PREFER: Using a Graph-Based Approach to Generate Paraphrases for Language Learning

Mei-Hua Chen[*], **Shih-Ting Huang**[+], **Chung-Chi Huang**[*], **Hsien-Chin Liou**[**], **Jason S. Chang**[+]
[*] Institute of Information Systems and Applications
[+] Department of Computer Science
[**] Department of Foreign Languages and Literature
National Tsing Hua University
HsinChu, Taiwan, R.O.C. 30013
{chen.meihua,koromiko1104,u901571,hsienchin,jason.jschang}@gmail.com

Abstract

Paraphrasing is an important aspect of language competence; however, EFL learners have long had difficulty paraphrasing in their writing owing to their limited language proficiency. Therefore, automatic paraphrase suggestion systems can be useful for writers. In this paper, we present PREFER[1], a paraphrase reference tool for helping language learners improve their writing skills. In this paper, we attempt to transform the paraphrase generation problem into a graphical problem in which the phrases are treated as nodes and translation similarities as edges. We adopt the PageRank algorithm to rank and filter the paraphrases generated by the pivot-based paraphrase generation method. We manually evaluate the performance of our method and assess the effectiveness of PREFER in language learning. The results show that our method successfully preserves both the semantic meaning and syntactic structure of the query phrase. Moreover, the students' writing performance improve most with the assistance of PREFER.

1. Introduction

Paraphrasing, or restating information using different words, is an essential part of productive language competence (Fuchs, 1980; Mel'čuk, 1992; Martinot, 2003). However, EFL learners have difficulty paraphrasing in their writing partly because of their insufficient lexical knowledge (Abasi et al. 2006; Chandrasoma et al. 2004). If they are provided with direct and substantial support while writing, they may be able to express their thoughts more fluently. Unfortunately, few paraphrase reference tools have been developed to provide instant assistance to learners in their writing process. In the light of the pressing need for paraphrase reference tools, we develop PREFER, a paraphrasing assistant system to help EFL learners vary their expression during writing.

Over the past decade, paraphrasing techniques have played an important role in many areas of Natural Language Processing, such as machine translation, and question answering. However, very few studies have been conducted concerning the application of automatic paraphrase generation techniques in language learning and teaching.

In this paper, we treat the paraphrase generation problem as a graph-related problem. We adopt the PageRank algorithm (Page et al., 1999) to generate paraphrases based on the assumption that a page with more incoming links is likely to receive a higher rank. Meanwhile, a page which is linked by a higher ranked page should transitively be ranked higher. We take advantage of transitivity of relevance to rank and filter the paraphrases generated by the pivot-based method (i.e., phrase are treated as paraphrases if they share the same translations) of Bannard and Callison-Burch (2005).

The advantage of the pivot approach is that the generated paraphrases are exactly semantically equivalent to the query phrase. However, its

[1] http://140.114.89.231/PREFER

The 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pages 80–85,
Montréal, Canada, June 3-8, 2012.

quality of the paraphrases highly correlates with that of the techniques of bilingual alignment. To overcome such limitation, we use the PageRank algorithm to refine the generated paraphrases. In other words, we leverage the PageRank algorithm to find more relevant paraphrases that preserve both meaning and grammaticality for language learners. The results of a manual evaluation and a system assessment show that our approach and system perform well.

2. Related Work

A number of studies have investigated EFL leaners' paraphrase competence. For example, Campbell (1987) reveals that language proficiency significantly affects paraphrasing competence. McInnis (2009) reports that paraphrasing task is more difficult for L2 students than that for L1 students. According to Milicevic (2011), L2 learners propose less valid paraphrases than native speakers. These findings indicate that EFL students have problems in paraphrasing. In view of this, we develop PREFER, a paraphrase reference tool, for helping English learners with their writing.

Paraphrase generation, on the other hand, has been an area of active research and the related work has been thoroughly surveyed in Androutsopoulos and Malakasiotis (2010) as well as in Madnani and Dorr (2010). In the rest of this section, we focus on reviewing the methods related to our work.

One prominent approach to paraphrase generation is based on bilingual parallel corpora. For example, Bannard and Callison-Burch (2005) propose the pivot approach to generate phrasal paraphrases from an English-German parallel corpus. With the advantage of its parallel and bilingual natures of such a corpus, the output paraphrases do preserve semantic similarity. Callison-Burch (2008) further places syntactic constraints on generated paraphrases to improve the quality of the paraphrases. In this paper, we generate paraphrases adopting the pivot-based method proposed by Bannard and Callison-Burch (2005) in the first round. Then we use a graph-based approach to further ensure paraphrase candidates preserve *both* meaning *and* grammaticality.

In a study more closely related to our work, Kok and Brockett (2010) take a graphical view of the pivot-based approach. They propose the Hitting Time Paraphrase algorithm (HTP) to measure similarities between phrases. The smaller the number of steps a random walker goes from one node to the other, the more likely these two nodes are paraphrases. The main difference between their work and ours lies in the definition of the graph. While they treat multilingual phrases as nodes, we treat only English phrases as nodes. Besides, we define the edges between nodes as semantic relation instead of bilingual alignment.

In contrast to the previous work, we present a graph-based method for refining the paraphrases generated by the pivoting approach. Our goal is to consolidate the relation between paraphrases to provide learners with more and better paraphrases which are helpful in expanding their lexical knowledge.

3. Graph-Based Paraphrase Generation

In this section, we describe how we use the PageRank algorithm to rank and filter the paraphrases generated by the pivot-based method.

3.1 Graph Construction

We first exploit the pivot-based method proposed by Bannard and Callison-Burch (2005) to populate our graph G using of candidate paraphrases *cP*={ $cp_1, cp_2, \dots, cp_n$ } from a bilingual parallel corpus B for a query phrase q. Each phrase in *cP* is also represented as a node in *G*. Note that the query phrase *q* is excluded from *cP*.

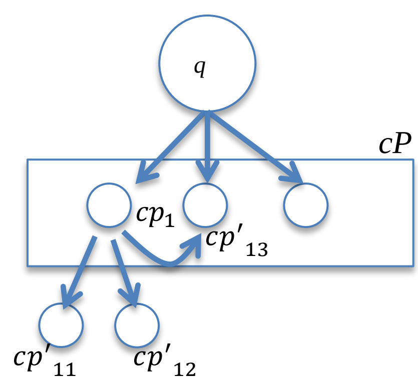

Figure 1. A simple graph *G*. Note that the cp_1 and cp'_{13} will be linked iff cp'_{13} is the paraphrase of q and is also the paraphrase of cp_1.

Graph G only contains the paraphrases cp_i whose probabilities are higher than a certain threshold ε^2 as nodes. In addition, each cp_i is linked to the query phrase q with edge e which is weighted by the probability $P(cp_i|q)$. Furthermore, we establish the edges among the phrases in cP. An example graph is shown in Figure 1. By repeating the previous steps, for each phrase cp_{1}, cp_{2},... in cP, we find their corresponding paraphrases, $cp'_{11}, cp'_{12}, cp'_{13}, \ldots$ and $cp'_{21}, cp'_{22}, cp'_{23}$, …., and discard the paraphrases that are not in cP. Once the phrases are linked with their paraphrases, the graph G is created.

In this paper, we also place a constraint that a paraphrase of a phrase q must neither be a substring nor a superstring of q. These strings are usually aligned with the same foreign language phrase while they are not paraphrases at all. For example, "*play an important*" and "*play an important role in*" are excluded for "*play an important role*". This has the effect of reducing some of the noise generated by the pivot-based method.

3.2 Graph-Based Paraphrase Generation

We then refine the generated paraphrases adopting the PageRank algorithm proposed by Page et al. (1999). Consider a graph consisting of a set of webpages on the Web V and a set of hyperlinks E. The PageRank algorithm assigns a value PR to each webpage as their importance measurement. The PR value of a certain page u is defined iteratively as the following equation:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \qquad (1)$$

where B_u is a set of pages linked to u and $L(.)$ denotes the number of *outbound links* from a page v.

Intuitively, by using formula (1) iteratively, we are able to calculate the PR values for all nodes and thus extract relatively important paraphrases. However, the original PageRank algorithm does not take the weight of each edge into consideration. That is, the PageRank algorithm treats all links equally when distributing rank scores. Treating all links equally in paraphrase generation task might lose some linguistic properties. For this, we consider the importance of edges of the nodes and weight the edges based on the paraphrase probability in the pivot-based approach using

$$w(u, v) = \sum_f P(f|v)P(u|f) \qquad (2).$$

Formula (2) represents the probability that the phrase u is the paraphrase of the phrase v. f refers to shared translations of v and u. Then for each iteration of the PageRank calculation, we reassign the PR value for all u in V to be $PR'(u)$ as:

$$PR'(u) = \sum_{v \in B_u} \frac{w(u,v)PR(v)}{L(v)} \qquad (3)$$

Instead of treating all edges equally, formula (3) integrates the weights of inbound link and outbound link edges (see Section 4 for the performance differences with and without weighting edges).

4. Results

In this section, we first present our experimental setting. Then evaluation results are reported.

4.1 Experimental Setting

In this paper, word alignments were produced by Giza++ toolkit (Och and Ney, 2003) over a set of Danish-English section (containing 1,236,427 sentences) of the Europarl corpus, version 2 (Koehn, 2002).

We compared our graph-based approach with a strong baseline, the pivot-based method with syntactic constraint (SBP) (Callison-Burch, 2008) utilizing the same Danish-English corpus. We also investigate the contribution of adding the edge weights to the PageRank algorithm by building two models, **PR** representing the method of the PageRank algorithm without weights and **PRw** representing the method of the weighted PageRank algorithm, for comparison.

To assess the performance of our method, we conducted a manual evaluation. We asked an experienced English lecturer to randomly select 100 most commonly used and meaningful phrases from 30 research articles in the discipline of Computer-Assisted Language Learning (CALL). A total of 88 unique phrases were used as our test set for evaluation excluding 12 phrases not existing in the Europarl corpus. For each phrase, we extracted

[2] We set ε to be 0.01.

the corresponding candidate paraphrases and chose top 5 for evaluation. Two raters, provided with a simplified scoring standard used by Callison-Burch (2008), manually evaluate the accuracy of the top ranked paraphrases of each phrase by score 0, 1 and 2. It is worth noting that the raters were asked to score each paraphrase candidate by considering its appropriateness in various contexts. In this evaluation, we strictly deemed a paraphrase to be correct if and only if both raters scored 2. The inter-annotator agreement was 0.63.

The coverage was measured by the number of correct answers within top 5 candidates. The precision was measured by the number of correct answers within the returned answers.

On the other hand, to assess the effectiveness of PREFER in language learning, we carried out an experiment with 55 Chinese-speaking EFL college freshmen, who had at least six years of formal instruction from junior to senior high schools and were estimated to be at the intermediate level regarding their overall English competence. The students were randomly divided into three groups. They were asked to paraphrase seven short paragraphs in the pre-test with no system support, and then paraphrase another seven short paragraphs in the post-test using three different tools: PREFER (P), *LONGMAN Dictionary of Contemporary English Online* (L), and *Thesaurus.com* (T). A total of 22 default phrases (http://140.114.75.22/share/examples.htm) were embedded in the paragraphs in the pre- and post-tests, targeted at comparing the quality and quantity of students' paraphrasing performance. Students were not restricted to paraphrase these embedded phrases. Instead, they were encouraged to replace any possible phrases or even restructure sentences. We had two experienced native-speaker TESL (Teaching English as a Second Language) lecturers to score the students' paraphrasing performance.

4.2 Experimental Results

4.2.1 Manual Evaluation

As shown in Table 1, **PRw** achieved both good precision and coverage. Moreover, **PR** and **PRw** performed better than **SBP** in both coverage and precision. Also, the result that the performance of **PRw** is better than that of **PR** implies that **PRw** is able to generate more semantically and syntactically correct paraphrases. However, the precision of 0.19 indicates that there is still room to improve the paraphrase generation model.

	PR	**PRw**	**SBP**
Coverage	0.17	**0.18**	0.07
Precision	0.17	**0.19**	0.10

Table 1: The measurement of paraphrases.

Additionally, Mean Reciprocal Rank (MRR) is also reported. Here, MRR is defined as a measure of how much effort needed to locate the first appropriate paraphrase for the given phrase in the ranked list of paraphrases. The MRR score of **PRw** (0.53) outperformed **PR** (0.51) and **SBP** (0.47). It demonstrated that the **PRw** model facilitates the high ranking of good paraphrases (i.e., paraphrases with meaning and grammaticality preserved would be ranked high).

4.2.2. Evaluation on Language Learning

The second evaluation is to assess the effectiveness of PREFER applied to CALL. We used a comparison method to measure the extent to which EFL learners achieved good performance in paraphrasing.

		P	L	T
improvement of paraphrasing task		**38.2%**	-31.6%	-6.2%
all paraphrasable phrases	rephrased	**38.4%**	-23.2%	9.5%
	correct	**53.3%**	-17.5%	4.6%
	correctness rate	**7.9%**	4.9%	-3.1%
22 default phrases	rephrased	**68%**	-16%	28%
	correct	**100%**	-5%	31%
	correctness rate	**13.6%**	7.9%	1.5%

Table 2. Comparison of paraphrasing performance among students using three different reference tools.

As seen in the first row of Table 2, the students' writing performance improved most with the assistance of PREFER (i.e., group P), compared with group L and group T. We further analyzed and compared the number of the rephrased phrases and the correct paraphrases, and the rate of correctness students achieved using different reference tools among our testing paraphrase candidates (see the middle and bottom panels of Table 2). Obviously, the students consulting PREFER achieved substantial paraphrasing improvement in all three aspects of both all and default phrases. But the other two groups seemed unable to manage well the paraphrasing task with traditional way of phrase information. This limited information seems insufficient to enable students to familiarize themselves with proper usages of phrases which might lead to improper paraphrasing.

In short, PREFER outperformed the other two reference tools in assisting EFL learners in their paraphrasing task.

5. Conclusion and Future Work

In this paper, we treat the paraphrase generation problem as a graphical problem. We utilize the PageRank algorithm to rank and filter the paraphrases generated using the pivot-based method. The results show that our method significantly produces better paraphrases in both precision and coverage compared with the syntactically-constrained pivot method of Callison-Burch (2008). Additionally, PREFER does benefit learners' writing performance.

In order to conduct a more comprehensive evaluation, we plan to adapt the in-context evaluation metric introduced by Callison-Burch et. al (2008). A larger test set would be generated manually to evaluate the performance of our paraphrase system. In addition, we will implement various kinds of baseline systems such as Kok and Brockett (2010) and Chan et al. (2011) to provide a more competitive comparison.

Many avenues exist for future research and improvement. For example, we would like to extend paraphrasing consecutive n-gram phrases to inconsecutive ones such as ones with incomplete transitive verbs (e.g., "***provide*** *someone* ***with*** *something*"). Besides, we are interested in weighting edges using syntactic and semantic relation in our graph-based method to further improve the quality of generated paraphrases.

References

Ali R. Abasi, Nahal Akbari, and Barbara Graves. 2006. Discourse appropriation, construction of identities, and the complex issue of plagiarism: ESL students writing in graduate school. *Journal of Second Language Writing*, 15, 102–117.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research* 38, 135–187.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*, pp. 597-604.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of EMNLP*, pp. 196–205.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2008. ParaMetric: an automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 97-104.

Cherry Campbell. 1990. Writing with others' words: Using background reading text in academic compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 211-30). Cambridge, UK: Cambridge University Press.

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*, pp. 33-42.

Ranamukalage Chandrasoma, Celia Thompson, and Alastair Pennycook. 2004. Beyond plagiarism: Transgressive and nontransgressive intertextuality. *Journal of Language, Identity, and Education*, *3*(3), 171–194.

Catherine Fuchs. 1980. *Paraphrase et théorie du langage. Contribition à une histoire des theories linguistiques contemporaines et à la construction d'une théorie énonciative de la paraphraase*. Thèse de doctorat. Paris: Université Paris VII.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished, http://www.isi.edu/~koehn/europarl/.

Stanley Kok and Chris Brockett. 2010. Hitting the right paraphrases in good time. In *Proceedings of NAACL/HLT*, pp. 145-153.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–388.

Claire Martinot. 2003. Pour une linguistique de l'acquisition. La reformulation: du concept descriptif au concept explicatif. *Langage et Société*, 2(104); 147-151.

Lara McInnis. 2009. *Analyzing English L1 and L2 Paraphrasing Strategies through Concurrent Verbal Report and Stimulated Recall Protocols*. MA Thesis. Toronto: University of Toronto.

Igor Mel'čuk. 1992. Paraphrase et lexique: la théorie Sens-Texte et le Dictionnaire explicatif et combinatoire. In: Mel'čuk, I. et al., *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III*. Montréal: Presses de l'Université de Montréal; 9-59.

Jasmina Milićević and Alexandra Tsedryk. 2011. Assessing and Improving Paraphrasing Competence in FSL. In *Proceedings of the 5th International Conference on Meaning- Text Theory*, pp. 175-184.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19-51.

Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. 1999. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report*. pp. 1999-66, Stanford University InfoLab.