Enlarging Paraphrase Collections through Generalization and Instantiation

Atsushi Fujita
Future University Hakodate
116-2 Kameda-nakano-cho,
Hakodate, Hokkaido, 041-8655, Japan
fujita@fun.ac.jp

Pierre Isabelle **Roland Kuhn**
National Research Council Canada
283 Alexandre-Taché Boulevard,
Gatineau, QC, J8X 3X7, Canada
{Pierre.Isabelle, Roland.Kuhn}@nrc.ca

Abstract

This paper presents a paraphrase acquisition method that uncovers and exploits generalities underlying paraphrases: paraphrase patterns are first induced and then used to collect novel instances. Unlike existing methods, ours uses both bilingual parallel and monolingual corpora. While the former are regarded as a source of high-quality seed paraphrases, the latter are searched for paraphrases that match patterns learned from the seed paraphrases. We show how one can use monolingual corpora, which are far more numerous and larger than bilingual corpora, to obtain paraphrases that rival in quality those derived directly from bilingual corpora. In our experiments, the number of paraphrase pairs obtained in this way from monolingual corpora was a large multiple of the number of seed paraphrases. Human evaluation through a paraphrase substitution test demonstrated that the newly acquired paraphrase pairs are of reasonable quality. Remaining noise can be further reduced by filtering seed paraphrases.

1 Introduction

Paraphrases are semantically equivalent expressions in the same language. Because "equivalence" is the most fundamental semantic relationship, techniques for generating and recognizing paraphrases play an important role in a wide range of natural language processing tasks (Madnani and Dorr, 2010).

In the last decade, automatic acquisition of knowledge about paraphrases from corpora has been drawing the attention of many researchers. Typically, the acquired knowledge is simply represented as pairs of semantically equivalent sub-sentential expressions as in (1).

(1) a. look like ⇔ resemble
b. control system ⇔ controller

The challenge in acquiring paraphrases is to ensure good coverage of the targeted classes of paraphrases along with a low proportion of incorrect pairs. However, no matter what type of resource has been used, it has proven difficult to acquire paraphrase pairs with both high recall and high precision.

Among various types of corpora, monolingual corpora can be considered the best source for high-coverage paraphrase acquisition, because there is far more monolingual than bilingual text available. Most methods that exploit monolingual corpora rely on the Distributional Hypothesis (Harris, 1968): expressions that appear in similar contexts are expected to have similar meaning. However, if one uses purely distributional criteria, it is difficult to distinguish real paraphrases from pairs of expressions that are related in other ways, such as antonyms and cousin words.

In contrast, since the work in (Bannard and Callison-Burch, 2005), bilingual parallel corpora have been acknowledged as a good source of high-quality paraphrases: paraphrases are obtained by putting together expressions that receive the same translation in the other language (pivot language). Because translation expresses a specific meaning more directly than context in the aforementioned approach, pairs of expressions acquired in this manner tend to be correct paraphrases. However, the coverage problem remains: there is much less bilingual parallel than monolingual text available.

Our objective in this paper is to obtain paraphrases that have high **quality** (like those extracted from bilingual parallel corpora via pivoting) but can be generated in large **quantity** (like those extracted

from monolingual corpora via contextual similarity). To achieve this, we propose a method that exploits general patterns underlying paraphrases and uses both bilingual parallel and monolingual sources of information. Given a relatively high-quality set of paraphrases obtained from a bilingual parallel corpus, a set of paraphrase patterns is first induced. Then, appropriate instances of such patterns, i.e., potential paraphrases, are harvested from a monolingual corpus.

After reviewing existing methods in Section 2, our method is presented in Section 3. Section 4 describes our experiments in acquiring paraphrases and presents statistics summarizing the coverage of our method. Section 5 describes a human evaluation of the quality of the acquired paraphrases. Finally, Section 6 concludes this paper.

2 Literature on Paraphrase Acquisition

This section summarizes existing corpus-based methods for paraphrase acquisition, following the classification in (Hashimoto et al., 2011): similarity-based and alignment-based methods.

2.1 Similarity-based Methods

Techniques that use monolingual (non-parallel) corpora mostly rely on the Distributional Hypothesis (Harris, 1968). Because a large quantity of monolingual data is available for many languages, a large number of paraphrase candidates can be acquired (Lin and Pantel, 2001; Paşca and Dienes, 2005; Bhagat and Ravichandran, 2008, etc.). The recipes proposed so far are based on three main ingredients, i.e., features used for representing context of target expression (contextual features), criteria for weighting and filtering features, and aggregation functions.

A drawback of relying only on contextual similarity is that it tends to give high scores to semantically related but non-equivalent expressions, such as antonyms and cousin words. To enhance the precision of the results, filtering mechanisms need to be introduced (Marton et al., 2011).

2.2 Alignment-based Methods

Pairs of expressions that get translated to the same expression in a different language can be regarded as paraphrases. On the basis of this hypothesis, Barzilay and McKeown (2001) and Pang et al. (2003) created monolingual parallel corpora from multiple human translations of the same source. Then, they extracted corresponding parts of such parallel sentences as sub-sentential paraphrases.

Leveraging recent advances in statistical machine translation (SMT), Bannard and Callison-Burch (2005) proposed a method for acquiring sub-sentential paraphrases from bilingual parallel corpora. As in SMT, a translation table is first built on the basis of alignments between expressions, such as words, phrases, and subtrees, across a parallel sentence pair. Then, pairs of expressions (e_1, e_2) in the same language that are aligned with the same expressions in the other language (pivot language) are extracted as paraphrases. The likelihood of e_2 being a paraphrase of e_1 is given by

$$p(e_2|e_1) \quad = \quad \sum_{f \in Tr(e_1,e_2)} p(e_2|f)p(f|e_1), \quad (1)$$

where $Tr(e_1, e_2)$ stands for the set of shared translations of e_1 and e_2. Each factor $p(e|f)$ and $p(f|e)$ is estimated from the number of times e and f are aligned and the number of occurrences of each expression in each language. Kok and Brockett (2010) showed how one can discover paraphrases that do not share any translation in one language by traversing a graph created from multiple translation tables, each corresponding to a bilingual parallel corpus.

This approach, however, suffers from a coverage problem, because both monolingual parallel and bilingual parallel corpora tend to be significantly smaller than monolingual non-parallel corpora. The acquired pairs of expressions include some non-paraphrases as well. Many of these come from erroneous alignments, which are particularly frequent when the given corpus is small.

Monolingual comparable corpora have also been exploited as sources of paraphrases using alignment-based methods. For instance, multiple news articles covering the same event (Shinyama et al., 2002; Barzilay and Lee, 2003; Dolan et al., 2004; Wubben et al., 2009) have been used. Such corpora have also been created manually through crowdsourcing (Chen and Dolan, 2011). However, the availability of monolingual comparable corpora is very limited for most languages; thus, approaches relying on these corpora have typically produced only very

small collections of paraphrases. Hashimoto et al. (2011) found a way around this limitation by collecting sentences that constitute explicit definitions of particular words or phrases from monolingual non-parallel Web documents, pairing sentences that define the same noun phrase, and then finding corresponding phrases in each sentence pair. One limitation of this approach is that it requires a considerable amount of labeled data for both the corpus construction and the paraphrase extraction steps.

2.3 Summary

Existing methods have investigated one of the following four types of corpora as their principal resource[1]: monolingual non-parallel corpora, monolingual parallel corpora, monolingual comparable corpora, and bilingual parallel corpora. No matter what type of resource has been used, however, it has proven difficult to acquire paraphrases with both high recall and precision, with the possible exception of the method in (Hashimoto et al., 2011) which requires large amounts of labeled data.

3 Proposed Method

While most existing methods deal with expressions only at the surface level, ours exploits generalities underlying paraphrases to achieve better coverage while retaining high precision. Furthermore, unlike existing methods, ours uses both bilingual parallel and monolingual non-parallel corpora as sources for acquiring paraphrases.

The process is illustrated in Figure 1. First, a set of high-quality seed paraphrases, P_{Seed}, is acquired from bilingual parallel corpora by using an alignment-based method. Then, our method collects further paraphrases through the following two steps.

Generalization (Step 2): Paraphrase patterns are learned from the seed paraphrases, P_{Seed}.

Instantiation (Step 3): A novel set of paraphrase pairs, P_{Hvst}, is finally harvested from monolingual non-parallel corpora using the learned patterns; each newly acquired paraphrase pair is assessed by contextual similarity.

[1]Chan et al. (2011) used monolingual corpora only for reranking paraphrases obtained from bilingual parallel corpora. To the best of our knowledge, bilingual comparable corpora have never been used as sources for acquiring paraphrases.

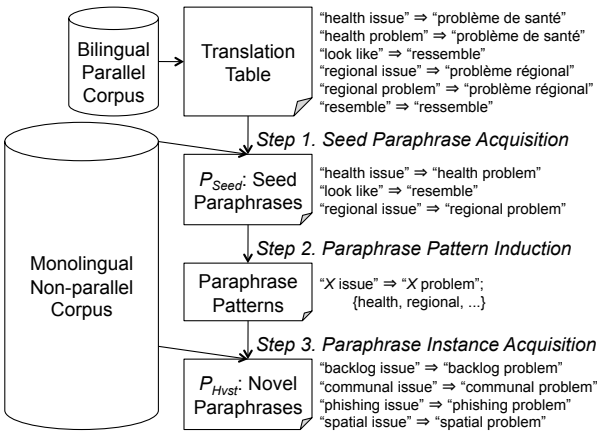


Figure 1: Process of paraphrase acquisition.

The set P_{Seed} acquired early in the process can be pooled with the set P_{Hvst} harvested in the last stage of the process.

3.1 Step 1. Seed Paraphrase Acquisition

The goal of the first step is to obtain a set of high-quality paraphrase pairs, P_{Seed}.

For this purpose, alignment-based methods with bilingual or monolingual parallel corpora are preferable to similarity-based methods applied to non-parallel corpora. Among various options, in this paper, we start from the standard technique proposed by Bannard and Callison-Burch (2005) with bilingual parallel corpora (see also Section 2.2). In particular, we assume the phrase-based SMT framework (Koehn et al., 2003). Then, we purify the results with several filtering methods.

The phrase pair extraction process of phrase-based SMT systems aims at high recall for increased robustness of the translation process. As a result, a naive application of the paraphrase acquisition method produces pairs of expressions that are not exact paraphrases. For instance, the algorithm explained in Koehn (2009, p.134) extracts both "dass" and ", dass" as counterparts of "that" from the sentence pair. To reduce that kind of noise, we apply some filtering techniques to the candidate translation pairs. First, statistically unreliable translation pairs (Johnson et al., 2007) are filtered out. Then, we also filter out phrases made up entirely of stop words (including punctuation marks), both in the language of interest and in the pivot language.

Let P_{Raw} be the initial set of paraphrase pairs extracted from the sanitized translation table. We first

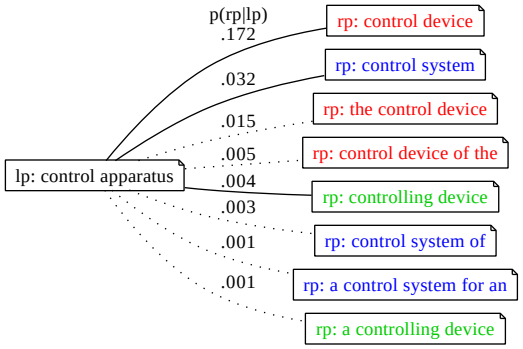


Figure 2: RHS-filtering for "control apparatus".

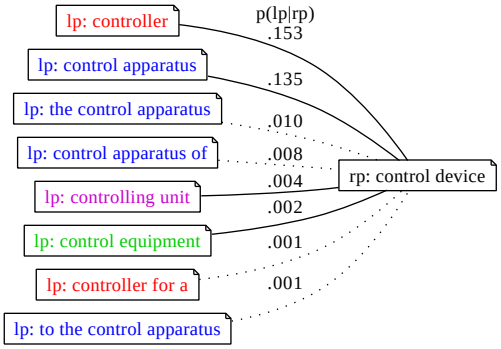


Figure 3: LHS-filtering for "control device".

discard pairs whose difference comprises only stop words, such as "the schools" ⇒ "schools and". We also remove pairs containing only singular-plural differences, such as "family unit" ⇒ "family units". Depending on the language of interest, other types of morphological variants, such as those shown in (2), may also be ignored.

(2) a. "européenne" ⇒ "européen"
(Gender in French)
b. "guten Lösungen" ⇒ "gute Lösungen"
(Case in German)

We further filter out less reliable pairs, such as those shown with dotted lines in Figures 2 and 3. This is carried out by comparing the right-hand side (RHS) phrases of each left-hand side (LHS) phrase, and vice versa[2]. Given a set of paraphrase pairs, RHS phrases corresponding to the same LHS phrase lp are compared. A RHS phrase rp is not licensed iff lp has another RHS phrase rp' ($\neq rp$) which satisfies the following two conditions (see also Figure 2).

- rp' is a word sub-sequence of rp
- rp' is a more likely paraphrase than rp, i.e., $p(rp'|lp) > p(rp|lp)$

LHS phrases for each RHS phrase rp are also compared in a similar manner, i.e., a LHS phrase lp is not qualified as a legitimate source of rp iff rp has another LHS phrase lp' ($\neq lp$) which satisfies the following conditions (see also Figure 3).

- lp' is a word sub-sequence of lp
- lp' is a more likely source than lp, i.e., $p(lp'|rp) > p(lp|rp)$

The two directions of filtering are separately applied and the intersection of their results is retained.

[2]cf. Denkowski and Lavie (2011); they only compared each RHS phrase to its corresponding LHS phrase.

Candidate pairs are finally filtered on the basis of their reliability score. Traditionally, a threshold (th_p) on the conditional probability given by Eq. (1) is used (Du et al., 2010; Max, 2010; Denkowski and Lavie, 2011, etc.). Furthermore, we also require that LHS and RHS phrases exceed a threshold (th_s) on their contextual similarity in a monolingual corpus. This paper neither proposes a specific recipe nor makes a comprehensive comparison of existing recipes for computing contextual similarity, although one particular recipe is used in our experiments (see Section 4.1).

3.2 Step 2. Paraphrase Pattern Induction

From a set of seed paraphrases, P_{Seed}, paraphrase patterns are induced. For instance, from paraphrases in (3), we induce paraphrase patterns in (4).

(3) a. "restraint system" ⇒ "restraint apparatus"
b. "movement against racism"
⇒ "anti-racism movement"
c. "middle eastern countries"
⇒ "countries in the middle east"

(4) a. "X system" ⇒ "X apparatus"
b. "X against Y" ⇒ "anti-Y X"
c. "X eastern Y" ⇒ "Y in the X east"

Word pairs of LHS and RHS phrases will be replaced with variable slots iff they are fully identical or singular-plural variants. Note that stop words are retained. While a deeper level of lexical correspondences, such as "eastern" and "east" in (3c) and "system" and "apparatus" in (3a), could be captured, this would require the use of rich language resources, thereby making the method less portable to resource-poor languages.

Note that our aim is to automatically capture general paraphrase patterns of the kind that have sometimes been manually described (Jacquemin, 1999; Fujita et al., 2007). This is different from approaches that attach variable slots to paraphrases for calculating their similarity (Lin and Pantel, 2001; Szpektor and Dagan, 2008) or for constraining the context in which they are regarded legitimate (Callison-Burch, 2008; Zhao et al., 2009).

3.3 Step 3. Paraphrase Instance Acquisition

Given a set of paraphrase patterns, such as those shown in (4), a set of novel instances, i.e., novel paraphrases, P_{Hvst}, will now be harvested from monolingual non-parallel corpora. In other words, a set of appropriate slot-fillers will be extracted.

First, expressions that match both elements of the pattern, except stop words, are collected from a given monolingual corpus. Pattern matching alone may generate inappropriate pairs, so we then assess the legitimacy of each collected slot-filler.

Let $LHS(\mathbf{w})$ and $RHS(\mathbf{w})$ be the expressions generated by instantiating the k variable slots in LHS and RHS phrases of the pattern with a k-tuple of slot-fillers $\mathbf{w}$ $(= w_1, \ldots, w_k)$, respectively. We estimate how likely $RHS(\mathbf{w})$ is to be a paraphrase of $LHS(\mathbf{w})$ based on the contextual similarity between them using a monolingual corpus; a pair of phrases is discarded if they are used in substantially dissimilar contexts. We use the same recipe and threshold value for th_s with Step 1 in our experiments.

Contextual similarity of antonyms and cousin words can also be high, as they are often used in similar contexts. However, this is not a problem in our framework, because semantic equivalence between $LHS(\mathbf{w})$ and $RHS(\mathbf{w})$ is almost entirely guaranteed as a result of the way the corresponding patterns were learned from a bilingual parallel corpus.

3.4 Characteristics

In terms of coverage, P_{Hvst} is expected to be greatly larger than P_{Seed}, although it will not cover totally different pairs of paraphrases, such as those shown in (1). On the other hand, the quality of P_{Hvst} depends on that of P_{Seed}. Unlike in the pure similarity-based method, P_{Hvst} is constrained by the paraphrase patterns derived from the set of high-quality paraphrases, P_{Seed}, and will therefore generally exclude the kind of semantically similar but non-equivalent pairs that contextual similarity alone tends to extract alongside real paraphrases.

As mentioned in Section 3.1, other types of methods can be used for obtaining high-quality seed paraphrases, P_{Seed}. For instance, the supervised method proposed by Hashimoto et al. (2011) uses the existence of shared words as a feature to determine whether the given pair of expressions are paraphrases, and thereby extracts many pairs sharing the same words. Thus, their output has a high potential to be used as an alternative seed for our method.

Another advantage of our method is that it does not require any labeled data, unlike the supervised methods proposed by Zhao et al. (2009) and Hashimoto et al. (2011).

4 Quantitative Impact

4.1 Experimental Settings

Two different sets of corpora were used as data sources; in both settings, we acquired English paraphrases.

Europarl: The English-French version of the Europarl Parallel Corpus[3] consisting of 1.8M sentence pairs (51M words in English and 56M words in French) was used as a bilingual parallel corpus, while its English side and the English side of the 10^9 French-English corpus[4] consisting of 23.8M sentences (649M words) were used as monolingual data.

Patent: The Japanese-English Patent Translation data (Fujii et al., 2010) consisting of 3.2M sentence pairs (122M morphemes in Japanese and 106M words in English) was used as a bilingual parallel corpus, while its English side and the 30.0M sentences (626M words) from the 2007 chapter of NTCIR unaligned patent documents were used as monolingual data.

To study the behavior of our method for different amounts of bilingual parallel data, we carried out learning curve experiments.

We used our in-house tokenizer for segmentation of English and French sentences and MeCab[5] for Japanese sentences.

[3] http://statmt.org/europarl/, release 6

[4] http://statmt.org/wmt10/training-giga-fren.tar

[5] http://mecab.sourceforge.net/, version 0.98

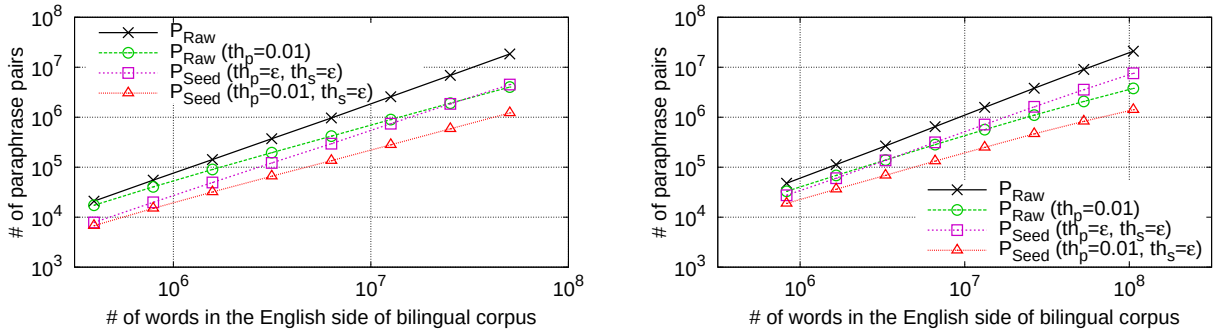


Figure 4: # of paraphrase pairs in P_{Seed} (left: Europarl, right: Patent).

Stop word lists for sanitizing translation pairs and paraphrase pairs were manually compiled: we enumerated 442 English words, 193 French words, and 149 Japanese morphemes, respectively.

From a bilingual parallel corpus, a translation table was created by our in-house phrase-based SMT system, PORTAGE (Sadat et al., 2005). Phrase alignments of each sentence pair were identified by the heuristic "grow-diag-final"[6] with a maximum phrase length 8. The resulting translation pairs were then filtered with the significance pruning technique of (Johnson et al., 2007), using $\alpha + \epsilon$ as threshold.

As contextual features for computing similarity of each paraphrase pair, all of the 1- to 4-grams of words adjacent to each occurrence of a phrase were counted. This is a compromise between less expensive but noisier approaches, such as bag-of-words, and more accurate but more expensive approaches that incorporate syntactic features (Lin and Pantel, 2001; Shinyama et al., 2002; Pang et al., 2003; Szpektor and Dagan, 2008). Contextual similarity is finally measured by taking cosine between two feature vectors.

4.2 Statistics on Acquired Paraphrases

Seed Paraphrases (P_{Seed})

Figure 4 shows the number of paraphrase pairs P_{Seed} obtained from the bilingual parallel corpora. The general trend is simply that the larger the corpus is, the more paraphrases are acquired.

Given the initial set of paraphrases, P_{Raw} ("×"), our filtering techniques ("□") discarded a large portion (63-75% in Europarl and 43-64% in Patent) of them. Pairs with zero similarity were also filtered out, i.e., $th_s = \epsilon$. This suggests that many incorrect and/or relatively useless pairs, such as those shown in Figures 2 and 3, had originally been acquired.

Lines with "○" show the results based on a widely-used threshold value on the conditional probability in Eq. (1), i.e., $th_p = 0.01$ (Du et al., 2010; Max, 2010; Denkowski and Lavie, 2011, etc.). The percentage of paraphrase pairs thereby discarded varied greatly depending on the corpus size (17-78% in Europarl and 31-82% in Patent), suggesting that the threshold value should be determined depending on the given corpus. In the following experiment, however, we conform to the convention $th_p = 0.01$ ("△") to ensure the quality of P_{Seed} that we will be using for inducing paraphrase patterns, even though this results in discarding some less frequent but correct paraphrase pairs, such as "control apparatus" ⇒ "controlling device" in Figure 2.

Paraphrase Patterns

Figures 5 and 6 show the number of paraphrase patterns that our method induced and their coverage against P_{Seed}, respectively. Due to their rather rigid form, the patterns covered no more than 15% of P_{Seed} in Europarl. In contrast, a higher proportion of P_{Seed} in Patent was generalized into patterns. We speculate it is because the patent domain contains many expressions, including technical terms, that have similar variations of constructions.

The acquired patterns were mostly one-variable patterns: 88-93% and 80-91% of total patterns for different variants of the Europarl and Patent settings, respectively. Given that there are far more one-variable patterns than other types, and that one-variable patterns are the simplest type, we henceforth focus on them. More complex patterns, including two-variable patterns (7-11% and 8-17% in each setting), will be investigated in our future work.

[6] http://statmt.org/moses/?n=FactoredTraining.AlignWords

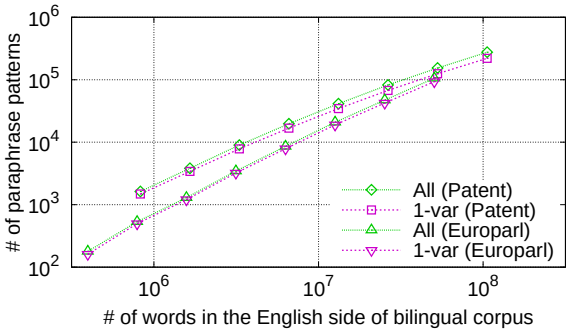


Figure 5: # of paraphrase patterns.

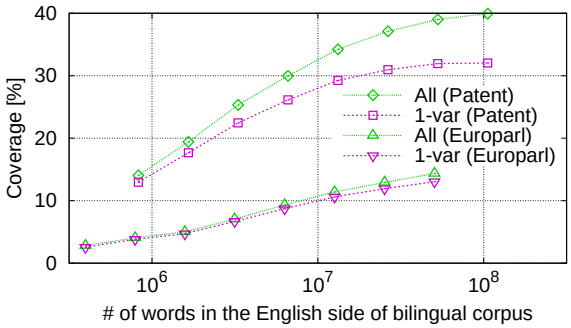
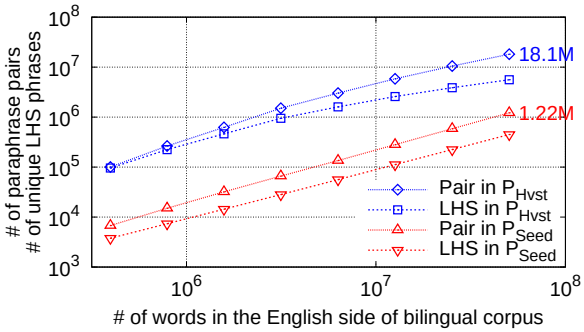


Figure 6: Coverage of the paraphrase patterns.



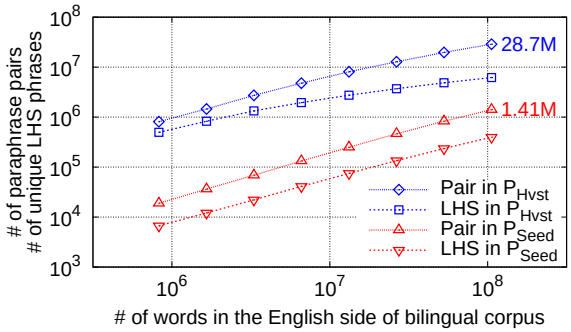


Figure 7: # of paraphrase pairs and unique LHS phrases in P_{Seed} and P_{Hvst} (left: Europarl, right: Patent).

Novel Paraphrases (P_{Hvst})

Using the paraphrase patterns, novel paraphrase pairs, P_{Hvst}, were harvested from the monolingual non-parallel corpora. In this experiment, we only retained one-variable patterns and regarded only single words as slot-fillers for them. Nevertheless, we managed to acquire a large number of paraphrase pairs as depicted in Figure 7, where pairs having zero similarity were excluded. For instance, when the full size of bilingual parallel corpus in Patent was used, we acquired 1.41M pairs of seed paraphrases, P_{Seed}, and 28.7M pairs of novel paraphrases, P_{Hvst}. In other words, our method expanded P_{Seed} by about 21 times. The number of unique LHS phrases that P_{Hvst} covers was also significantly larger than that of P_{Seed}.

Figure 8 highlights the remarkably large ratio of P_{Hvst} to P_{Seed} in terms of the number of paraphrase pairs and the number of unique LHS phrases. The smaller the bilingual corpus is, the higher the ratio is, except when there is only a very small amount of Europarl data. This demonstrates that our method is quite powerful, given a minimum amount of data.

Another striking difference between P_{Seed} and P_{Hvst} is the average number of RHS phrases per unique LHS phrase, i.e., their relative yield. As displayed in Figure 9, the yield for P_{Hvst} increased rapidly with the scaling up of the bilingual corpus, while that of P_{Seed} only grew slowly. The alignment-based method with bilingual corpora cannot produce very many RHS phrases per unique LHS phrase due to its reliance on conditional probability and the surface level processing. In contrast, our method does not limit the number of RHS phrases: each RHS phrase is separately assessed by its similarity to the corresponding LHS phrase. One limitation of our method is that it cannot achieve high yield for P_{Hvst} whenever only a small number of paraphrase patterns can be extracted from the bilingual corpus (see also Figure 5).

Both the ratio of P_{Hvst} to P_{Seed} and the relative yield could probably be increased by scaling up the monolingual corpus. For instance, in the patent domain, monolingual documents 10 times larger than the one used in the above experiments are available at the NTCIR project[7]. It would be interesting to compare the relative gains brought by in-domain versus general-purpose corpora.

[7] http://ntcir.nii.ac.jp/PatentMT-2/

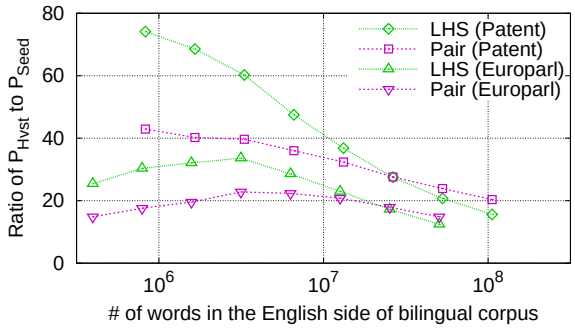


Figure 8: Ratio of P_{Hvst} to P_{Seed}.

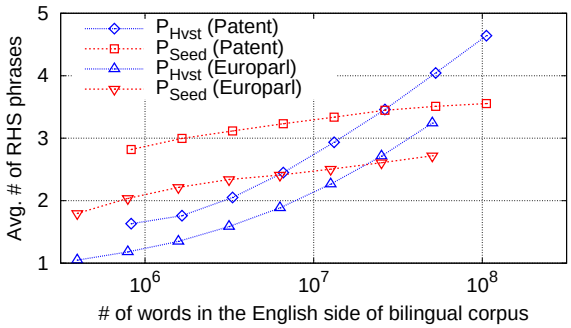
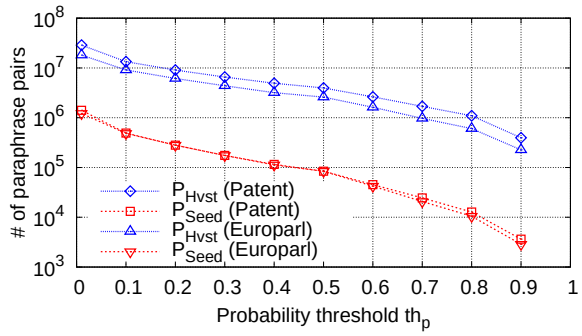


Figure 9: Average # of RHS phrases per LHS phrase.



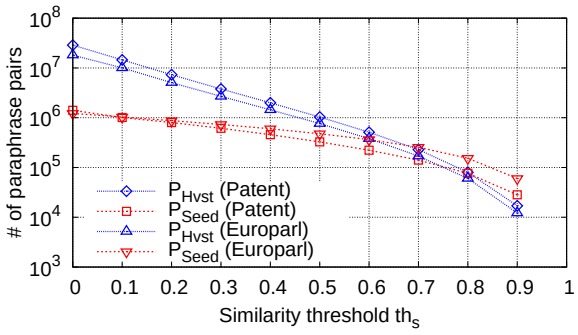


Figure 10: # of acquired paraphrase pairs against threshold values.
(left: probability-based ($0.01 \leq th_p \leq 0.9$, $th_s = \epsilon$), right: similarity-based ($\epsilon \leq th_s \leq 0.9$, $th_p = 0.01$))

Finally, we investigated how the number of paraphrase pairs varies depending on the values for the two thresholds, i.e., th_p on the conditional probability and th_s on the contextual similarity, respectively. Figure 10 shows the results when the full sizes of bilingual corpora are used. When the pairs were filtered only with th_p, the number of paraphrase pairs in P_{Hvst} decreased more slowly than that of P_{Seed} according to the increase of the threshold value. This is a benefit from our generalization and instantiation method. The same paraphrase pattern is often induced from more than one paraphrase pair in P_{Seed}. Thus, as long as at least one of them has a probability higher than the given threshold value, corresponding novel paraphrases can be harvested.

On the other hand, as a results of assessing each individual paraphrase pair by the contextual similarity, many pairs in P_{Hvst}, which are supposed to be incorrect instances of their corresponding pattern, are filtered out by a larger threshold value for th_s. In contrast, many pairs in P_{Seed} have a relatively high similarity, e.g., 40% of all pairs have similarity higher than 0.4. This indicates the quality of P_{Seed} is highly guaranteed by the shared translations.

5 Human Evaluation of Quality

To confirm that the quality of P_{Hvst} is sufficiently high, we carried out a substitution test.

First, by substituting sub-sentential paraphrases to existing sentences in a given test corpus, pairs of slightly different sentences were automatically generated. For instance, by applying "looks like" $\Rightarrow$ "resembles" to (5), (6) was generated.

(5) The roof *looks like* a prehistoric lizard's spine.

(6) The roof *resembles* a prehistoric lizard's spine.

Human evaluators were then asked to score each pair of an original sentence and a paraphrased sentence with the following two 5-point scale grades proposed by Callison-Burch (2008):

Grammaticality: whether the paraphrased sentence is grammatical (1: horrible, 5: perfect)

Meaning: whether the meaning of the original sentence is properly retained by the paraphrased sentence (1: totally different, 5: equivalent)

To make results more consistent and reduce the human labor, evaluators were asked to rate at the same time several paraphrases for the same source phrase. For instance, given a source sentence (5), the

evaluators might be given the following sentences in addition to a paraphrased sentence (6).

(7) The roof *seems like* a prehistoric lizard's spine.

(8) The roof *would look like* a prehistoric lizard's spine.

In this experiment, we showed five paraphrases per source phrase, assuming that evaluators would get confused if too large a number of paraphrase candidates were presented at the same time.

5.1 Data for Evaluation

As in previous work (Callison-Burch, 2008; Chan et al., 2011), we evaluated paraphrases acquired from the Europarl corpus on news sentences. Paraphrase examples were automatically generated from the English part of WMT 2008-2011 "newstest" data (10,050 unique sentences) by applying the union of P_{Seed} and P_{Hvst} of the Europarl setting (19.3M paraphrases for 5.95M phrases).

On the other hand, paraphrases acquired from patent documents are much more difficult to evaluate due to the following reasons. First, they may be too domain-specific to be of any use in general areas such as news sentences. However, conducting an in-domain evaluation would be difficult without enrolling domain experts. We expect that paraphrases from a domain can be used safely in that domain. Nevertheless, deciding under what circumstances they can be used safely in another domain is an interesting research question.

To reduce the human labor for the evaluation, sentences were restricted to those with moderate length: 10-30 words, which are expected to provide sufficient but succinct context. To propose multiple paraphrase candidates at the same time, we also restricted phrases to be paraphrased (LHS phrases) to those having at least five paraphrases including ones from P_{Hvst}. This resulted in 60,421 paraphrases for 988 phrase tokens (353 unique phrases).

Finally, we randomly sampled 80 unique phrase tokens and five unique paraphrases for each phrase token (400 examples in total), and asked six people having a high level of English proficiency to evaluate them. Inter-evaluator agreement was calculated from five different pairs of evaluators, each judging the same 10 examples. The remaining 350 examples were divided into six chunks of slightly unequal length, with each chunk being judged by one of the six evaluators.

		5-point		Binary		
	n	G	M	G	M	Both
P_{Seed}	55	4.60	4.35	0.85	0.93	0.78
P_{Hvst}	295	4.22	3.35	0.74	0.67	0.55
Total	350	4.28	3.50	0.76	0.71	0.58

Table 1: Avg. score and precision of binary classification.

5.2 Results

Table 1 shows the average of the original 5-point scale scores and the percentage of examples that are judged correct based on a binary judgment (Callison-Burch, 2008): an example is considered to be correct iff the grammaticality score is 4 or above and/or the meaning score is 3 or above. Paraphrases based on P_{Seed} achieved a quite high performance in both grammaticality ("G") and meaning ("M") in part because of the effectiveness of our filtering techniques. The performance of paraphrases drawn from P_{Hvst} was reasonably high and similar to the scores 0.68 for grammaticality, 0.61 for meaning, and 0.55 for both, of the best model reported in (Callison-Burch, 2008), although it was inferior to P_{Seed}.

Despite the fact that all of our evaluators had a high-level command of English, the agreement was not very high. This was true even when the collected scores were mapped into binary classes. In this case, the κ values (Cohen, 1960) for each criterion were 0.45 and 0.45, respectively, which indicate the agreement was "fair". To obtain a better κ value, the criteria for grading will need to be improved. However, we think that was not too low either[8].

The most promising way for improving the quality of P_{Hvst} is to ensure that paraphrase patterns cover only legitimate paraphrases. We investigated this by filtering the manually scored paraphrase examples with two thresholds for cleaning seed paraphrases P_{Seed}: th_p on the conditional probability estimated using the bilingual parallel corpus and th_s on the contextual similarity in the monolingual nonparallel corpus. Figure 11 shows the average score of the examples whose corresponding paraphrase is obtainable with the given threshold values. Note that the points in the figure with higher threshold values are less reliable than the others, because filtering reduces the number of the manually scored examples

[8]Note that Callison-Burch (2008) might possibly underestimate the chance agreement and overestimate the κ values, because the distribution of human scores would not be uniform.

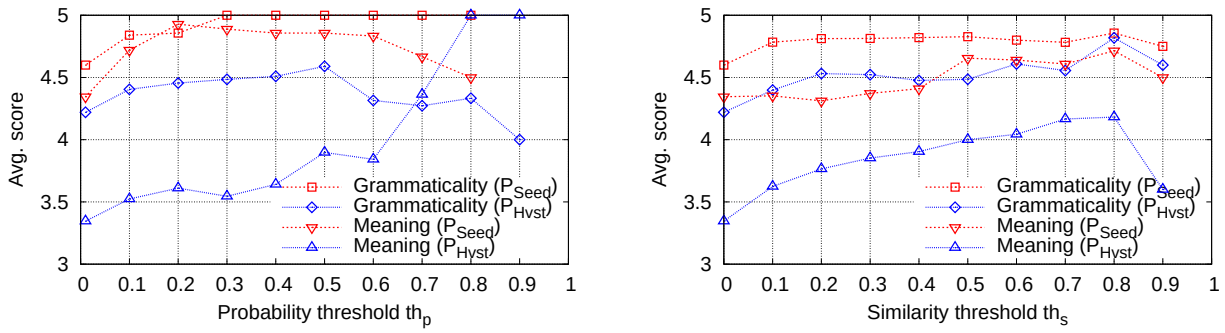


Figure 11: Average score of paraphrase examples against threshold values.
(left: probability-based ($0.01 \leq th_p \leq 0.9, th_s = \epsilon$), right: similarity-based ($\epsilon \leq th_s \leq 0.9, th_p = 0.01$))
The points with higher threshold values are less reliable than the others,
because filtering reduces the number of the manually scored examples used to calculate scores.

used to calculate scores. Nevertheless, it indicates that better filtering of P_{Seed} with higher threshold values is likely to produce a better-quality set of paraphrases P_{Hvst}. For instance, an inappropriate paraphrase pattern (9a) was excluded with $th_p = 0.1$ or $th_s = 0.1$, while correct ones (9b) and (9c) remained even when a large threshold value is used.

(9) a. "X years" $\Rightarrow$ "turn X"
b. "X supplied" $\Rightarrow$ "X provided"
c. "main X" $\Rightarrow$ "most significant X"

Kendall's correlation coefficient τ_B (Kendall, 1938) between the contextual similarity and each of the human scores were 0.24 for grammaticality and 0.21 for meaning, respectively. Although they are rivaling the best results reported in (Chan et al., 2011), i.e., 0.24 and 0.21, similarity metrics should be further investigated to realize a more accurate filtering.

6 Conclusion

In this paper, we exploited general patterns underlying paraphrases to acquire automatically a large number of high-quality paraphrase pairs using both bilingual parallel and monolingual non-parallel corpora. Experiments using two sets of corpora demonstrated that our method is able to leverage information in a relatively small bilingual parallel corpus to exploit large amounts of information in a relatively large monolingual non-parallel corpus. Human evaluation through a paraphrase substitution test revealed that the acquired paraphrases are generally of reasonable quality. Our original objective was to extract from monolingual corpora a large **quantity** of paraphrases whose **quality** is as high as that of paraphrases from bilingual parallel corpora. We have met the quantity part of the objective, and have come close to meeting the quality part.

There are three main directions for our future work. First, we intend to carry out in-depth analyses of the proposed method. For instance, while we showed that the performance of phrase substitution could be improved by removing noisy seed paraphrases, this also strongly affected the quantity. We will therefore investigate similarity metrics in our future work. Other interesting questions related to the work presented here are, as mentioned in Section 4.2, exploitation of patterns with more than one variable, learning curve experiments with different amounts of monolingual data, and comparison of in-domain and general-purpose monolingual corpora. Second, we have an interest in exploiting sophisticated paraphrase patterns; for instance, by inducing patterns hierarchically (recursively) and incorporating lexical resources such as those exemplified in (4). Finally, the developed paraphrase collection will be attested through applications, such as sentence compression (Cohn and Lapata, 2008; Ganitkevitch et al., 2011) and machine translation (Callison-Burch et al., 2006; Marton et al., 2009).

Acknowledgments

We are deeply grateful to our colleagues at National Research Council Canada, especially George Foster, Eric Joanis, and Samuel Larkin, for their technical support. The first author is currently a JSPS (the Japan Society for the Promotion of Science) Postdoctoral Fellow for Research Abroad.

References

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* (*ACL*), pages 597–604.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 50–57.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pages 16–23.

Rahul Bhagat and Deepak Ravichandran. 2008. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 161–170.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pages 17–24.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 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 Workshop on Geometrical Models of Natual Language Semantics* (*GEMS*), pages 33–42.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 190–200.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING*), pages 137–144.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the 6th Workshop on Statistical Machine Translation* (*WMT*), pages 85–91.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING*), pages 350–356.

Jinhua Du, Jie Jiang, and Andy Way. 2010. Facilitating translation using source language paraphrase lattices. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 420–429.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata. 2010. Overview of the patent translation task at the NTCIR-8 workshop. In *Proceedings of NTCIR-8 Workshop Meeting*, pages 371–376.

Atsushi Fujita, Shuhei Kato, Naoki Kato, and Satoshi Sato. 2007. A compositional approach toward dynamic phrasal thesaurus. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (*WTEP*), pages 151–158.

Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 1168–1179.

Zellig Harris. 1968. *Mathematical Structures of Language*. John Wiley & Sons.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi. 2011. Extracting paraphrases from definition sentences on the Web. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 1087–1097.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 341–348.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL*), pages 967–975.

Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Asso-*

ciation for Computational Linguistics (*HLT-NAACL*), pages 48–54.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

Stanley Kok and Chris Brockett. 2010. Hitting the right paraphrases in good time. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), pages 145–153.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 381–390.

Yuval Marton, Ahmed El Kholy, and Nizar Habash. 2011. Filtering antonymous, trend-contrasting, and polarity-dissimilar distributional paraphrases for improving statistical machine translation. In *Proceedings of the 6th Workshop on Statistical Machine Translation* (*WMT*), pages 237–249.

Aurélien Max. 2010. Example-based paraphrasing for improved phrase-based statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 656–666.

Marius Paşca and Péter Dienes. 2005. Aligning needles in a haystack: Paraphrase acquisition across the Web. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing* (*IJCNLP*), pages 119–130.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pages 102–109.

Fatiha Sadat, Howard Johnson, Akakpo Agbago, George Foster, Roland Kuhn, Joel Martin, and Aaron Tikuisis. 2005. PORTAGE: A phrase-based machine translation system. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 129–132.

Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of the 2002 Human Language Technology Conference* (*HLT*).

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary templates. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING*). 849-856.

Sander Wubben, Antal van den Bosch, Emiel Krahmer, and Erwin Marsi. 2009. Clustering and matching headlines for automatic paraphrase acquisition. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 122–125.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2009. Extracting paraphrase patterns from bilingual parallel corpora. *Natural Language Engineering*, 15(4):503–526.