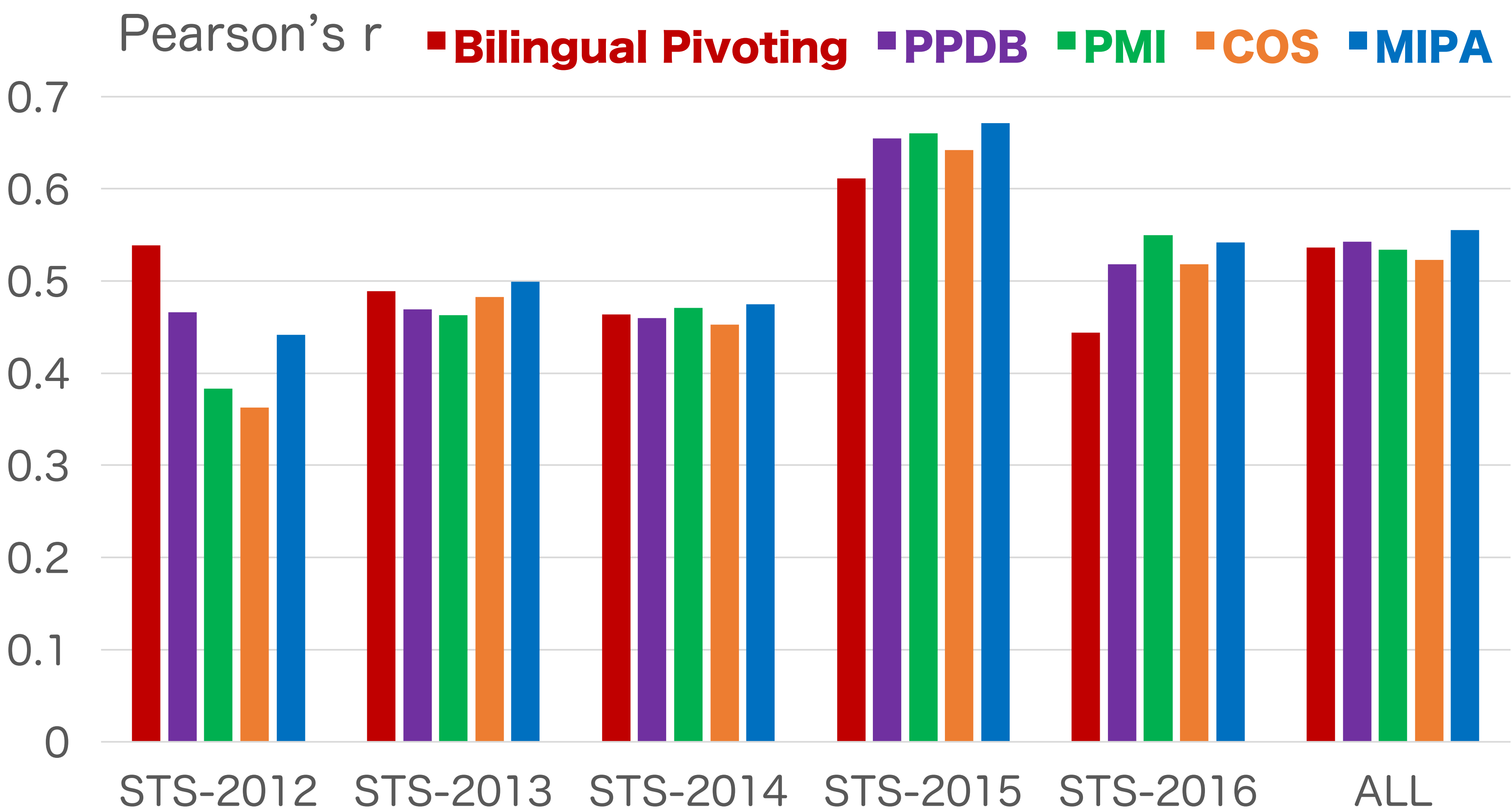
Birdie is washing itself in the water basin .

$$\mathrm{PAS}(x, y) = \frac{\mathrm{PA}(x, y) + \mathrm{PA}(y, x)}{|x| + |y|}$$

$$\mathrm{PA}(x, y) = \sum_{i=1}^{|x|} \begin{cases} 1 & \exists j : x_i \Leftrightarrow y_j \in y \\ 0 & \text{otherwise} \end{cases}$$

where $x_i \Leftrightarrow y_j$ holds if and only if the word pair (x_i, y_j) is included in PPDB

PAS with Top-10 paraphrases



MIPA: Mutual Information Based Paraphrase Acquisition via Bilingual Pivoting

- We generalized lexical synonymity using weighted PMI.

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

- The complementary nature of information from bilingual corpora and from monolingual corpora helps MIPA on paraphrase acquisition accurately.