Augmentable Paraphrase Extraction Framework

Mei-Hua Chen[ac], **Yi-Chun Chen**[b], **Shih-Ting Huang**[b], **Jason S. Chang**[b]
[a]Institute of Information Systems and Applications and [b]Department of Computer Science, National Tsing Hua University, HsinChu, Taiwan, R.O.C. 30013
[c]Department of Foreign Languages and Literature, Hua Fan University, Taipei, Taiwan, R.O.C. 22301
{chen.meihua, pieyaaa, koromiko1104,jason.jschang}@gmail.com

Abstract

Paraphrase extraction relying on a single factor such as distribution similarity or translation similarity might lead to the loss of some linguistic properties. In this paper, we propose a paraphrase extraction framework, which accommodates various linguistically motivated factors to optimize the quality of paraphrase extraction. The major contributions of this study lie in the augmentable paraphrasing framework and the three kinds of factors conducive to both semantic and syntactic correctness. A manual evaluation showed that our model achieves more successful results than the state-of-the-art methods.

1. Introduction

Paraphrasing provides an alternative way to express an idea using different words. Early work on paraphrase acquisition has been mainly based on either distributional similarity (e.g., Lin and Pantel, 2001) or the pivot-based approach (e.g., Bannard and Callison-Burch, 2005). Both methods have their strengths and limitations. Distributional similarity is capable of extracting syntactically correct paraphrases, but may risk including antonymous phrases as paraphrases. On the other hand, the pivot approach has the advantage of preserving semantic similarity among the generated paraphrases; however, the quality and quantity of the paraphrases closely correlates with the techniques of bilingual phrase alignment.

Considering single factors, existing paraphrasing methods could lose some linguistic properties. In view of this, we attempt to differentiate the importance of the paraphrase candidates based on various factors. In this paper, we take a graphical view of the paraphrasing issue. To achieve the goal mentioned above, we adopt the Weighted PageRank Algorithm (Xing and Ghorbani, 2004). English phrases are treated as nodes. The edge weights are determined by various factors such as semantic similarity or syntactic similarity between nodes. It means that the performance of the ranked paraphrase candidates depends on the factors we selected and added. In other words, our framework is augmentable and is able to accommodate various factors to optimize the quality of paraphrase extraction.

In this case, we propose three linguistically motivated factors to improve the performance of the paraphrase extraction. Lexical distributional similarity is used to ensure that the contexts in which the generated paraphrases appear are similar whereas syntactic distributional similarity is adopted for the purpose of maintaining the syntactic correctness. Translation similarity, one more factor, is capable of preserving semantic equivalence. These three selected factors adopted together effectively achieve better performance on paraphrase extraction. The evaluation shows that our model achieves more satisfactory results than the state-of-the-art pivot-based methods and graph-based methods.

2. Related Work

Several approaches have been proposed to extract paraphrases. Earlier studies have focused on extracting paraphrases from monolingual corpora. Barzilay and Mckeown (2001) determine that the phrases in a monolingual parallel corpus are paraphrases of one another only if they appear in similar contexts. Lin and Pantel (2001) derive

International Joint Conference on Natural Language Processing, pages 706–711,
Nagoya, Japan, 14-18 October 2013.

paraphrases using parse tree paths to compute distributional similarity. Another prominent approach to paraphrase extraction is based on bilingual parallel corpora. For example, Bannard and Callison-Burch (2005) propose the pivot approach to extract phrasal paraphrases from an English-German parallel corpus. With the advantage of its parallel and bilingual natures of such a corpus, the output paraphrases preserve semantic equivalence. Callison-Burch (2008) further places syntactic constraints on extracted paraphrases to improve the quality of the paraphrases. Chan et al. (2011) use monolingual distributional similarity to rank paraphrases generated by the syntactically-constrained pivot method.

Recently, some studies take a graphical view of the pivot-based approach. Kok and Brockett (2010) propose the Hitting Time Paraphrase algorithm (HTP) to measure the similarities between phrases. Chen et al. (2012) adopt the PageRank algorithm to find more relevant paraphrases that preserve both meaning and grammaticality for language learners. In this paper, we, similarly, present the state-of-the-art approach as a graph. However, unlike Kok and Brockett (2010), we treat English phrases (instead of multilingual phrases) as nodes. On the other hand, different from Chen et al. (2012), our model is augmentable by involving varied linguistic information or domain knowledge.

3. Method

Typically, the state-of-the-art paraphrase extraction models only deal with single factors such as distribution similarity or translation similarity. However, different linguistic factors could facilitate the paraphrase extraction in various ways. With this in mind, we propose an augmentable paraphrase extraction framework based on a graph-based method, which can be modeled with multiple linguistically motivated factors.

In the following section, we describe the graph construction (Section 3.1). Then the paraphrase extraction framework is outlined in Section 3.2. Section 3.3 introduces the three factors we proposed for optimizing the quality of paraphrase extraction. Finally, we utilize the grid search method to fine-tune the parameters of our model.

3.1 Graph Construction

We transform the paraphrase generation problem into a graph-based problem. First, we generate a graph $G \equiv (V,E)$, in which an English phrase is a node $v \in V$ and two nodes are connected by an edge $e \in E$. A set of paraphrase candidates $CP = \{cp_1, cp_2, \ldots, cp_n\}$ is generated for a query phrase q from a bilingual corpus based on the pivot method (Bannard and Callison-Burch, 2005). We further generate a set of transitive paraphrases $CP' = \{ cp'_1, cp'_2, \ldots, cp'_m \}$ of the phrase q, namely, paraphrases cp_i and their paraphrases cp'_j in the same manner. We truncate the paraphrase candidates whose translation similarities are smaller than the threshold ε[1]; we also exclude cp_i that consists only of a stopword or contains q or is contained in q. Thus, some noisy paraphrases are easily eliminated.

Consider the example graph for the query phrase "*on the whole*" shown in Figure 1. We first find its set of candidate paraphrases *CP*, including "*generally speaking*", "*in general*", "*in a nutshell*", using the pivot-based method mentioned above. Then for each phrase in *CP*, we extract the corresponding paraphrases respectively. For example, "*in brief*", "*broadly speaking*", "*in general*" are paraphrases of the first phrase "*generally speaking*" in *CP*. During the process, we keep the extracted paraphrases whose translation similarities are larger than δ[2]. By linking the phrases with their transitive paraphrases, the graph G is created.

3.2 Augmentable Paraphrase Extraction Framework

In this sub-section, we propose an augmentable paraphrase extraction framework, which can be modeled by multiple factors. Considering a graph $G \equiv (V,E)$, the PageRank algorithm assigns a value *PR* to each node as their importance measurement. We further adopt the Weighted PageRank algorithm (Xing and Ghorbani, 2004) to state the relatedness between nodes. We calculate the weight W of the edge which links node v to node u using various factor functions $\mathcal{F}_i$, the weight function is described as follow,

$$W(u,v) = \sum_{i=1}^{N} \lambda_i \mathcal{F}_i(u,v,q)$$

[1] We set ε=0.01.

[2] We set δ = 0.0001.

where q is a query phrase, $\mathcal{F}_i(u,v,q)$ is a factor function and λ_i is the weight of the factor.
The weighted PR value of a certain node u is defined iteratively as:

$$PR(u) = \sum_{v \in R(v)} PR(v)W(u,v)$$

where $R(v)$ is a set of nodes that point to u.

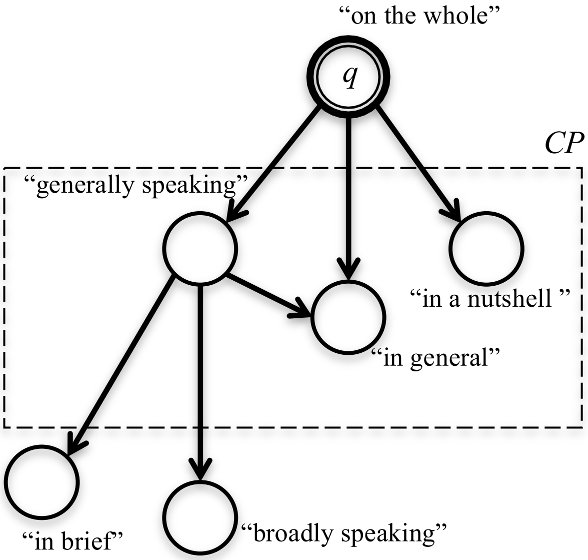

Figure 1. Example graph for the phrase "*on the whole*".

3.3 Linguistically Motivated Factors

Our model enables linguistically motivated factors to optimize the performance of paraphrase extraction. In this sub-section, we introduce three decisive factors: lexical distributional similarity, syntactic distributional similarity and translation similarity.

Lexical distributional similarity factor
Lexical distributional information is to ensure that the contexts in which the generated paraphrases appear are similar. For each phrase p in G, we extract three kinds of context vectors, v_L, v_R, v_{LR} and calculate vector similarities. Vectors v_L and v_R represent two sets of adjacent words which occur in the left and right of p respectively. Words appear simultaneously in both left and right sides of p are also extracted as the feature vector v_{LR}. Each item in vectors is an associated score calculated by pointwise mutual information of the phrase p (Cover and Thomas, 1991).

Given the query phrase q, for each paraphrase candidate u in G, we calculate the cosine similarity of the context vectors, v_L, v_R, v_{LR} between q and u. That is, three factors $\mathcal{F}_{v_L}$, $\mathcal{F}_{v_R}$and $\mathcal{F}_{v_{LR}}$ are described as a cosine similarity function:

$$\mathcal{F}_k = \frac{v_{u_k} \cdot v_{q_k}}{|v_{u_k}||v_{q_k}|}$$

where v_{u_k} denotes a context vector of u, and v_{q_k} a context vector of q and $k \in \{L, R, LR\}$.

Syntactic distributional similarity factor
Calculating the extrinsic syntactic similarity between nodes is used to maintain the syntactic correctness of the generated paraphrases. For each phrase p, we extract three vectors s_L, s_R, s_{LR}, which represents the <POS tag, frequency> pairs that appear on the left, right and both left and right sides of the phrase p. We use the GENIA tagger to obtain POS tags surrounding the phrase p. Each item in vectors is paired with the frequency of the corresponding tag. For each paraphrase candidate u of the query phrase q, we calculate the similarities $\mathcal{F}_{s_L}$, $\mathcal{F}_{s_R}$ and $\mathcal{F}_{s_{LR}}$ between the vectors of u and q using cosine similarity.

$$\mathcal{F}_k = \frac{s_{u_k} \cdot s_{q_k}}{|s_{u_k}||s_{q_k}|}$$

where s_{u_k} denotes a vector of u, and s_{q_k} a vector of q, and $k \in \{L, R, LR\}$.

Translation similarity factor
Next, we calculate the intrinsic translation similarity which is capable of preserving semantic equivalence. Translation similarity factor for an edge connecting node v and u is defined as:

$$\mathcal{F}_{tran} = \sum_{f \in T(v)} P(f|v)P(u|f)$$

where u is one paraphrase of phrase v, $T(v)$ denotes a set of the foreign-language alignment of v, and $P(.)$ the translation probability. Both of the alignment and translation probability are described in Och and Ney (2003).

3.4 Parameter Optimization

Once the factors are selected, we have to determine the weights of the factors, (i.e., λ_i in Section 3.2). In other words, we train the weights of factors such that the performance is optimal for a given developing data set. We use Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen, 2002) to measure the quality of paraphrases. From the top to the bottom of the result list, the DCG score is accumulated with the gain of each result discounted at lower ranks. The DCG score is defined as:

$$DCG(r,c) = \sum_{i=1}^{k} \frac{2^{score_i} - 1}{log_2(i+1)}$$

where r represents a set of manually labeled paraphrase scores, c is a set of paraphrases to be evaluated, and $score_i$ is the paraphrase score at rank i of c.

The parameters[3] are selected in order to maximize the DCG scores in a total of S query phrases from the developing data set:

$$\hat{\lambda}_1^N = arg\ max_{\lambda_1^N} \left\{ \sum_{s=1}^{S} DCG\left(r_s, \hat{c}(p_s, \lambda_1^N)\right) \right\}$$

where $\hat{c}$ is a set of paraphrases of the query phrase p_s, extracted from our model under the parameter values λ_1^N.

In the process, we first assign each parameter a random value ranging from 0 to 1 and use a grid-based line optimization method to optimize the parameters. While optimizing a parameter, we maximize the parameter of certain dimension while the parameters of other dimensions are fixed. The process stops when the values of the parameters do not change in two iterations.

4. Results

4.1 Experimental Setting

In this paper, we adopted the Danish-English section (containing 1,236,427 sentences) of the Europarl corpus, version 2 (Koehn, 2002) for computing distributional similarity and translation similarity. Word alignments were produced by Giza++ toolkit (Och and Ney, 2003). We randomly selected 50 phrases as the developing set for optimizing parameters. For each phrase, three distinct sentences which containing the phrase are randomly sampled. A total of 6073 paraphrases have been labeled score 0 (incorrect), 1 (partially correct), and 2 (correct) by considering the fluency of each sentence for developing optimization.

We compared our augmentable paraphrase extraction framework (**APF**) with three baselines: the syntactically-constrained pivot method (**SBiP**) (Callison-Burch, 2008), syntactically-constrained pivot method using monolingual distributional similarity (**SBiP-MonoDS**) (Chan et al., 2011) and the graph-based method (**GB**) (Chen et al., 2012). To assess the contribution of the parameter optimization, we built another model based on APF with identical weights of factors (**APF-avgW**).

We evaluated the paraphrase quality through a substitution test. We randomly selected 133 most commonly used phrases from 30 research articles. For each phrase, we extracted the corresponding paraphrase candidates and evaluated its top 5 candidates. At the same time, three or less distinct sentences containing the phrase were randomly sampled (a total of 398 sentences were evaluated) from the New York Times section of the English Gigaword (LDC2003T05) to capture the fact that paraphrases are valid in some contexts but not others (Szpektor et al., 2007). Two native speaker judges evaluated whether the candidates are syntactically and semantically appropriate in various contexts. They assigned two values corresponding to the semantic and syntactic considerations to each sentence by score 0, (not acceptable), 1 ("acceptable") and 2 ("acceptable and correct"). The inter-annotator agreement was 0.67.

It is worth noting that we include two measurement schemes for comprehensive analysis. The strict scheme considers a paraphrase as "correct" if and only if both of the two judges scored 2 points, whereas the other one considers a paraphrase as "acceptable" if it is given scores of 1 or 2.

4.2 Experimental Results

We compared the performance of the five models, **SBiP**, **SBiP-MonoDS**, **GB**, **APF-avgW** and **APF**, using the precision, coverage, MRR and DCG. Because the number of paraphrases generated by **SBiP**, **SBip-DS** (101 phrases) and **GB, APF-avgW**, **APF** (131 phrases) are varied, we

[3] In this paper, the parameters are λ_{v_L} = 0.03, λ_{v_R} = 0.01, $\lambda_{v_{LR}}$ = 0.99, λ_{s_L} = 0.00001, λ_{s_R} = 0.00001, $\lambda_{s_{LR}}$ = 0.18 and λ_{tran} = 0.06.

decided to analyze the results of 99 phrases involving 295 sentences which were generated by all five models. Top-*k* precision indicates the percentage of the sentences in which correct paraphrase(s) appear in the top-*k* paraphrase candidates. The coverage was measured by the number of sentences in which at least one out of five paraphrases is correct within all 398 sentences.

Table 1 shows the results of precision and coverage in overall consideration. As can be seen, the **APF** achieved higher precision and coverage than the other four methods.

Additionally, we evaluated the results using MRR. MRR is defined as a measure of how much effort needed for a user to locate the first appropriate paraphrase for the given phrase in the ranked list of paraphrases. As shown in Table 2, the **APF** model performed better than the other models in both correct and acceptable measures. Moreover, Table 3 showed that the **APF** model outperformed the other models in both correct and acceptable measures based on either overall or individual consideration. DCG comprehensively considers both the number of good quality paraphrases and the ranking of these paraphrases. Overall, the **APF** model achieved better performance in paraphrase extraction.

	Top-1 precision	Top-5 precision	Coverage
SBiP	0.13/0.29	0.29/0.59	0.22/0.45
SBip-DS	0.15/0.36	0.32/0.57	0.25/0.44
GB	0.15/0.33	0.29/0.54	0.26/0.51
APF-avgW	0.14/0.39	0.36/**0.66**	0.34/**0.61**
APF	**0.16/0.42**	**0.38**/0.65	**0.36/0.61**

Table 1. Performance of the five models. Note that the former value indicates **correct** measures and the latter one **acceptable** measures.

	Semantic	Syntactic	Both
SBiP	0.19/0.41	0.45/0.52	0.18/0.40
SBip-DS	0.22/0.46	0.48/0.54	0.22/0.45
GB	0.21/0.43	0.51/0.54	0.21/0.42
APF-avgW	0.22/0.50	0.57/0.64	0.21/0.49
APF	**0.24/0.51**	**0.58/0.65**	**0.23/0.50**

Table 2. MRR scores of the five models. Note that the former value indicates **correct** measures and the latter **acceptable** measures.

	Semantic	Syntactic	Both
SBiP	0.24/0.68	0.75/0.89	0.23/0.64
SBip-DS	0.29/0.78	0.86/1.01	0.29/0.73
GB	0.27/0.81	0.96/1.09	0.26/0.75
APF-avgW	0.31/0.93	1.12/1.28	0.31/0.90
APF	**0.33/0.96**	**1.14/1.31**	**0.33/0.93**

Table 3. DCG scores of the five models. Note that the former value indicates **correct** measures and the **latter** acceptable measures.

5. Conclusion

In this paper, we propose a paraphrase extraction framework. Accommodating various linguistically motivated factors, the framework is capable of extracting better paraphrases carrying linguistic features. The results of the manual evaluation demonstrated that the proposed methods achieved performance improvement in terms of precision, coverage, MRR and DCG. The optimized parameters show that the lexical and syntactic distributional similarity factors make a substantial contribution to our model. Specifically, the words as well as the POS tags appear in both left and right sides show satisfactory performance.

However, some further analyses could be conducted in the future. Although the weights of parameters carry the linguistic properties, the proposed factors could be considered separately for examining and comparing the individual effectiveness in our framework. On the other hand, other factors could be taken in consideration. For example, parsing information could be added to the framework to investigate whether or to what extent it contributes to the paraphrasing task.

References

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

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the ACL*, pp. 50–57.

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

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.

Mei-Hua Chen, Shih-Ting Huang, Chung-Chi Huang, Hsien-Chin Liou and Jason S. Chang. 2012. PREFER: Using a Graph-Based Approach to Generate Paraphrases for Language Learning. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pp. 80-85.

Thomas M. Cover, Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley & Sons.

Jane Frodesen. 2002. Developing paraphrasing skills: A pre-paraphrasing mini-lesson. Retrieved September 19, 2012 from www.ucop.edu/dws/lounge/dws_ml_pre_paraphrasing.pdf.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*. 20(4), pp. 422-446.

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.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. In *Proceedings of ACM SIGKDD Conference on Knowledge Discov- ery and Data Mining*, pp. 323–328.

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

Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance based evaluation of entailment rule acquisition. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics,* pp.456-463.

Wenpu Xing and Ali Ghorbani. 2004. Weighted pagerank algorithm. In *Proceedings of the 2nd Annual Conference on Communication Networks and Services Research*, pp. 305–314.