MIPA: Mutual Information Based Paraphrase Acquisition via Bilingual Pivoting

Tomoyuki Kajiwara* **Mamoru Komachi†** **Daichi Mochihashi‡**

*Tokyo Metropolitan University, Tokyo, Japan, kajiwara-tomoyuki@ed.tmu.ac.jp
†Tokyo Metropolitan University, Tokyo, Japan, komachi@tmu.ac.jp
‡The Institute of Statistical Mathematics, Tokyo, Japan, daichi@ism.ac.jp

Abstract

We present a pointwise mutual information (PMI) based approach for formalizing paraphrasability and propose a variant of PMI, called mutual information based paraphrase acquisition (MIPA), for paraphrase acquisition. Our paraphrase acquisition method first acquires lexical paraphrase pairs by bilingual pivoting and then reranks them by PMI and distributional similarity. The complementary nature of information from bilingual corpora and from monolingual corpora renders the proposed method robust. Experimental results show that the proposed method substantially outperforms bilingual pivoting and distributional similarity themselves in terms of metrics such as mean reciprocal rank, mean average precision, coverage, and Spearman's correlation.

1 Introduction

Paraphrases are useful for flexible language understanding in many NLP applications. For example, the usefulness of the paraphrase database PPDB (Ganitkevitch et al., 2013; Pavlick et al., 2015), a publicly available large-scale resource for lexical paraphrasing, has been reported for tasks such as learning word embeddings (Yu and Dredze, 2014) and semantic textual similarity (Sultan et al., 2015). In PPDB, paraphrase pairs are acquired via word alignment on a bilingual corpus by a process called bilingual pivoting (Bannard and Callison-Burch, 2005). Figure 1 shows an example of English language paraphrase acquisition using the German language as a pivot.

Although bilingual pivoting is widely used for paraphrase acquisition, it always includes noise due to unrelated word pairs caused by word alignment errors on the bilingual corpus. Distributional similarity, another well-known method for paraphrase acquisition, is free of alignment errors, but includes noise due to antonym pairs that share the same contexts on the monolingual corpus (Mohammad et al., 2013).

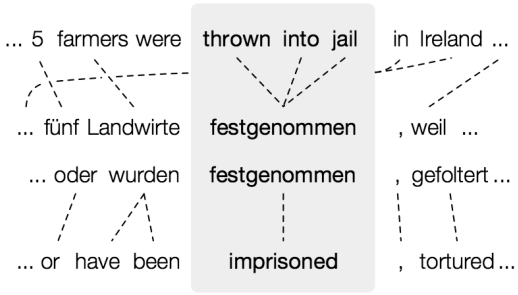

Figure 1: Paraphrase acquisition via bilingual pivoting (Ganitkevitch et al., 2013).

In this study, we formalize the paraphrasability of paraphrase pairs acquired via bilingual pivoting using pointwise mutual information (PMI) and reduce the noise by reranking the pairs using distributional similarity. The proposed method extends Local PMI (Evert, 2005), which is a variant of weighted PMI that aims to avoid low-frequency bias in PMI, for paraphrase acquisition. Since bilingual pivoting and distributional similarity have different advantages and disadvantages, we combine them to construct a complementary paraphrase acquisition method, called mutual information based paraphrase acquisition (MIPA). Experimental results show that MIPA outperforms bilingual pivoting and distributional similarity themselves in terms of metrics such as mean reciprocal rank (MRR), mean average precision (MAP), coverage, and Spearman's correlation.

The contributions of our study are as follows.

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 80–89,
Taipei, Taiwan, November 27 – December 1, 2017

- Bilingual pivoting-based lexical paraphrase acquisition is generalized using PMI.
- Lexical paraphrases are acquired robustly using both bilingual and monolingual corpora.
- We release our lexical paraphrase pairs[1].

2 Bilingual Pivoting

Bilingual pivoting (Bannard and Callison-Burch, 2005) is a method used to acquire large-scale lexical paraphrases by two-level word alignment on a bilingual corpus. Bilingual pivoting employs a conditional paraphrase probability $p(e_2|e_1)$ as a paraphrasability measure, when word alignments exist between an English phrase e_1 and a foreign language phrase f, and between the foreign language phrase f and another English phrase e_2 on a bilingual corpus. It calculates the probability from an English phrase e_1 to another English phrase e_2 using word alignment probabilities $p(f|e_1)$ and $p(e_2|f)$; here, the foreign language phrase f is used as the pivot.

$$\begin{aligned} p(e_2|e_1) &= \sum\nolimits_f p(e_2|f,e_1)\,p(f|e_1) \\ &\approx \sum\nolimits_f p(e_2|f)\,p(f|e_1) \end{aligned} \quad (1)$$

It assumes conditional independence of e_1 and e_2 given f, so that the equation above can be estimated easily using phrase-based statistical machine translation models. One of its advantages is that it requires only two translation models to acquire paraphrases on a large scale. However, since the conditional probability is asymmetric, it may introduce irrelevant paraphrases that do not hold the same meaning as the original one. In addition, owing to the data sparseness problem in the bilingual corpus, paraphrase probabilities may be overestimated for low-frequency word pairs.

To mitigate this, PPDB (Ganitkevitch et al., 2013) defined the symmetric paraphrase score $s_{bp}(e_1,e_2)$ using bi-directional bilingual pivoting.

$$s_{bp}(e_1,e_2) = -\lambda_1 \log p(e_2|e_1) - \lambda_2 \log p(e_1|e_2) \quad (2)$$

Unlike Equation (1), s_{bp} enforces mutual paraphrasability of e_1 and e_2. As discussed later, this does not necessarily increase the performance of paraphrase acquisition, because the symmetric constraint may be too strict to allow the extraction of broad-coverage paraphrases. In this study, without loss of generality, we set[2] $\lambda_1 = \lambda_2 = -1$.

$$s_{bp}(e_1,e_2) = \log p(e_2|e_1) + \log p(e_1|e_2) \quad (3)$$

Although these paraphrase acquisition methods can extract large-scale paraphrase knowledge, the results may contain many fragments caused by word alignment error.

3 MIPA: Mutual Information Based Paraphrase Acquisition

To mitigate overestimation, we acquire lexical paraphrases with the conditional paraphrase probability by using Kneser-Ney smoothing (Kneser and Ney, 1995) and reranking them using information theoretic measure from a bilingual corpus and distributional similarity calculated from a large-scale monolingual corpus.

3.1 Smoothing of Bilingual Pivoting

Since bilingual pivoting adopts the conditional probability $p(e_2|e_1)$ as paraphrasability, we can mitigate the problem of overestimation by applying a smoothing method.

In the hierarchical Bayesian model, the conditional probability $p(y|x)$ is expressed using the Dirichlet distribution with parameter α_y and maximum likelihood estimation $\hat{p}_{y|x}$ as follows.

$$\begin{aligned} p(y|x) &= \frac{n(y|x)+\alpha_y}{\sum_y (n(y|x)+\alpha_y)} \\ &\simeq \frac{n(y|x)}{n(x)+\sum_y \alpha_y} \qquad \because \alpha_y \ll 1 \\ &= \frac{n(x)}{n(x)+\sum_y \alpha_y} \cdot \frac{n(y|x)}{n(x)} \\ &= \frac{n(x)}{n(x)+\sum_y \alpha_y} \cdot \hat{p}_{y|x} \end{aligned} \quad (4)$$

Here, $n(x)$ indicates the frequency of a word x and $n(y|x)$ indicates the co-occurrence frequency of word y following x. As $\sum_y \alpha_y$ is too large to be ignored, especially when the frequency $n(x)$ is small, Equation (4) shows that the maximum likelihood estimation $\hat{p}_{y|x}$ estimates the probability to be excessively large.

Therefore, we propose using Kneser-Ney smoothing (Kneser and Ney, 1995), which is considered to be an extension of the Dirichlet smoothing above, to mitigate overestimation of paraphrase probability in bilingual pivoting.

[1] https://github.com/tmu-nlp/pmi-ppdb

[2] PPDB[3]: $\lambda_1 = \lambda_2 = 1$

[3] http://www.cis.upenn.edu/~ccb/ppdb/

$$
\begin{aligned}
p_{\mathrm{KN}}(e_2|e_1) &= \frac{n(e_2|e_1)-\delta}{n(e_1)} + \gamma(e_1)p_{\mathrm{KN}}(e_2) \\
\delta &= \frac{N_1}{N_1+2N_2} \\
\gamma(e_1) &= \frac{\delta}{n(e_1)}N(e_1) \\
p_{\mathrm{KN}}(e_2) &= \frac{N(e_2)}{\sum_i N(e_i)}
\end{aligned} \tag{5}
$$

Here, N_n indicates the number of types of word pairs of frequency n and $N(e_1)$ indicates the number of types of paraphrase candidates of word e_1.

3.2 Generalization of Bilingual Pivoting using Mutual Information

The bi-directional bilingual pivoting of PPDB (Ganitkevitch et al., 2013) constrains paraphrase acquisition to be strictly symmetric. However, although it is extremely effective for extracting synonymous expressions, it tends to give high scores to frequent but irrelevant phrases, since bilingual pivoting itself contains noisy phrase pairs because of word alignment errors.

To address the problem of frequent phrases, we smooth paraphrasability by bilingual pivoting in Equation (3) using word probabilities $p(e_1)$ and $p(e_2)$ from a monolingual corpus that is sufficiently larger than the bilingual corpus.

$$
\begin{aligned}
s_{pmi}(e_1,e_2) = &\log p(e_2|e_1) + \log p(e_1|e_2) \\
&- \log p(e_1) - \log p(e_2)
\end{aligned} \tag{6}
$$

Thus, we can interpret the bi-directional bilingual pivoting as an unsmoothed version of PMI. Since the difference in the logarithms of the numerator and denominator is equal to the logarithm of the quotient, we can transform Equation (6) as

$$
\begin{aligned}
s_{pmi}(e_1,e_2) &= \log\frac{p(e_2|e_1)}{p(e_2)} + \log\frac{p(e_1|e_2)}{p(e_1)} \\
&= 2\mathrm{PMI}(e_1,e_2)
\end{aligned} \tag{7}
$$

since we can transform PMI into the following forms using Bayes' theorem.

$$
\begin{aligned}
\mathrm{PMI}(x,y) &= \log\frac{p(x,y)}{p(x)p(y)} \\
&= \log\frac{p(y|x)p(x)}{p(x)p(y)} = \log\frac{p(y|x)}{p(y)} \\
&= \log\frac{p(x|y)p(y)}{p(x)p(y)} = \log\frac{p(x|y)}{p(x)}
\end{aligned} \tag{8}
$$

Plugging Equation (8) into Equation (7), we can interpret PMI as a geometric mean of two models.

$$
\begin{aligned}
\mathrm{PMI}(x,y) &= \frac{1}{2}\mathrm{PMI}(x,y) + \frac{1}{2}\mathrm{PMI}(x,y) \\
&= \frac{1}{2}\log\frac{p(y|x)}{p(y)} + \frac{1}{2}\log\frac{p(x|y)}{p(x)} \\
&= \log\left[\left\{\frac{p(y|x)}{p(y)}\right\}^{\frac{1}{2}} \cdot \left\{\frac{p(x|y)}{p(x)}\right\}^{\frac{1}{2}}\right]
\end{aligned} \tag{9}
$$

Bilingual pivoting in Equation (1) can be regarded as a mixture model that considers only the $e_1 \rightarrow e_2$ direction. However, as shown in Equation (9), our proposed method can be regarded as a product model (Hinton, 2002) that considers both directions. PPDB (Pavlick et al., 2015) also considers the paraphrase probability in both directions, but the authors did not regard it as a product model; instead the paraphrase probability in each direction is treated as one of the features of supervised learning.

3.3 Incorporating Distributional Similarity

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, Evert (2005) proposed Local PMI, which assigns weights to PMI depending on the co-occurrence frequency of word pairs.

$$
\mathrm{LocalPMI}(x,y) = n(x,y)\cdot\mathrm{PMI}(x,y) \tag{10}
$$

In this study, however, it was difficult to directly calculate the weight corresponding to $n(x,y)$ in Equation (10) on the bilingual corpus. Furthermore, our aim was to calculate not the strength of co-occurrence (relation) between words, but the paraphrasability. Therefore, it is not appropriate to count the co-occurrence frequency on a monolingual corpus such as Local PMI.

Alternatively, we use as a weight the distributional similarity, which is frequently used for paraphrase acquisition from a monolingual corpus (Chan et al., 2011; Glavaš and Štajner, 2015).

$$
\begin{aligned}
s_{lpmi}(e_1,e_2) &= cos(e_1,e_2)\cdot s_{pmi}(e_1,e_2) \\
&= cos(e_1,e_2)\cdot 2\mathrm{PMI}(e_1,e_2)
\end{aligned} \tag{11}
$$

Here, $cos(e_1,e_2)$ indicates cosine similarity between vector representations of word e_1 and word e_2. Equation (11) simultaneously considers paraphrasability based on the monolingual corpus (distributional similarity) and on the bilingual corpus

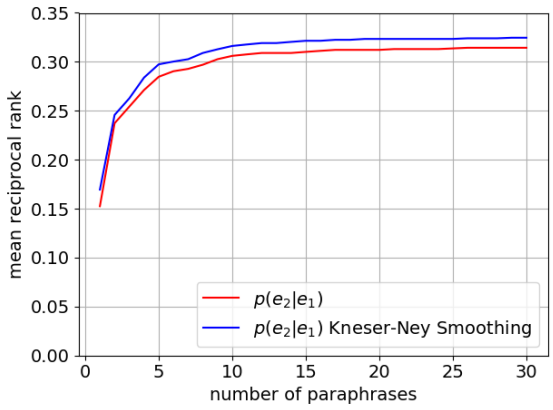

Figure 2: Effectiveness of smoothing of bilingual pivoting evaluated by paraphrase ranking in MRR.

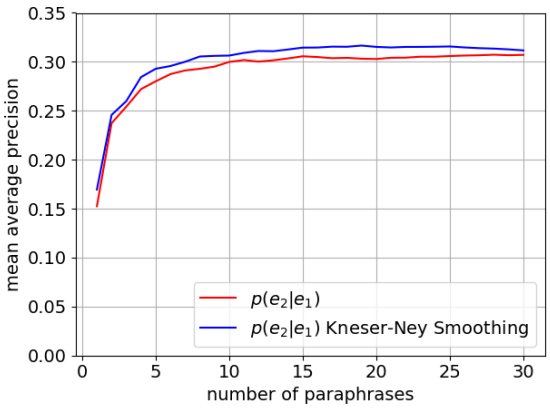

Figure 3: Effectiveness of smoothing of bilingual pivoting evaluated by paraphrase ranking in MAP.

(bilingual pivoting). Distributional similarity, as opposed to bilingual pivoting, is robust against noise associated with unrelated word pairs. Bilingual pivoting is robust against noise arising from antonym pairs, unlike distributional similarity. Therefore, $s_{lpmi}(e_1, e_2)$ can perform paraphrase acquisition robustly by compensating the disadvantages. Hereinafter, we refer to $s_{lpmi}(e_1, e_2)$ as MIPA, mutual information based paraphrase acquisition via bilingual pivoting.

4 Experiments

4.1 Settings

We used French-English parallel data[4] from Europarl-v7 (Koehn, 2005) and GIZA++ (Och and Ney, 2003) (IBM model 4) to calculate the conditional paraphrase probability $p(e_2|e_1)$ and $p(e_1|e_2)$. We also used the English Gigaword 5th Edition[5] and KenLM (Heafield, 2011) to calculate the word probability $p(e_1)$ and $p(e_2)$. For $cos(e_1, e_2)$, we used the CBOW model[6] of word2vec (Mikolov et al., 2013a). Finally, we acquired paraphrase candidates of 170,682,871 word pairs, excepting the paraphrase of itself ($e_1 = e_2$).

We employed the conditional paraphrase probability of bilingual pivoting given in Equation (1), the symmetric paraphrase score of PPDB given by Equation (3), and distributional similarity as baselines, and compared them with PMI shown in Equation (7) and the MIPA score given in Equation (11). Note that distributional similarity implies that the paraphrase pairs acquired via bilingual pivoting were reranked by distributional similarity rather than by using the top-k distributionally similar words among all the vocabularies.

4.2 Evaluation Datasets and Metrics

For evaluation, we used two datasets included in Human Paraphrase Judgments[7] published by Pavlick et al. (2015); hereafter, we call these datasets HPJ-Wikipedia and HPJ-PPDB, respectively.

First, Human Paraphrase Judgments includes a paraphrase list of 100 words or phrases randomly extracted from Wikipedia and processed using a five-step manual evaluation for each paraphrase pair (HPJ-Wikipedia). A correct paraphrase is a word that gained three or more evaluations in manual evaluation. We used this dataset to evaluate the acquired paraphrase pairs by MRR and MAP, following Pavlick et al. (2015). Furthermore, we evaluated the coverage of the top-k paraphrase pairs. Function words such as "as" have more than 50,000 types of paraphrase candidates, because they are sensitive to word alignment errors in bilingual pivoting. However, since many of these paraphrase candidates are word pairs that are not in fact paraphrases, we evaluated the coverage in terms of the extent to which they can reduce unnecessary candidates while preserving the correct paraphrases.

Second, Human Paraphrase Judgments also includes a five-step manual evaluation of 26,456 word pairs sampled from PPDB (Ganitkevitch et al., 2013) (HPJ-PPDB)

[4] http://www.statmt.org/europarl/
[5] https://catalog.ldc.upenn.edu/LDC2011T07
[6] https://code.google.com/archive/p/word2vec/
[7] http://www.seas.upenn.edu/~epavlick/data.html

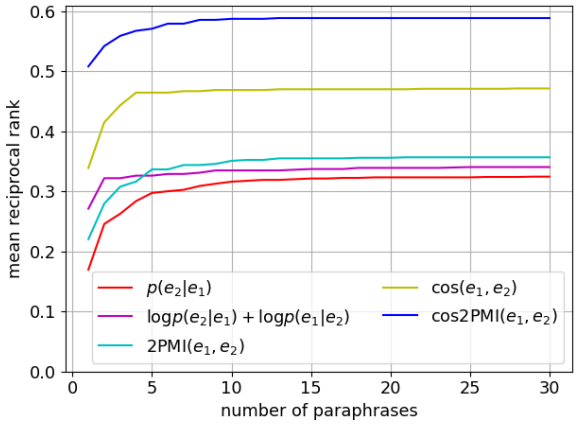

Figure 4: Paraphrase ranking in MRR.

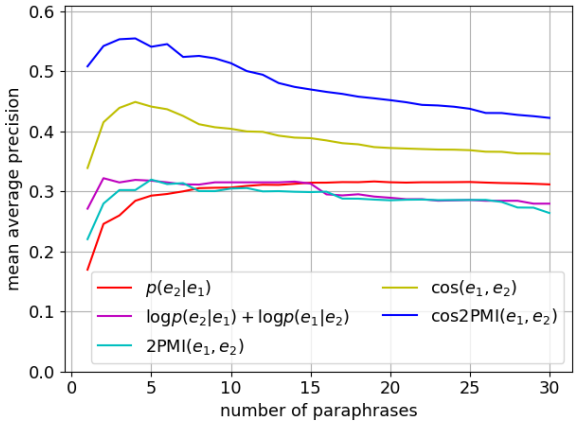

Figure 5: Paraphrase ranking in MAP.

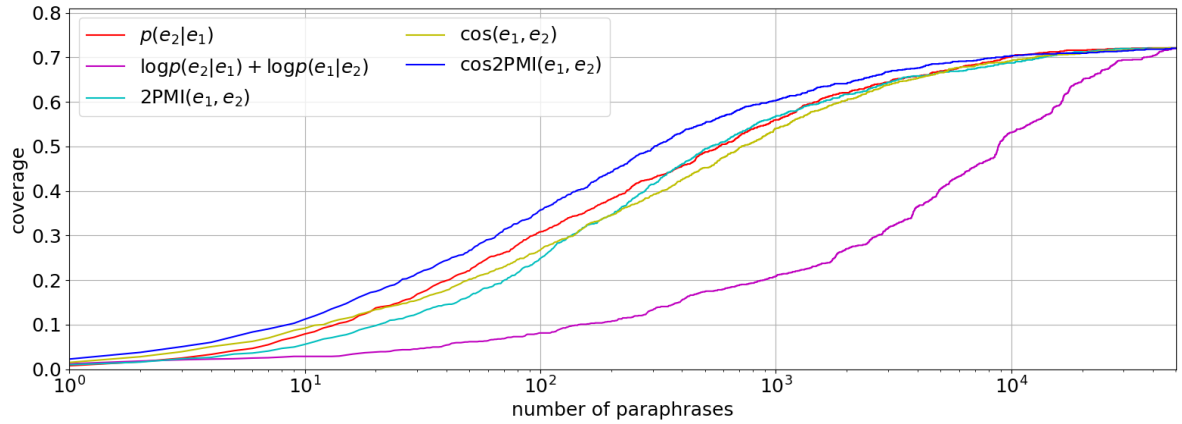

Figure 6: Coverage of the top-k paraphrase pairs.

together with the paraphrase list of 100 words. We used this dataset to evaluate the overall paraphrase ranking based on Spearman's correlation coefficient, as in Pavlick et al. (2015).

4.3 Results

Figures 2 and 3 show the effectiveness of adopting Kneser-Ney smoothing for bilingual pivoting in terms of MRR and MAP on HPJ-Wikipedia. The horizontal axis of each graph represents the evaluation of the paraphrase up to the top-k of the paraphrase score. The results confirm that the ranking of paraphrases acquired via bilingual pivoting was improved by applying Kneser-Ney smoothing. In the rest of this study, we always applied Kneser-Ney smoothing to conditional paraphrase probability.

Figures 4 and 5 show the comparison of paraphrase rankings in MRR and MAP on HPJ-Wikipedia. The evaluation by MRR, shown in Figure 4, demonstrates that the ranking performance of paraphrase pairs is improved by making bilingual pivoting symmetric. PMI slightly outperforms the baselines of bilingual pivoting below the top five. Furthermore, MIPA shows the highest performance, because reranking by distributional similarity greatly improves bilingual pivoting.

The evaluation using MAP, shown in Figure 5, also reinforced the same result, i.e., reranking by distribution similarity improved bilingual pivoting, and MIPA showed the highest performance.

Figure 6 shows the coverage of the top-ranked paraphrases on HPJ-Wikipedia. Despite the fact that the symmetric paraphrase score is better than the conditional paraphrase probability in the ranking performance of the top three in MRR and MAP, it shows a poor performance in terms of coverage. Although there is not a significant difference between MIPA and the other methods, MIPA was shown to outperform them.

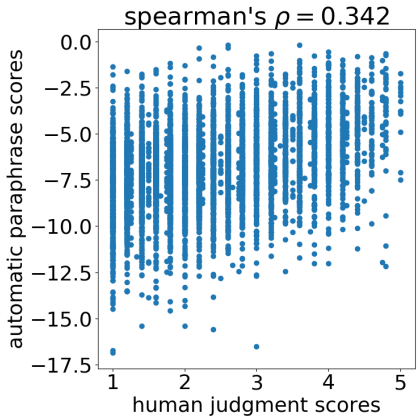

Figure 7: $\rho : \log p(e_2 | e_1)$.

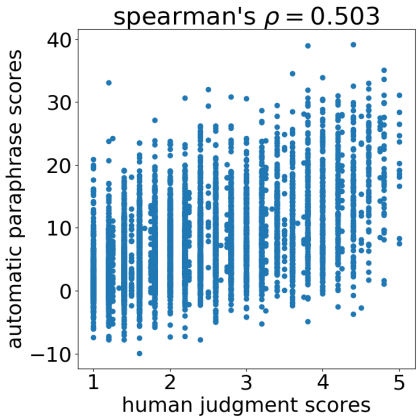

Figure 8: $\rho : \mathrm{MIPA}(e_1, e_2)$.

	$p(e_2 \mid e_1)$	$\log p(e_2 \mid e_1) + \log p(e_1 \mid e_2)$	$2\mathrm{PMI}(e_1, e_2)$	$\cos(e_1, e_2)$	$\cos(e_1, e_2)2\mathrm{PMI}(e_1, e_2)$
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

Table 1: Paraphrase examples of *cultural*. Italicized words are the correct paraphrases.

Figures 7 and 8 show the scatter plots and Spearman's correlation coefficient of each paraphrase score and manual evaluation (average value of five evaluators) on HPJ-PPDB. As in the previous experimental results, MIPA showed a high correlation. In particular, the noise generated by false positives at the upper left of the scatter plot can be reduced by combining PMI and distributional similarity.

5 Discussion

5.1 Qualitative Analysis

Table 1 shows examples of the top 10 in paraphrase rankings. In the paraphrase examples of *cultural*, conditional paraphrase probability does not score the correct paraphrase as top-ranked words. Although the symmetric paraphrase score ranked the correct paraphrase at the top, words other than the top words are less reliable, as shown by the previous experimental results. PMI is strongly influenced by low-frequency words, and many of the top-ranked words are singleton words in the bilingual corpus. MIPA, in contrast, mitigates the problem of low-frequency bias, and many of the top-ranked words are correct paraphrases. Distributional similarity-based methods include relatively numerous correct paraphrases at the top, and the other top-ranked words are also strongly related to *cultural*. From the viewpoint of paraphrases, 3 of the top 10 words of the proposed method are incorrect, but these words may also be useful for applications such as learning word embeddings (Yu and Dredze, 2014) and semantic textual similarity (Sultan et al., 2015).

Table 2 shows correct examples of the paraphrase rankings. In the paraphrase examples of *labourers*, there were 20 correct paraphrases that received a rating of 3 or higher in manual evaluation. With respect to the conditional paraphrase probability and PMI, it is necessary to consider up to the 400th place to cover all correct paraphrases. However, distributional similarity-based methods have correct paraphrases of higher rank. In particular, MIPA was able to include 10 words of correct paraphrases in the top 20 words; that is, our method can obtain paraphrases with high coverage by using only the highly ranked paraphrases.

$p(e_2\|e_1)$		$\log p(e_2\|e_1) + \log p(e_1\|e_2)$		2PMI(e_1, e_2)		$\cos(e_1, e_2)$		$\cos(e_1, e_2)$2PMI(e_1, e_2)	
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

Table 2: Correct paraphrase examples of *labourers*.

	$p(e_2\|e_1)$	$\log p(e_2\|e_1) + \log p(e_1\|e_2)$	2PMI(e_1, e_2)	$\cos(e_1, e_2)$	$\cos(e_1, e_2)$2PMI(e_1, e_2)
STS-2012	**0.539**	0.466	0.383	0.363	0.442
STS-2013	0.489	0.469	0.463	0.483	**0.499**
STS-2014	0.464	0.460	0.471	0.453	**0.475**
STS-2015	0.611	0.655	0.660	0.642	**0.671**
STS-2016	0.444	0.518	**0.550**	0.518	0.542
ALL	0.536	0.543	0.534	0.523	**0.555**

Table 3: Evaluation by Pearson's correlation coefficient in semantic textual similarity task.

5.2 Quantitative Analysis

Next, we applied the acquired paraphrase pairs to the semantic textual similarity task and evaluated the extent to which the acquired paraphrases improve downstream applications. The semantic textual similarity task deals with calculating the semantic similarity between two sentences. In this study, we conducted the evaluation by applying Pearson's correlation coefficient with a five-step manual evaluation using five datasets constructed by SemEval (Agirre et al., 2012, 2013, 2014, 2015, 2016). We applied the acquired paraphrase pairs to the unsupervised method of DLC@CU (Sultan et al., 2015), which achieved excellent results using PPDB in the semantic textual similarity task of SemEval-2015 (Agirre et al., 2015). DLS@CU performs word alignment (Sultan et al., 2014) using PPDB, and calculates sentence similarity according to the ratio of aligned words:

$$sts(s_1, s_2) = \frac{n_a(s_1) + n_a(s_2)}{n(s_1) + n(s_2)} \qquad (12)$$

Here, $n(s)$ indicates the number of words in sentence s and $n_a(s)$ indicates the number of aligned words. Although DLS@CU targets all the paraphrases of PPDB, we used only the top 10 words of the paraphrase score for each target word and compared the performance of the paraphrase scores.

Table 3 shows the experimental results of the semantic textual similarity task. "ALL" is the weighted mean value of the Pearson's correlation coefficient over the five datasets. MIPA achieved the highest performance on three out of the five datasets. In other words, the proposed method extracted paraphrase pairs useful for calculating sentence similarity at the top-rank.

5.3 Reranking PPDB 2.0

Finally, we reranked paraphrase pairs from a publicly available state-of-the-art paraphrase database.[8] PPDB 2.0 (Pavlick et al., 2015) scores paraphrase pairs using supervised learning with

[8] http://paraphrase.org/

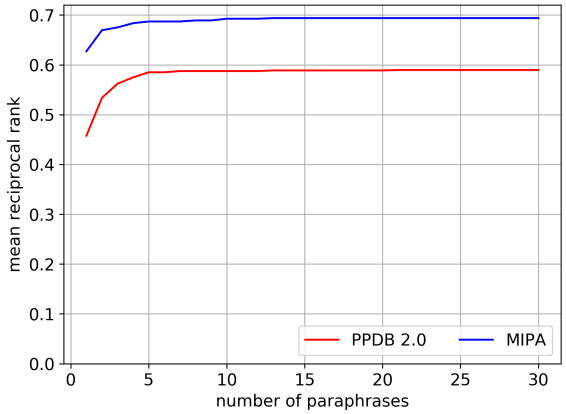

Figure 9: Reranking PPDB 2.0 in MRR.

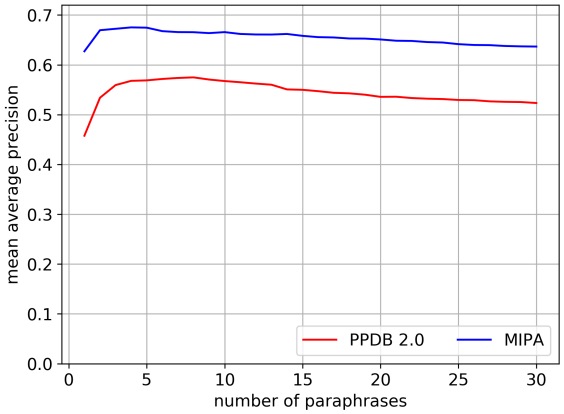

Figure 10: Reranking PPDB 2.0 in MAP.

26,455 labeled data and 209 features. We sorted the paraphrase pairs from PPDB 2.0 using the MIPA instead of the PPDB 2.0 score and used the same evaluation means as described in Section 4. Surprisingly, our unsupervised approach outperformed the paraphrase ranking performance of PPDB 2.0's supervised approach in terms of MRR (Figure 9) and MAP (Figure 10).

6 Related Work

Levy and Goldberg (2014) explained a well-known representation learning method for word embeddings, the skip-gram with negative-sampling (SGNS) (Mikolov et al., 2013a,b), as a matrix factorization of a word-context co-occurrence matrix with shifted positive PMI. In this paper, we explained a well-known method for paraphrase acquisition, bilingual pivoting (Bannard and Callison-Burch, 2005; Ganitkevitch et al., 2013), as a (weighted) PMI.

Chan et al. (2011) reranked paraphrase pairs acquired via bilingual pivoting using distributional similarity. The main idea of reranking paraphrase pairs using information from a monolingual corpus is similar to ours, but Chan et al.'s method failed to acquire semantically similar paraphrases. We succeeded in acquiring semantically similar paraphrases because we effectively combined information from a bilingual corpus and a monolingual corpus by using weighted PMI.

In addition to English, paraphrase databases are constructed in many languages using bilingual pivoting (Bannard and Callison-Burch, 2005). Ganitkevitch and Callison-Burch (2014) constructed paraphrase databases[8] in 23 languages, including European languages and Chinese. Furthermore, Mizukami et al. (2014) constructed the Japanese version[9]. In this study, we improved bilingual pivoting using a monolingual corpus. Since large-scale monolingual corpora are easily available for many languages, our proposed method may improve paraphrase databases in each of these languages.

PPDB (Ganitkevitch et al., 2013) constructed by bilingual pivoting is used in many NLP applications, such as learning word embeddings (Yu and Dredze, 2014), semantic textual similarity (Sultan et al., 2015), machine translation (Mehdizadeh Seraj et al., 2015), sentence compression (Napoles et al., 2016), question answering (Sultan et al., 2016), and text simplification (Xu et al., 2016). Our proposed method may improve the performance of many of these NLP applications supported by PPDB.

7 Conclusion

We proposed a new approach for formalizing lexical paraphrasability based on weighted PMI and acquired paraphrase pairs using information from both a bilingual corpus and a monolingual corpus. Our proposed method, MIPA, uses bilingual pivoting weighted by distributional similarity to acquire paraphrase pairs robustly, as each of the methods complements the other. Experimental results using manually annotated datasets for lexical paraphrase showed that the proposed method outperformed bilingual pivoting and distributional similarity in terms of metrics such as MRR, MAP, coverage, and Spearman's correlation. We also confirmed the effectiveness of the proposed method

[9] http://ahclab.naist.jp/resource/jppdb/

by conducting an extrinsic evaluation on a semantic textual similarity task. In addition to the semantic textual similarity task, we hope to improve the performance of many NLP applications based on the results of this study.

Acknowledgements

This research was (partly) supported by Grant-in-Aid for Research on Priority Areas, Tokyo Metropolitan University, "Research on social bigdata."

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation*. pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. pages 497–511.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*. pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics*. pages 32–43.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. pages 597–604.

Tsz Ping Chan, Chris Callison-Burch, and Benjamin Van Durme. 2011. Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. pages 33–42.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The Multilingual Paraphrase Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. pages 4276–4283.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 758–764.

Goran Glavaš and Sanja Štajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. pages 63–68.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. pages 187–197.

Geoffrey E. Hinton. 2002. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation* 14(8):1771–1800.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-off for M-gram Language Modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. volume 1, pages 181–184.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Machine Translation Summit*. pages 79–86.

Omer Levy and Yoav Goldberg. 2014. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems*. pages 2177–2185.

Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. 2015. Improving Statistical Machine Translation with a Multilingual Paraphrase Database. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 1379–1390.

Tomas Mikolov, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at the International Conference on Learning Representations*. pages 1–12.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems*. pages 3111–3119.

Masahiro Mizukami, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Building a Free, General-Domain Paraphrase Database for Japanese. In *Proceedings of the 17th Oriental COCOSDA Conference*. pages 129 – 133.

Saif M. Mohammad, Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing Lexical Contrast. *Computational Linguistics* 39(3):555–590.

Courtney Napoles, Chris Callison-Burch, and Matt Post. 2016. Sentential Paraphrasing as Black-Box Machine Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*. pages 62–66.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1):19–51.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. pages 425–430.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the Association for Computational Linguistics* 2:219–230.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. pages 148–153.

Md Arafat Sultan, Vittorio Castelli, and Radu Florian. 2016. A Joint Model for Answer Sentence Ranking and Answer Extraction. *Transactions of the Association for Computational Linguistics* 4:113–125.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics* 4:401–415.

Mo Yu and Mark Dredze. 2014. Improving Lexical Embeddings with Semantic Knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. pages 545–550.