Expanding paraphrase lexicons by exploiting generalities
Fujita, Atsushi; Isabelle, Pierre

Publisher's version / Version de l'éditeur:
https://doi.org/10.1145/3160488
ACM Transactions on Asian and Low-Resource Language Information Processing, 17, 2, pp. 1-36, 2018-02-05

NRC Publications Record / Notice d'Archives des publications de CNRC:
https://nrc-publications.canada.ca/eng/view/object/?id=b2cf9aa4-4513-4c40-b6e3-e64a2243598f
https://publications-cnrc.canada.ca/fra/voir/objet/?id=b2cf9aa4-4513-4c40-b6e3-e64a2243598f

Expanding Paraphrase Lexicons by Exploiting Generalities

ATSUSHI FUJITA, National Institute of Information and Communications Technology
PIERRE ISABELLE, National Research Council Canada

Techniques for generating and recognizing paraphrases, i.e., semantically equivalent expressions, play an important role in a wide range of natural language processing tasks. In the last decade, the task of automatic acquisition of subsentential paraphrases, i.e., words and phrases with (approximately) the same meaning, has been drawing much attention in the research community. The core problem is to obtain paraphrases of high quality in large quantity. This article presents a method for tackling this issue by systematically expanding an initial seed lexicon made up of high-quality paraphrases. This involves automatically capturing morpho-semantic and syntactic generalizations within the lexicon and using them to leverage the power of large-scale monolingual data. Given an input set of paraphrases, our method starts by inducing paraphrase patterns that constitute generalizations over corresponding pairs of lexical variants, such as "amending" and "amendment," in a fully empirical way. It then searches large-scale monolingual data for new paraphrases matching those patterns. The results of our experiments on English, French, and Japanese demonstrate that our method manages to expand seed lexicons by a large multiple. Human evaluation based on paraphrase substitution tests reveals that the automatically acquired paraphrases are also of high quality.

CCS Concepts: • **Computing methodologies → Information extraction**; *Language resources*;

Additional Key Words and Phrases: Paraphrase, semantic similarity, lexical variants, knowledge acquisition

ACM Reference format:
Atsushi Fujita and Pierre Isabelle. 2018. Expanding Paraphrase Lexicons by Exploiting Generalities. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 17, 2, Article 13 (January 2018), 36 pages.
https://doi.org/10.1145/3160488

1 INTRODUCTION

One of the characteristics of human languages is that the same semantic content can be expressed using multiple different linguistic expressions, i.e., **paraphrases**. The notion of paraphrase covers a variety of linguistic phenomena, such as those in the sentences of (1).

(1) a. There would be better approaches than amending the regulation.
b. Amending the regulation would not be the most appropriate approach.
c. Amending the regulation would not be the best approach.
d. Amendment of the regulation would not be the best approach.

Authors' addresses: A. Fujita, National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan; email: atsushi.fujita@nict.go.jp; P. Isabelle, National Research Council Canada, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada; email: pierre.isabelle@cnrc-nrc.gc.ca.

ACM 2375-4699/2018/01-ART13 $15.00
https://doi.org/10.1145/3160488

Sentence (1a) shares some content words with the other sentences, but the sentence structure is completely different. We can say that it constitutes a sentential paraphrase of the others. On the other hand, sentences (1b) and (1c) exhibit an alternation between the phrase "most appropriate" and the single word "best." Sentences (1c) and (1d) exhibit structural variation in a nominal phrase, i.e., "amending the regulation" versus "amendment of the regulation." In this article, we refer to these substituted words and phrases that have approximately the same meaning as mutual **subsentential paraphrases**. Dealing with paraphrases is an important issue relevant to a broad range of natural language processing (NLP) tasks [1, 41]. For instance, technologies that accurately recognize and/or generate paraphrases promise to improve NLP applications as diverse as information retrieval, machine translation, question answering, text summarization, and text simplification.

A large-scale knowledge base of subsentential paraphrases appears indispensable for dealing robustly and accurately with paraphrase phenomena.[1] Thus, in recent years, the task of automatically creating such **paraphrase lexicons** has been drawing the attention of many researchers (see Section 2). The challenge is to ensure broad coverage alongside high accuracy. Given the sheer size of available monolingual corpora,[2] there is no doubt that they contain more paraphrases than any other resources. However, it is proving extremely difficult to automatically extract a sizable fraction of them without bringing into play some additional resources. For instance, methods based solely on co-occurrence statistics in monolingual corpora have trouble distinguishing paraphrases from other types of semantic relations, such as antonymy and sibling words. In contrast, it is much easier to extract highly accurate paraphrases from parallel corpora, taking advantage of alignment constraints between subparts of aligned documents and sentences. However, the availability of such corpora is far more limited, with the result that their paraphrase coverage is fairly limited.

In this article, we propose a method for expanding a preexisting low-coverage but high-quality set of paraphrases. By exploiting the morpho-semantic and syntactic generality underlying paraphrases as a means of leveraging the coverage of large-scale monolingual data, we aim to significantly extend the coverage without sacrificing quality. Given seed paraphrase pairs, our method first induces **paraphrase patterns** by abstracting away from corresponding word stems that appear on each side of a paraphrase pair. For instance, from seed pair (2a), pattern (2b) is induced:

(2) a. amending the regulation ⇔ amendment of the regulation
 b. X_1:ing the X_2:ϵ ⇔X_1:ment of the X_2:ϵ

where each variable represents a word or stem that, together with specific (possibly null) affixes, captures a lexical correspondence. For instance, the pair ("X_1:ing", "X_1:ment") captures not only the correspondence between "amending" and "amendment" but also between a large number of other verb/noun pairs, such as "developing" and "development." As for the pair ("X_2:ϵ", "X_2:ϵ"), it just abstracts over any identical word pair. We call them **affix patterns** (see Section 2.3).

The resulting paraphrase patterns are then used to acquire from a monolingual corpus new paraphrase pairs that constitute instances of the paraphrase patterns learned from the seed paraphrases. This makes it possible to acquire new paraphrase pairs that have no lexical overlap with the seed paraphrases. For instance, paraphrase pairs in (3) would be obtained by using the pattern (2b).

[1] Distributed representations of words and phrases constitute alternative means of dealing with paraphrases, but their superiority to traditional symbolic approaches such as ours has yet to be demonstrated.

[2] While monolingual parallel corpora have also been used as a source of paraphrases, the term "monolingual corpora" in this article refers to monolingual nonparallel corpora unless otherwise explicitly noted.

Table 1. Comparison of Prior Arts in Automatic Paraphrase Acquisition

	Corpus	Approach	Coverage	Accuracy
(a)	Monolingual	Contextual similarity	√ Most promising	× Low in general
(b)	Monolingual parallel	Alignment	× Extremely limited	√ Relatively high
(c)	Monolingual comparable	Alignment	× Relatively limited	√ Relatively high
(d)	Bilingual parallel	Pivoting	× Relatively limited	√ Relatively high

(3) a. investing the resources ⇔ investment of the resources
b. recruiting the engineers ⇔ recruitment of the engineers

Some generalities underlying paraphrases have been exploited manually [19, 29, 32, 46] or empirically [21], but there is no general method for identifying and exploiting a wide variety of lexical correspondences and paraphrase patterns. Our method tackles this issue through a fully empirical exploitation of morphologically based affix patterns discovered in a high-quality seed set of paraphrases. The method is thus potentially applicable to any of the numerous languages that possess a relatively rich morphology.

The remainder of this article is organized as follows. Section 2 reviews existing methods for creating paraphrase lexicons and those for dealing with various lexical correspondences. Section 3 provides some tips for further improving the quality of the low-coverage seed set of paraphrases produced by existing methods, and then presents our method for systematically expanding such seed sets. Section 4 describes our experiments on expanding seed paraphrase lexicons in English, French, and Japanese, focusing primarily on the quantitative impact with regard to the quantity of seed paraphrases. Section 5 reports on our manual assessment of the quality of the created paraphrase lexicons, and Section 6 summarizes the contributions of this work and points out directions for future work.

2 LITERATURE REVIEW

Methods for automatically creating paraphrase lexicons using various types of corpora have been extensively studied. As summarized in Table 1, there are two major streams: one that uses monolingual corpora, i.e., (a), and one that uses parallel or comparable corpora, i.e., (b), (c), and (d). After reviewing each approach, we summarize previous work that has focused on the generalities underlying paraphrases.

2.1 Automatic Paraphrase Acquisition from Monolingual Corpora

Monolingual corpora constitute the richest resource when targeting high coverage, given their availability on a massive scale, e.g., on the Web.

Techniques for acquiring paraphrases from such corpora are mostly based on the **contextual similarity** stemming from the Distributional Hypothesis [28]. There are various recipes for computing the contextual similarity of two given expressions, but all of them comprise three ingredients: (i) extraction of contextual features for each expression, (ii) weighting and filtering such features, and (iii) similarity measurement based on the two sets of contextual features. The first step is to represent each given expression with a set of co-occurring expressions in the relevant corpus. For instance, adjacent word n-grams [7, 43, 47], nominal arguments of verb phrases [13, 40, 52, 54, 55], modifiers and modified words [26, 56], and even indirect dependencies [27] have been used. Then, the weight for each feature is adjusted. Pointwise mutual information [39] and relative feature focus [24] are among the better-known examples of weighting methods. Latent variable models and distributed representations are useful in alleviating the data-sparseness problem occurring when processing surface forms of words. Finally, contextual similarity of two expressions is computed by comparing their corresponding feature sets. Several measures, including cosine similarity, Jaccard's coefficient, and Kullback–Leibler divergence, have been used [31, 38]. It is worth noting that most of the measures focus on the overlap of the two feature sets; thus, the similarity will be zero if two sets have no overlap.

Despite the quantitative advantage, this approach tends to result in low accuracy, because contextual information alone often fails to differentiate paraphrases from expressions that entertain other semantic relations, such as antonyms, hypernym–hyponym pairs and sibling words [44].

Instead of exhaustively collecting all candidate pairs of expressions, several methods for relation extraction collect linguistic patterns that are semantically similar to a given set of seed patterns [13, 49, 50, 52, 54]. First, slot-fillers of the patterns and new patterns are iteratively collected in a bootstrapping manner. Then, pairs of the patterns are regarded as paraphrase patterns. Each of the acquired patterns holds a particular semantic relation, which is virtually defined by the given seed patterns. However, they do not capture any generality between pairs of paraphrases, including lexical correspondences exhibited by, for instance, (2b).

2.2 Automatic Paraphrase Acquisition from Parallel and Comparable Corpora

Much effort has gone into compiling monolingual parallel corpora and extracting paraphrases from them by identifying corresponding parts of aligned sentences. Barzilay and McKeown [6] and Pang et al. [48] have collected multiple human translations of the same source text. Multiple verbalizations of mathematical proofs have also been used [5]. These methods rely on solid anchors that guarantee the semantic equivalence of sentences or text fragments.

Monolingual comparable corpora are also useful sources of paraphrases. For instance, articles from different newswire services describing the same event can be used for that purpose [4, 15, 51, 57]. Chen and Dolan [10] created such a corpus by collecting multiple descriptions for the same short movies through crowdsourcing. Web-harvested definition sentences of the same term are not necessarily parallel at sentence level but often contain subsentential paraphrases [30, 58].

Bilingual parallel corpora have been recognized as sources of paraphrases since the work by Bannard and Callison-Burch [3]. First, a translation table is created using techniques developed for statistical machine translation (SMT). Then, pairs of expressions in one language are extracted as paraphrases if they share identical translations in the other language, i.e., the **pivot language**. For instance, a pair ("under control", "in check") will be extracted on the basis of its shared linkage with the German translation "unter controlle." Each paraphrase pair (e_1, e_2) is assigned forward and backward **paraphrase probabilities**, $p(e_2|e_1)$ and $p(e_1|e_2)$, estimated by marginalizing over all of the translations F shared by e_1 and e_2, as follows:

$$p(e_2|e_1) = \sum_{f \in F} \phi(e_2|f)\phi(f|e_1), \tag{1}$$

where $\phi(e|f)$ and $\phi(f|e)$ are the backward and forward translation probabilities between e in the language of interest and f in the pivot language, respectively. These are estimated from the number of times e and f are aligned and the number of occurrences of each expression in the corresponding side of bilingual data. This **bilingual pivoting** approach has inspired further techniques, such as the use of syntactic information as the basis of constraints [8, 59], learning patterns using synchronous grammar [21], uncovering missing links by combining multiple translation tables and other lexical resources [36], and reranking candidate pairs on the basis of contextual similarity [9]. Using this approach, Ganitkevitch and Callison-Burch [20] compiled a set of paraphrase lexicons for various languages called the Paraphrase Database (PPDB).

While parallel and comparable corpora constitute useful sources of highly accurate paraphrases, their limited availability for most language pairs precludes the extraction of high-coverage paraphrase lexicons solely on their basis. It should also be mentioned that the accuracy of the relevant methods is not perfect. Some spurious paraphrase pairs will also be extracted, notably as a result of errors in the automatic word alignment involved in the extraction process.

2.3 Generality Exhibited by Paraphrases

As reviewed above, the existing computational methods extract individual pairs of paraphrases without trying to generalize in any way over the extracted set. Yet, as suggested by our examples (2) and (3), many types of paraphrases can be captured by generalizing over specific surface forms. In this article, we consider the following three types of morpho-semantically related word groups and refer to them as **lexical variants**.

Derivational morphology. Different words that share the same stem and a large part of their meaning, such as {"develop", "developer", "development", . . .}. Words in this group may differ in part of speech (POS).

Inflectional morphology. Different surface forms of the same word, such as {"amend", "amends", "amending",. . .}. Such sets of forms result from language-dependent linguistic processes, such as verb conjugation, noun pluralization, and case marking.

Orthographic variants. Different spellings of the same inflectional and conjugational form of the same word, such as {"color", "colour"} and {"authorize", "authorise"}.

Several traditional linguistic theories, such as transformational grammar [29] and Meaning-Text Theory [46], have proposed representing each set of paraphrases by a unique canonical forms of lexical variants. However, they did not provide any automated means of computing paraphrases. Jacquemin [32] and Fujita et al. [19] have attempted to capture various kinds of paraphrases using manually described syntactic transformation patterns in combination with dictionaries of lexical derivations. Such attempts capture only limited types of subsentential paraphrases, such as technical terms and short verb phrases, and their real coverage has not been evaluated.

CELEX [2] provides useful tools for computing but is available for only a couple of languages. WordNet [17] also contains information of that kind and is currently available for a somewhat larger number of languages. Catvar [25] is a more comprehensive lexical derivation database but only available for English. While such manually created resources tend to be highly accurate, their creation requires a great deal of human effort. Gaussier [23] and Fujita et al. [19] have automatically extracted lexical derivation sets from a list of headwords using **affix patterns**, such as ("X:ment", "X:er") for ("development", "developer"), as clues. While such approaches significantly reduce human effort and retain reasonable accuracy, their coverage is still limited as a result of their reliance on manually compiled and POS-tagged word lists.

3 EXPANDING PARAPHRASE LEXICONS THROUGH GENERALIZATION

This section describes our corpus-based paraphrase acquisition method. As mentioned above, the idea is to exploit generalizations underlying high-quality seed sets of paraphrases in order to achieve broader coverage without sacrificing high accuracy using monolingual data. The process comprises the following three steps (see also Figure 1).

Step 1. Acquiring seed paraphrase pairs. High-quality seed paraphrase pairs, S_{Seed}, are acquired using existing methods, such as those reviewed in Section 2.

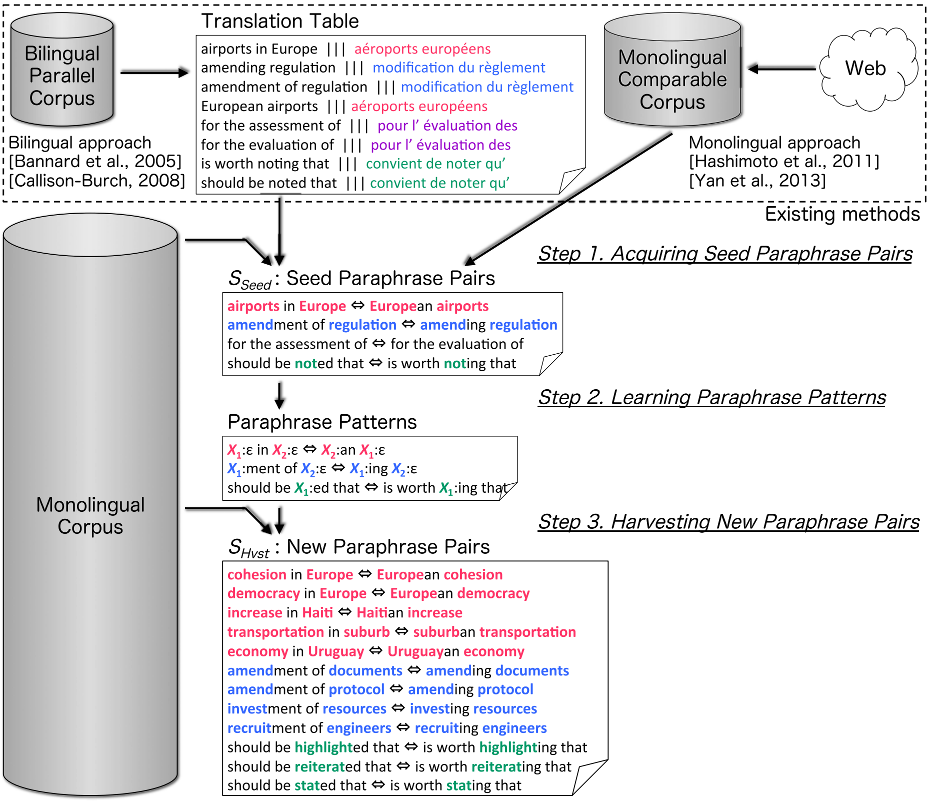
Translation Table
Bilingual Parallel Corpus
airports in Europe ||| aéroports européens
amending regulation ||| modification du règlement
amendment of regulation ||| modification du règlement
European airports ||| aéroports européens
for the assessment of ||| pour l' évaluation des
for the evaluation of ||| pour l' évaluation des
is worth noting that ||| convient de noter qu'
should be noted that ||| convient de noter qu'
Monolingual Comparable Corpus
Web
Bilingual approach
[Bannard et al., 2005]
[Callison-Burch, 2008]
Monolingual approach
[Hashimoto et al., 2011]
[Yan et al., 2013]
Existing methods
Step 1. Acquiring Seed Paraphrase Pairs
S_{Seed} : Seed Paraphrase Pairs
airports in Europe ⇔ European airports
amendment of regulation ⇔ amending regulation
for the assessment of ⇔ for the evaluation of
should be noted that ⇔ is worth noting that
Step 2. Learning Paraphrase Patterns
Paraphrase Patterns
X_1:ε in X_2:ε ⇔ X_2:an X_1:ε
X_1:ment of X_2:ε ⇔ X_1:ing X_2:ε
should be X_1:ed that ⇔ is worth X_1:ing that
Monolingual Corpus
Step 3. Harvesting New Paraphrase Pairs
S_{Hvst} : New Paraphrase Pairs
cohesion in Europe ⇔ European cohesion
democracy in Europe ⇔ European democracy
increase in Haiti ⇔ Haitian increase
transportation in suburb ⇔ suburban transportation
economy in Uruguay ⇔ Uruguayan economy
amendment of documents ⇔ amending documents
amendment of protocol ⇔ amending protocol
investment of resources ⇔ investing resources
recruitment of engineers ⇔ recruiting engineers
should be highlighted that ⇔ is worth highlighting that
should be reiterated that ⇔ is worth reiterating that
should be stated that ⇔ is worth stating that

Fig. 1. Overview of our method for expanding a given seed paraphrase lexicon.

Step 2. Learning paraphrase patterns. Paraphrase patterns are induced from S_{Seed}. This involves abstracting away from some specific stems and words, to uncover various types of lexical variants in each paraphrase pair.

Step 3. Harvesting new paraphrase pairs. New paraphrase pairs, S_{Hvst}, are extracted from monolingual data by exploiting the induced paraphrase patterns.

The generalization-and-instantiation approach of steps 2 and 3 is intended to build upon the existing methods used in step 1.

Previous work has established the usefulness of linguistic annotations, such as POS tags and syntactic information, for paraphrase extraction. On the other hand, we chose to avoid reliance on such tools to make our method applicable regardless of their availability. All we need are a tokenizer and a list of stop words. In fact, even the latter can be dispensed with, as it can easily be replaced by a list of the most frequent words.

3.1 Step 1. Acquiring Seed Paraphrase Pairs

The goal of the first step is to acquire a seed set of paraphrase pairs, S_{Seed}. Any method can be used provided that it yields enough high-quality paraphrase pairs distributed over a wide variety of types of lexical variants and paraphrase patterns. Given their greater accuracy, alignment-based methods applied to bilingual or monolingual parallel corpora (Section 2.2) are preferable to similarity-based methods applied to monolingual corpora (Section 2.1). For the sake

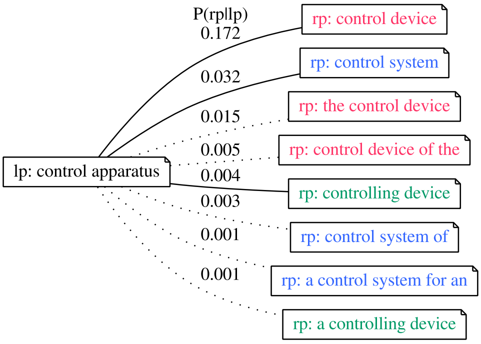

Fig. 2. RHS filtering for "control apparatus."

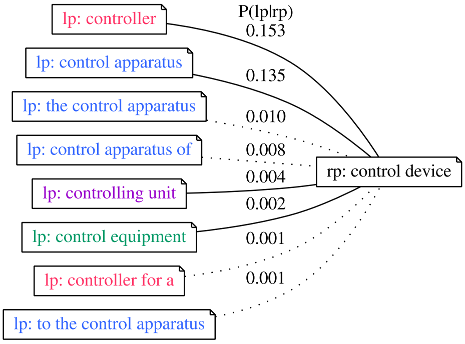

Fig. 3. LHS filtering for "control device."

of reproducibility, we use here the bilingual pivoting method [3]. More specifically, we use the phrase-based SMT framework [35] and offer additional filtering methods as a way to further improve the resulting set of extracted paraphrase pairs.

3.1.1 *Cleaning Translation Pairs.*

The phrase pair extraction process of phrase-based SMT systems aims at high recall for increased robustness of the translation process. As a result, unfiltered results of the bilingual pivoting method will include many pairs of phrases that are not accurate paraphrases. For instance, as discussed in Koehn [34], applying the phrase-extraction algorithm to a sentence pair that contains unaligned words often leads to multiple noninterchangeable translations for a single source phrase, such as "dass" and ", dass" in German for "that" in English.

In order to reduce this kind of noise, we apply some filtering techniques to the translation pairs. First, statistically unreliable translation pairs [33] are filtered out. Then, we further filter out phrases made up entirely of stop words or punctuation marks, both in the language of interest and in the pivot language.

3.1.2 *Filtering Seed Paraphrase Pairs.*

Let S_{Raw} be the initial set of paraphrase pairs extracted from the sanitized translation table using the bilingual pivoting method. We extract seed paraphrase pairs, S_{Seed}, from S_{Raw}, relying on a stoplist and paraphrase probabilities provided by Equation (1) in Section 2.2. We start our filtering process by discarding pairs whose difference comprises stop words only, such as ("the schools", "schools and"). Furthermore, to filter out unlikely pairs of the kind shown with dotted lines in Figures 2 and 3, we compare the right-hand side (RHS) phrases of each left-hand side (LHS) phrase and vice versa. Given a set of paraphrase pairs, the RHS-filtering (Figure 2) filters RHS phrases corresponding to the same LHS phrase lp. An RHS phrase rp is not licensed *iff* lp has another RHS phrase rp' ($\neq rp$) that satisfies the following conditions.

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

The LHS-filtering (Figure 3) works in the same manner; LHS phrases for each RHS phrase rp are compared. An LHS phrase lp is not qualified as a legitimate source of rp *iff* rp has another LHS phrase lp' ($\neq lp$) that satisfies the following conditions.

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

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

3.1.3 Measuring Reliability. Candidate pairs are finally filtered on the basis of their reliability scores. A threshold (th_p) on paraphrase probability has been used in relevant literature [14, 16, 45, and others]. Furthermore, we measure the contextual similarity between the phrases of each paraphrase pair, $Sim(lp, rp)$, using monolingual data, as in previous studies [9, 22], which have established that information derived from additional monolingual data can help assess the quality of paraphrases extracted from bilingual data. We discard paraphrase pairs whose similarity is lower than a specific threshold (th_{s1}). A variety of recipes for computing contextual similarity (see Section 2.1) can be used; we have selected one for our experiment in Section 4. However, we do not provide here any comparisons or recommendations regarding that particular aspect.

3.2 Step 2. Learning Paraphrase Patterns

Given a seed set of paraphrase pairs, S_{Seed}, paraphrase patterns are then induced. For instance, from the paraphrase pairs in (4), we obtain the paraphrase patterns in (5).

(4) a. airports in Europe ⇔ European airports
b. amendment of regulation ⇔ amending regulation
c. should be noted that ⇔ is worth noting that

(5) a. X_1:ϵ in X_2:ϵ ⇔ X_2:an X_1:ϵ
b. X_1:ment of X_2:ϵ ⇔ X_1:ing X_2:ϵ
c. should be X_1:ed that ⇔ is worth X_1:ing that

where each pair of the same variable slots accompanied by affix information, such as ("X_1:ing", "X_1:ment") and ("X_2:ϵ", "X_2:ϵ"), represents a corresponding pair of words that are captured by affix patterns introduced in previous work (see Section 2.3).

This approach aims to automatically capture general paraphrase patterns of the kind that have often been targeted by means of handcrafted rules [19, 32]. The problem with handcrafted rules is the difficulty of covering the extensive variety of such paraphrase patterns robustly and accurately. While our data-driven method does not need any manually created resources, such as dictionaries, it is still able to take advantage of whenever they are available.

Note that our use of variable slots is crucially different from their use in other paraphrase acquisition methods. In [13, 40, 52, 54, 55], variable slots serve to calculate the contextual similarity of the original fully lexicalized parts; in [8, 59], they serve to restrict the context in which the paired phrases are regarded as legitimate paraphrases.

3.2.1 Learning Affix Patterns. First, affix patterns of lexical variants, such as ("X:ϵ", "X:an") and ("X:ment", "X:ing") in (5), are learned from S_{Seed}. While previous studies [19, 23] have considered only suffix patterns, we also deal with identical word forms shared by paraphrase pairs, denoted as ("X:ϵ", "X:ϵ"), and prefix patterns. For instance, we obtain ("un:X", "ϵ:X") and ("co:X", "ϵ:X") from ("unreliable", "reliable") and ("coexist", "exist") in (6).

(6) a. is unreliable ⇔ is not reliable
b. coexist with ⇔ exist together with

We currently do not consider combinations of prefix and suffix, such as ("ϵ:X:ly", "in:X:ϵ") and ("ϵ:X:ed", "un:X:able") exhibited by ("directly", "indirect") and ("believed", "unbelievable"), respectively, and types of affixes other than prefixes and suffixes.

ALGORITHM 1: Extraction of Candidate Lexical Variants

```
Input: A paraphrase pair comprising $N$ and $M$ words $a = (a_1, a_2, \ldots, a_N)$ and $b = (b_1, b_2, \ldots, b_M)$
Input: A set of stop words $W$
Output: A set of candidate lexical variants with corresponding affix patterns $C$
  procedure EXTRACT_CANDIDATE_LEXICAL_VARIANTS($a, b, W$)
      $C \leftarrow \{\}$
      for each $(a_i, b_j)$ such that $1 \leq i \leq N \wedge 1 \leq j \leq M$ do
          if $a_i \notin W \wedge b_j \notin W$ then                              ▷ Stop words are ignored
              $prefix \leftarrow$ FIND_LONGEST_COMMON_PREFIX($a_i, b_j$)
              if $prefix \neq$ “” then              ▷ If $a_i$ and $b_j$ share prefix of at least one character
                  $suffix_1 \leftarrow$ VARIABLIZE_PREFIX($a_i, prefix$)
                  $suffix_2 \leftarrow$ VARIABLIZE_PREFIX($b_j, prefix$)
                  $C \leftarrow C \cup \{(a_i, b_j, suffix_1, suffix_2, prefix)\}$   ▷ e.g., (“noted,” “noting,” “$X$:ed,” “$X$:ing,” “not”)
              end if
              $suffix \leftarrow$ FIND_LONGEST_COMMON_SUFFIX($a_i, b_j$)
              if $suffix \neq$ “” then              ▷ If $a_i$ and $b_j$ share suffix of at least one character
                  $prefix_1 \leftarrow$ VARIABLIZE_SUFFIX($a_i, suffix$)
                  $prefix_2 \leftarrow$ VARIABLIZE_SUFFIX($b_j, suffix$)
                  $C \leftarrow C \cup \{(a_i, b_j, prefix_1, prefix_2, suffix)\}$     ▷ e.g., (“coexist,” “exist,” “co:$X$,” “$\epsilon$:$X$,” “exist”)
              end if
          end if
      end for
  end procedure
```

In previous work, affix patterns have been acquired from lists of headwords in manually compiled dictionaries. Here, we acquire them from actual paraphrase pairs by over-generation and filtering. First, candidate pairs of lexical variants are extracted from S_{Seed} using Algorithm 1, on the following assumption.

> Words appearing on opposite sides of a paraphrase pair are very likely to be semantically related whenever they share the same stem.

We do not rely on any language resources to determine word stems. Instead, given a word pair, we regard their longest common prefix or suffix as their shared stem and generate a candidate affix pattern using the remaining parts of the words. For instance, from paraphrase pair (7), we extract four pairs of words and their corresponding affix patterns, as shown in Table 2.

(7) is aimed at achieving ⇔ aims to achieve

A list of stop words is used to prevent associating unlikely pairs of words, such as (“alleviate”, “all”) and (“compare”, “are”). This also excludes pairs of related words, such as (“that”, “this”) and (“neither”, “either”). However, we believe that this does not significantly reduce the coverage of affix patterns. Whenever a pattern is indeed useful, it will be possible to derive it from other word pairs.

When processing Japanese data, we Romanize the cursive forms of syllabographs, i.e., *hiragana*, as in the linguistics literature to extract more appropriate candidates of affix patterns at the phoneme level. For instance, a naïve application of the above method to a pair of transitive and intransitive verbs in (8a) outputs a candidate affix pattern (8b). However, through Romanization, the method can include “k” in the shared stem and obtains a more general affix pattern (8c).

Table 2. Candidate Pairs of Lexical Variants and Corresponding Affix Patterns Extracted from Example (7)

Word$_1$	Word$_2$	Affix$_1$	Affix$_2$	Stem
aimed	aims	X:ed	X:s	aim
aimed	achieve	X:imed	X:chieve	a
achieving	aims	X:chieving	X:ims	a
achieving	achieve	X:ing	X:e	achiev

Table 3. Examples of Filtering Affix Patterns

Affix$_1$	Affix$_2$	# of unique stems		Result
		length≥5	length<5	
X:an	X:ϵ	7	0	Retained
X:ment	X:ing	18	2	Retained
X:ing	X:ed	245	53	Retained
un:X	ϵ:X	43	7	Retained
co:X	ϵ:X	8	0	Retained
X:chieve	X:imed	0	1	Eliminated
X:chieving	X:ims	0	1	Eliminated
X:ed	X:s	69	22	Retained
X:ing	X:e	330	70	Retained

Note: The numbers of unique stems are taken from our experimental results using the entire data for the Europarl English setting (Section 4).

(8) a. (“近づける” (*tikadukeru*, to put close), “近づく” (*tikaduku*; to get close))
b. Stem: “近づ” (*tikadu*), affix pattern: (“X:ける” (*keru*), “X:く” (*ku*))
c. Stem: “近 duk” (*tikaduk*), affix pattern: (“X:eru”, “X:u”)

As it turns out, not every candidate affix pattern generated by the above method proves to be appropriate. This is why we perform filtering based on the following criterion [23].

> An affix pattern is retained *iff* it is associated with at least n unique stems that are of length at least k characters.

This criterion relies on two parameters. Parameter n assesses whether a pattern is sufficiently productive. The more word pairs a pattern is associated with, the more likely the pattern is to be useful in finding new word pairs that match it. The other parameter, k, is motivated by the observation that a genuine pattern is more likely to be used even for long stems because affixation, inflection, and conjugation are fundamental operations for producing lexical variants. The optimal value for these two parameters can vary across different corpora and languages. Table 3 shows whether each affix pattern in the above examples is retained or eliminated with $k = 5$ and $n = 2$, as proposed in [23].

3.2.2 Generating Paraphrase Patterns. Paraphrase patterns are generated from the seed paraphrase pairs in S_{Seed} by using the affix patterns acquired in the previous step. In this step, as presented in Algorithm 2, we exhaustively consider all combinations of lexical variants that match

ALGORITHM 2: Generation of Paraphrase Patterns

```
Input: A pair of phrases $a$ and $b$
Input: A set of stop words $W$
Input: A set of affix patterns $L$          ▷ For all $(affix_1, affix_2)$, $(affix_2, affix_1)$ is also included in $L$
Output: A set of paraphrase patterns $P$
  procedure GENERATE_PARAPHRASE_PATTERNS($a, b, W, L$)
      $C \leftarrow$ EXTRACT_CANDIDATE_LEXICAL_VARIANTS($a, b, W$)          ▷ See Algorithm 1
      $C' \leftarrow \{p = (\cdot, \cdot, affix_1, affix_2, \cdot) \in C \mid (affix_1, affix_2) \in L\}$          ▷ Retain only reliable affix patterns
      $P \leftarrow$ ENUMERATE_PATTERNS($a, b, C', 1$)          ▷ Find all generalizations
      $P \leftarrow$ UNIQUE($P$)          ▷ Take unique generalizations ignoring index variations of variable slots
  end procedure
  procedure ENUMERATE_PATTERNS($a, b, C, n$)
      $P = \{\}$
      for each $p = (word_1, word_2, affix_1, affix_2, stem) \in C$ do
          $a' \leftarrow$ GENERALIZE_WORDS($a, word_1, affix_1, n$)          ▷ Replace all words $word_1$ in $a$ with $X_n$ and $affix_1$
          $b' \leftarrow$ GENERALIZE_WORDS($b, word_2, affix_2, n$)          ▷ Replace all words $word_2$ in $b$ with $X_n$ and $affix_2$
          if $a' \neq a \land b' \neq b$ then          ▷ $(a, b)$ is generalized by $p$
              $P' \leftarrow$ ENUMERATE_PATTERNS($a', b', C \backslash \{p\}, n + 1$)          ▷ Depth-first search with recursion
              if $P' = \{\}$ then          ▷ If $(a, b)$ is no more generalizable
                  $P' \leftarrow \{(a', b')\}$
              end if
              $P \leftarrow P \cup P'$
          end if
      end for
      return $P$          ▷ Return {} if none of $C$ is applicable
  end procedure
```

one of these generated affix patterns. For instance, from the paraphrase pair (7), the following two patterns can be generated.[3]

(9) a. is X_1:ed at X_2:ing ⇔ X_1:s to X_2:e
b. is X_1:imed at X_2:chieving ⇔X_2:ims to X_1:chieve

3.3 Step 3. Harvesting New Paraphrase Pairs

The paraphrase patterns induced in the previous step have better coverage than S_{Seed} as a consequence of generalizing corresponding pairs of words. However, at the same time, some information, such as selectional restrictions of verbs and other types of idiosyncracy of words, is lost in the generalization process. Thus, even if the given seed paraphrase pairs are all correct, the induced paraphrase patterns do not necessarily guarantee that corresponding pairs of phrases are always correct paraphrases. Instead of leaving such issues of pattern matching to the potential users, we collect new pairs of phrases from a large-scale monolingual data and assess each of them on the basis of yet another type of information, i.e., contextual similarity.

First, all those pairs of phrases that instantiate each side of any given pattern are collected. A particular affix pattern alone cannot guarantee that matching pairs will always preserve the semantic relations exhibited by the seed paraphrase pair from which it has been induced. For

[3] Note that this example is used only for explaining the complete pattern generation process. Owing to the aforementioned filtering method, spurious patterns, such as (9b), get filtered out in practice.

instance, pattern (10b) is learned from seed paraphrase pair (10a), where the "part-of" relation is held between the corresponding variable slots ("X_1:ϵ", "X_1:an").

(10) a. countries of Europe ⇔ European nations
b. countries of X_1:ϵ ⇔ X_1:an nations

The variable slots can be instantiated by word pairs with different semantic relations. For instance, the variable slot in (10b) will match inappropriate pairs, such as ("uncle", "unclean") and ("beg", "began"), alongside appropriate ones, such as ("Haiti", "Haitian") and ("suburb", "suburban"). However, we assume that the fixed part of each paraphrase pair, such as "countries of" and "nations" in (10b), provides a strong enough constraint to filter out inappropriate matches, guaranteeing also the grammaticality of the entire phrase to some degree.

Pattern matching alone would collect pairs of phrases that are not suitable paraphrases in any context. We therefore assess the legitimacy of each collected pair of phrases by calculating contextual similarity between phrases in the same way as we have done for filtering S_{Seed}. Then, a pair is eliminated whenever its two phrases are used in significantly dissimilar contexts, i.e., $Sim(lp, rp) < th_{s2}$. The contextual similarity of antonyms and sibling words also tends to be high. However, we expect that this is not a problem within our process, assuming that semantic equivalence between each collected pair of phrases is virtually assured because of the fact that the corresponding pattern has been learned from high-quality seed paraphrase pairs.

As illustrated in Figure 1, the paraphrase pairs we can collect from monolingual data include ones that have no word overlap with the seed paraphrase pairs. For instance, using the paraphrase pattern (5b) induced from a seed paraphrase pair (4b), we can obtain the following pairs.

(11) a. investment of resources ⇔ investing resources
b. recruitment of engineers ⇔ recruiting engineers

3.4 Limitation

One limitation of our proposed method is that it only considers prefixation and suffixation as clues for discovering lexical variants. Extensions are needed in order to capture a wider range of lexical variants. In some languages, infixation and circumfixation need to be considered. Also, Gaussier [23] has pointed out that some lexical derivations in French involve character-level alternations, such as "c" and "ç."

Another limitation of our method is that it does not generalize paraphrase pairs comprising words that have completely unrelated surface forms, such as ("look like", "resemble") and ("kick the bucket", "pass away"). To create more complete paraphrase lexicons, we will need some additional mechanisms that are able to deal with such idiosyncratic paraphrases.

Finally, as our method considers only corpora as the sources of paraphrase pairs, it will never acquire paraphrases that do not occur in its input corpora. Querying Web search engines might be a way to overcome this limitation, but might also result in noisier output.

4 QUANTITATIVE IMPACT

In this section, we evaluate the extent to which our proposed method can expand a given paraphrase lexicon, using English, French, and Japanese as target languages. We conducted the following three experiments.

Experiment 1. The seed set of paraphrase pairs, S_{Seed}, was created using the bilingual pivoting method. We conducted a learning-curve experiment in order to study the characteristics of our method relative to the quantity of input bilingual data.

Table 4. The Bilingual Data

Setting	# of sentences	# of tokens		
		English	French	Japanese
Europarl	2.01M	55.7M	61.9M	-
NTCIR	3.19M	107.6M	-	115.5M

Table 5. The Additional Monolingual Data Used with the Corresponding Side of Bilingual Data

Setting	Additional monolingual data	Language	# of sentences	# of tokens
Europarl	News Crawl 2011–2013	English	52.0M	1,203M
		French	19.4M	479M
NTCIR	NTCIR unaligned 2006–2007	English	39.9M	1,360M
		Japanese	136.5M	5,849M

Experiment 2. Another learning-curve experiment was conducted with regard to the size of input monolingual data.

Experiment 3. To demonstrate the versatility of our method, we applied it to several existing paraphrase lexicons created by other researchers.

4.1 Experiment 1: Learning Curve Relative to the Size of Input Seed Paraphrases

4.1.1 Data and Tools. As sources for acquiring seed paraphrase pairs, S_{Seed}, we used two different bilingual parallel corpora: the English–French version of the Europarl parallel corpus[4] and the NTCIR Japanese–English patent translation data.[5] Table 4 summarizes the number of sentences and tokens on each side of the bilingual data. For the learning-curve experiment, we created smaller subcorpora, maintaining their inclusion relation. First, a half-size subcorpus was created by randomly sampling 50% of the sentence pairs from the entire bilingual data. This process was repeated on each half-sized version until we obtained subcorpora of 1/128 size.

As for the monolingual data, we used News Crawl[6] 2011–2013 in combination with Europarl. For the NTCIR experiment, the 2006–2007 chapters of NTCIR unaligned patent documents were also used. Table 5 summarizes the numbers of sentences and tokens in the entire data. The concatenation of these monolingual data and the corresponding side of bilingual data were used to extract new paraphrase pairs and to compute contextual similarity of phrase pairs.

The English and French data were tokenized using the Moses toolkit[7] and the Japanese data were tokenized using MeCab.[8] To perform phrase-table cleaning and additional filtering steps, we used stoplists available on the Web:[9] 571 English and 463 French words. For Japanese, we manually listed 160 different morphemes.

4.1.2 Seed Paraphrase Pairs. The seed paraphrase pairs, S_{Seed}, were acquired using the bilingual pivoting method described in Section 3.1. The process began with the creation of a translation table from the input bilingual parallel corpus. IBM2 word alignment was determined using SyMGIZA++.[10] Phrase alignments of each sentence pair were identified using the "grow-diag-final"

[4] http://statmt.org/europarl/, release 7.
[5] http://ntcir.nii.ac.jp/PatentMT-2/.
[6] http://statmt.org/wmt14/translation-task.html.
[7] http://statmt.org/moses/, RELEASE-2.1.1.
[8] http://taku910.github.io/mecab/, version 0.996
[9] http://members.unine.ch/jacques.savoy/clef/.
[10] https://github.com/emjotde/symgiza-pp/.

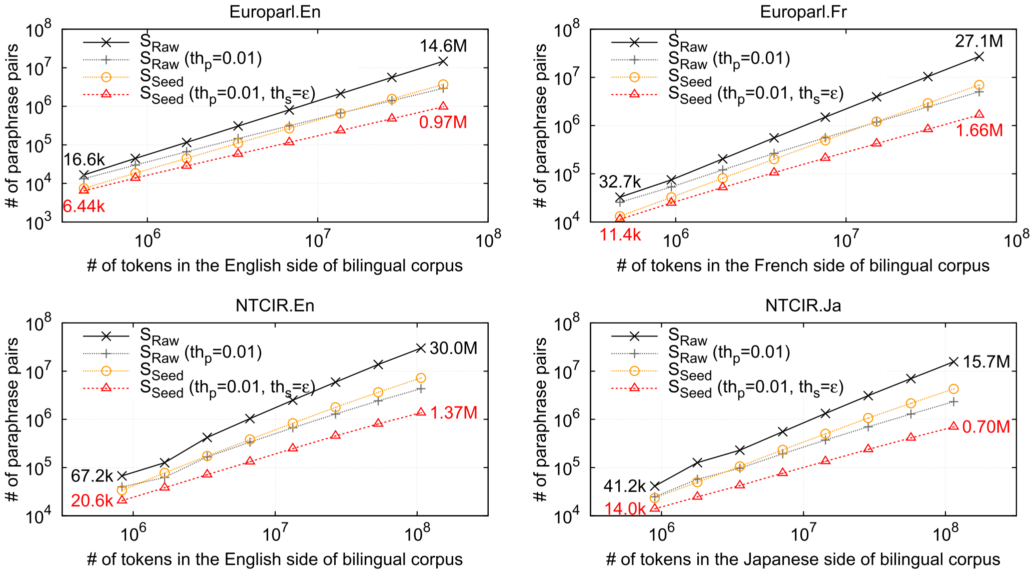

Fig. 4. Number of seed paraphrase pairs.

heuristic[11] with a maximum phrase length of eight. The resulting translation pairs were then filtered through significance pruning with $\alpha + \epsilon$ as threshold [33]. After extracting the initial set of paraphrase pairs, S_{Raw}, we performed extensive filtering as described in Section 3.1.

To compute the contextual similarity between the phrases of each paraphrase pair, we first extracted adjacent-word 1–4 grams of each occurrence of the phrase. For instance, from the preprocessed sentence in (12a), a total of eight features in (12b) were extracted for this occurrence of the phrase "amending the regulation."

(12) a. the commission is now amending the regulation and will take a final vote.
b. L1:now, L2:is:now, L3:commission:is:now, L4:the:commission:is:now,
R1:and, R2:and:will, R3:and:will:take, R4:and:will:take:a

Thereafter, each phrase was assigned a vector summing up all features recorded for that phrase together with their respective frequencies. This was meant as a compromise between less expensive but noisier approaches, such as bag-of-words, and more accurate but more expensive approaches that resort to syntactic features. Finally, the contextual similarity of any pair of phrases was calculated as the cosine of the respective feature vector of each phrase. As the reliability thresholds, we used $th_p = 0.01$ for paraphrase probability and $th_{s1} = \epsilon$ for contextual similarity. While the former is in line with common practice, the latter was chosen in order to discard only paraphrase pairs that were used in completely dissimilar contexts.

The numbers of paraphrase pairs acquired from the bilingual data are depicted in Figure 4. The general trend is simply that more bilingual data resulted in more paraphrase pairs. The lines with

[11] We used the "grow-diag-final" heuristic, while the "grow-diag-final-and" heuristic had been used more widely in the SMT community. The additional "and" constraint trusts a smaller number of word alignments, consequently producing larger numbers of phrase pairs involving nonaligned words and significantly skewing the estimates. As explained in Sections 3.1.1 and 3.1.2, most are incorrect as paraphrases; thus, we should avoid them proactively. See the following document for reference: http://statmt.org/moses/?n=FactoredTraining.AlignWords.

"◯" demonstrate that the filtering techniques described in Section 3.1.2 discarded a large portion of the raw results of bilingual pivoting, i.e., S_{Raw}, depicted with "×." More bilingual data resulted in a higher ratio of discarded pairs, suggesting that many incorrect and/or relatively useless pairs, such as those shown in Figures 2 and 3, had originally been acquired.

The lines with "+" show the results based on a widely used threshold value on the paraphrase probability, i.e., $th_p = 0.01$, directly applied to S_{Raw}. The percentage of discarded paraphrase pairs varies greatly depending on corpus size, suggesting that the threshold value should in principle be corpus dependent. However, to ensure the quality of S_{Seed}, we decided to adopt the standard threshold value, even though it turned out to discard some less frequent but perfectly good paraphrase pairs, such as ("control apparatus", "controlling device") in Figure 2. The resulting number of paraphrase pairs is labeled with "△" in Figure 4.

4.1.3 Paraphrase Patterns. Given our seed set of paraphrase pairs, S_{Seed}, we first induced affix patterns and then paraphrase patterns. The affix patterns in the English and French experiments were filtered with $k = 5$ and $n = 2$, following Gaussier [23]. In the Japanese experiments, we set $k = 2$ because stems of lexical variants in Japanese are mostly written using ideographical characters, i.e., *kanji*, and consequently tend to be short. There are actually many pairs of lexical variants that share only one *kanji*, as exemplified in (13), but we reluctantly determined that we needed to abandon them, because $k = 1$ would have yielded too many inappropriate affix patterns.

(13) a. ("近 i" (*tikai*; be close), "近 duku" (*tikaduku*; to get close))
b. ("高 ku" (*takaku*; be high), "高 me" (*takame*; to raise / be relatively high))
c. ("残 ru" (*nokoru*; to remain), "残 su" (*nokosu*; to leave))

Tables 6 and 7, respectively, show the most frequent suffix and prefix patterns obtained from the entire bilingual and monolingual data in each of the four settings. The "#*sup*" columns show the numbers of corresponding word pairs. To facilitate understanding of the induced patterns, we examined samples of corresponding word pairs of each one so that we could manually annotate it with the POS pairs observed in that sample. The top-ranked suffix patterns were typical inflection patterns, including verb conjugation, but some of the identified suffix patterns turned out to mark derivational patterns, i.e., patterns that relate different words, including those with different POS, such as ("X:ng", "X:on") and ("X:ation", "X:ing") in English, as opposed to different forms of the same word. In contrast, most of the prefix patterns acquired in our experiment reflected derivational processes, even though in most cases the POS remained the same, e.g., the verbs "do" and "redo." Since our acquisition process of affix patterns did not make use of POS information, some affix patterns could be derived simultaneously from lexical variants with different POS. For instance, the English pattern ("X:s", "X:ϵ") was acquired not only from pairs of plural/singular nouns (NNS, NN) but also from pairs made up of the third-person singular verb and its corresponding base form (VBZ, VB). For the reasons discussed in Section 3.3, we do not expect such ambiguity to seriously hurt accuracy. We will examine in our future work whether the use of a POS tagger brings worthwhile gains.

English affix patterns derived from different domains, i.e., those from Europarl and NTCIR settings, showed a notably large overlap, demonstrating the generality of lexical variants across domains. For instance, from the entire data in Europarl and NTCIR settings, our method derived 595 and 1,203 affix patterns, respectively, 308 of which were common. Our method also identified affix patterns specific to each domain, such as ("euro:X", "ϵ:X") in the Europarl setting, and ("X:ing", "X:or") and ("photo:X", "ϵ:X") in the NTCIR setting, each derived from, for instance, ("eurosceptics", "sceptics"), ("detecting", "detector"), and ("photoresist", "resist").

Table 6. The 20 Most Frequent Suffix Patterns Obtained from the Entire Data in Each Setting

Europarl.En

Pattern	#*sup*	Corresponding POS pairs
(“X:s”, “X:ϵ”)	1,300	(NNS, NN), (VBZ, VB)
(“X:ing”, “X:ϵ”)	373	(VBG, VB)
(“X:ing”, “X:e”)	330	(VBG, VB)
(“X:ing”, “X:ed”)	245	(VBG, VBD), (VBG, VBN)
(“X:ly”, “X:ϵ”)	234	(RB, JJ)
(“X:ed”, “X:ϵ”)	193	(VBD, VB), (VBN, VB)
(“X:d”, “X:ϵ”)	187	(VBD, VB), (VBN, VB)
(“X:ies”, “X:y”)	126	(NNS, NN)
(“X:ing”, “X:s”)	126	(VBG, VBZ)
(“X:ing”, “X:es”)	125	(VBG, VBZ)
(“X:d”, “X:s”)	110	(VBD, VBZ), (VBN, VBZ)
(“X:ng”, “X:on”)	99	(VBG, NN)
(“X:ation”, “X:ing”)	73	(NN, VBG)
(“X:ion”, “X:e”)	71	(NN, VB)
(“X:ed”, “X:s”)	69	(VBD, VBZ), (VBN, VBZ)
(“X:ion”, “X:ed”)	67	(NN, VBD), (NN, VBN)
(“X:n”, “X:ϵ”)	55	(JJ, NNP)
(“X:al”, “X:ϵ”)	48	(JJ, NN), (JJ, JJ)
(“X:ity”, “X:ϵ”)	46	(NN, JJ)
(“X:es”, “X:ϵ”)	44	(NNS, NN), (VBZ, VB)

Europarl.Fr

Pattern	#*sup*	Corresponding POS pairs
(“X:s”, “X:ϵ”)	3,893	(NCp, NCs), (VKp, VKs), (AQp, AQs)
(“X:e”, “X:ϵ”)	1,062	(VKfs, VKms), (AQfs, AQms)
(“X:es”, “X:ϵ”)	804	(VKfp, VKms), (AQfp, AQms)
(“X:es”, “X:s”)	754	(VKfp, VKmp), (AQfp, AQmp)
(“X:e”, “X:s”)	753	(AQfs, AQmp)
(“X:r”, “X:ϵ”)	375	(VW, VP1s)
(“X:nt”, “X:ϵ”)	328	(VP3p, VP1s)
(“X:nt”, “X:r”)	319	(VP3p, VW)
(“X:ant”, “X:er”)	241	(VG, VW)
(“X:er”, “X:é”)	240	(VW, VKms)
(“X:ment”, “X:ϵ”)	231	(Adv, AQ)
(“X:a”, “X:ϵ”)	213	(VF3s, VW)
(“X:er”, “X:ée”)	177	(VW, VKfs)
(“X:er”, “X:és”)	175	(VW, VKmp)
(“X:e”, “X:é”)	171	(VP1s, VKms), (VP1s, JJ)
(“X:ons”, “X:er”)	165	(VP1p, VW)
(“X:ont”, “X:a”)	160	(VF3p, VF3s)
(“X:ons”, “X:e”)	155	(VP1p, VP1s)
(“X:ant”, “X:e”)	152	(VG, VP1s)
(“X:ant”, “X:ent”)	140	(VG, VP3p)

NTCIR.En

Pattern	#*sup*	Corresponding POS pairs
(“X:s”, “X:ϵ”)	1,542	(NNS, NN), (VBZ, VB)
(“X:ing”, “X:ed”)	552	(VBG, VBD), (VBG, VBN)
(“X:ing”, “X:ϵ”)	425	(VBG, VB)
(“X:ed”, “X:ϵ”)	316	(VBD, VB), (VBN, VB)
(“X:d”, “X:ϵ”)	312	(VBD, VB), (VBN, VB)
(“X:ly”, “X:ϵ”)	278	(RB, JJ)
(“X:ing”, “X:e”)	257	(VBG, VB)
(“X:d”, “X:s”)	228	(VBD, VBZ), (VBN, VBZ)
(“X:ng”, “X:on”)	180	(VBG, NN)
(“X:ing”, “X:es”)	179	(VBG, VBZ)
(“X:ion”, “X:ed”)	155	(NN, VBD), (NN, VBN)
(“X:ing”, “X:er”)	147	(VBG, NN)
(“X:ing”, “X:s”)	139	(VBG, VBZ)
(“X:ed”, “X:s”)	133	(VBD, VBZ), (VBN, VBZ)
(“X:al”, “X:ϵ”)	108	(JJ, NN), (JJ, JJ)
(“X:on”, “X:ve”)	93	(NN, JJ)
(“X:ation”, “X:ing”)	84	(NN, VBG)
(“X:ing”, “X:or”)	82	(VBG, NN)
(“X:ion”, “X:or”)	82	(NN, NN)
(“X:ies”, “X:y”)	78	(NNS, NN)

NTCIR.Ja

Pattern	#*sup*	Corresponding POS pairs
(“X:ー”, “X:ϵ”)	243	(N, N)★
(“X:ru”, “X:ϵ”)	137	(V-TE, V-CO)
(“X:i”, “X:u”)	129	(V-CO, V-TE)
(“X:a”, “X:u”)	94	(V-IR, V-TE)
(“X:a”, “X:i”)	85	(V-IR, V-CO)
(“X:e”, “X:u”)	36	(V-HY, V-TE)
(“X:ホルダ”, “X:ϵ”)	36	(N+“holder”, N)★
(“X:つ”, “X:ru”)	32	(V-CO, V-TE)
(“X:e”, “X:i”)	29	(V-HY, V-CO)
(“X:ku”, “X:i”)	24	(A-CO, A-TE), (V-TE, V-CO)
(“X:ドライバ”, “X:ϵ”)	24	(N+“driver”, N)★
(“X:センサ”, “X:ϵ”)	20	(N+“sensor”, N)★
(“X:a”, “X:e”)	19	(V-IR, V-HY)
(“X:つ”, “X:ri”)	19	(V-CO, V-CO)
(“X:リング”, “X:ル”)	18	(N, V)★
(“X:uru”, “X:i”)	17	(V-TE, V-CO)
(“X:aru”, “X:e”)	16	(V-TE, V-CO)
(“X:aつ”, “X:e”)	16	(V-CO, V-CO)
(“X:eru”, “X:u”)	15	(V-TE, V-TE)
(“X:re”, “X:ϵ”)	15	(V-HY, V-CO)

Note: POS tags for English follow those used in the Penn Treebank (ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz). Acronyms for French are as follows: “NC” for common noun, {“VW”, “VG”, “VK”, “VP”, “VF”} for infinitive, present participle, past participle, present indicative, and future indicative of verb, respectively, “AQ” for qualificative adjective, “Adv” for adverb, “m” for masculine, “f” for feminine, “s” for single, “p” for plural, “1” for first person, and “3” for third person. See http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php for the comprehensive list. Acronyms for Japanese are as follows: N for noun, V for verb, A for adjective, “IR” for irrealis form, “CO” for continuative form, “TE” for terminal form, and “HY” for hypothetical form. “★” indicates patterns specific to loan words written in *katakana* (the square forms of kana).

A comparison with the 7,699 affix patterns in Catvar[12] revealed that 277 (126 suffix and 151 prefix) and 561 (265 suffix and 296 prefix) affix patterns in the above two respective settings were not covered by Catvar. As a consequence of relying on the surface forms of words instead of their base forms, our method collected many complex suffix patterns, such as (“X:ing”, “X:ors”) from

[12] https://clipdemos.umiacs.umd.edu/catvar/, version2.1.

Table 7. The 10 Most Frequent Prefix Patterns Obtained from the Entire Data in Each Setting

Europarl.En

Pattern	#*sup*	Corresponding POS pairs
(“re:X”, “ϵ:X”)	51	(NN, NN), (VB, VB)
(“un:X”, “ϵ:X”)	43	(NN, NN), (VB, VB), (JJ, JJ)
(“in:X”, “ϵ:X”)	23	(NN, NN), (JJ, JJ), (RB, RB)
(“mis:X”, “ϵ:X”)	15	(NN, NN), (VB, VB)
(“pre:X”, “ϵ:X”)	13	(NN, NN), (JJ, JJ)
(“dis:X”, “ϵ:X”)	11	(NN, NN), (VB, VB)
(“inter:X”, “ϵ:X”)	11	(NN, NN), (VB, VB), (JJ, JJ)
(“over:X”, “ϵ:X”)	9	(NN, NN), (JJ, NN)
(“co:X”, “ϵ:X”)	8	(NN, NN), (VB, VB)
(“euro:X”, “ϵ:X”)	7	(NN, NN)

Europarl.Fr

Pattern	#*sup*	Corresponding POS pairs
(“re:X”, “ϵ:X”)	91	(V, V), (N, N), (A, A)
(“in:X”, “ϵ:X”)	67	(A, A), (N, N), (V, V)
(“r:X”, “ϵ:X”)	30	(V, V)
(“dé:X”, “ϵ:X”)	26	(V, V), (N, N)
(“ré:X”, “ϵ:X”)	23	(V, V), (N, N)
(“pré:X”, “ϵ:X”)	22	(V, V), (N, N), (A, A)
(“inter:X”, “ϵ:X”)	16	(A, A), (N, N)
(“sur:X”, “ϵ:X”)	16	(N, N), (V, V)
(“e:X”, “é:X”)	15	(N, N), (A, A)
(“con:X”, “ϵ:X”)	13	(N, N), (A, A)

NTCIR.En

Pattern	#*sup*	Corresponding POS pairs
(“sub:X”, “ϵ:X”)	53	(NN, NN)
(“re:X”, “ϵ:X”)	49	(NN, NN), (VB, VB)
(“photo:X”, “ϵ:X”)	47	(NN, NN), (JJ, JJ)
(“micro:X”, “ϵ:X”)	42	(NN, NN)
(“un:X”, “ϵ:X”)	42	(NN, NN), (VB, VB), (JJ, JJ)
(“inter:X”, “ϵ:X”)	29	(NN, NN), (VB, VB), (JJ, JJ)
(“de:X”, “ϵ:X”)	25	(NN, NN), (VB, VB)
(“pre:X”, “ϵ:X”)	25	(NN, NN), (JJ, JJ)
(“electro:X”, “ϵ:X”)	22	(NN, NN), (JJ, JJ)
(“in:X”, “ϵ:X”)	20	(NN, NN), (JJ, JJ), (RB, RB)

NTCIR.Ja

Pattern	#*sup*	Corresponding POS pairs
(“インナ:”, “ϵ:X”)	17	(“inner”+N, N)★
(“データ:X”, “ϵ:X”)	17	(“data”+N, N)★
(“取 ri:X”, “ϵ:X”)	16	(“take”+V, V)
(“取 ri:X”, “取:X”)	16	orthographic variant
(“リヤ:X”, “ϵ:X”)	15	(“rear”+N, N)★
(“ロウ:X”, “ϵ:X”)	15	(“low”+N, N)★, (“raw”+N, N)★
(“引 ki:X”, “ϵ:X”)	15	(“pull”+V, V)
(“ロータリ:X”, “ϵ:X”)	14	(“rotary”+N, N)★
(“デ:X”, “ϵ:X”)	12	(N, N), (V, V)★
(“不:X”, “ϵ:X”)	12	(N, N), (A, A)

Note: Acronyms follow those in Table 6.

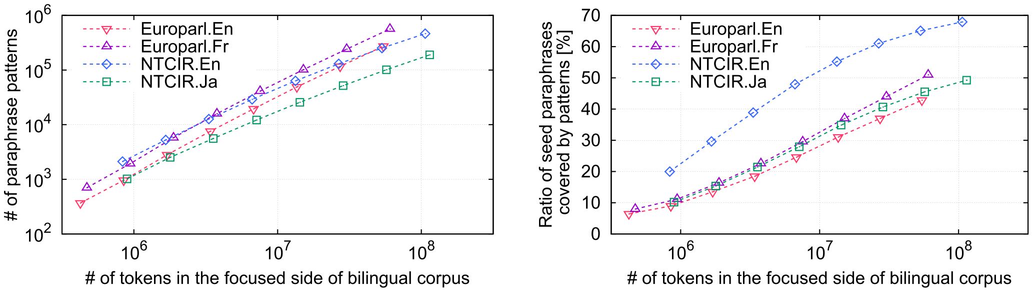

Fig. 5. Paraphrase patterns (left: number; right: coverage).

(“generating”, “generators”) and (“X:fully”, “X:ϵ”) from (“peacefully”, “peace”), and some regular suffixations, such as (“X:est”, “X:ϵ”) for superlative and base forms of adjectives, that Catvar ignored. On the other hand, the 10 most frequent paraphrase patterns presented in Table 7 were all missing in Catvar. We also noticed the limitation of our method; apparently it cannot capture the generality that is exhibited less frequently. For instance, in the Europarl setting, we had only one pair of words, (“observation”, “observer”), of (“X:ation”, “X:er”), and no pair of words of (“X:sm”, “X:ze”), which was exhibited by (“criticism”, “criticize”) and (“formalism”, “formalize”) in Catvar.

Using the acquired affix patterns to capture lexical variants, we then induced paraphrase patterns. Figure 5 shows the number of acquired paraphrase patterns and their coverage, i.e., the percentage of seed paraphrase pairs that get generalized into a pattern. More seed paraphrase pairs resulted in more patterns and higher coverage. When the entire bilingual and monolingual data were used, 40–70% of the paraphrase pairs in S_{Seed} were generalized into patterns. In the Europarl setting, we obtained more patterns and higher coverage for French than for English. This can be explained by the fact that French morphology is significantly richer than that in English. In the

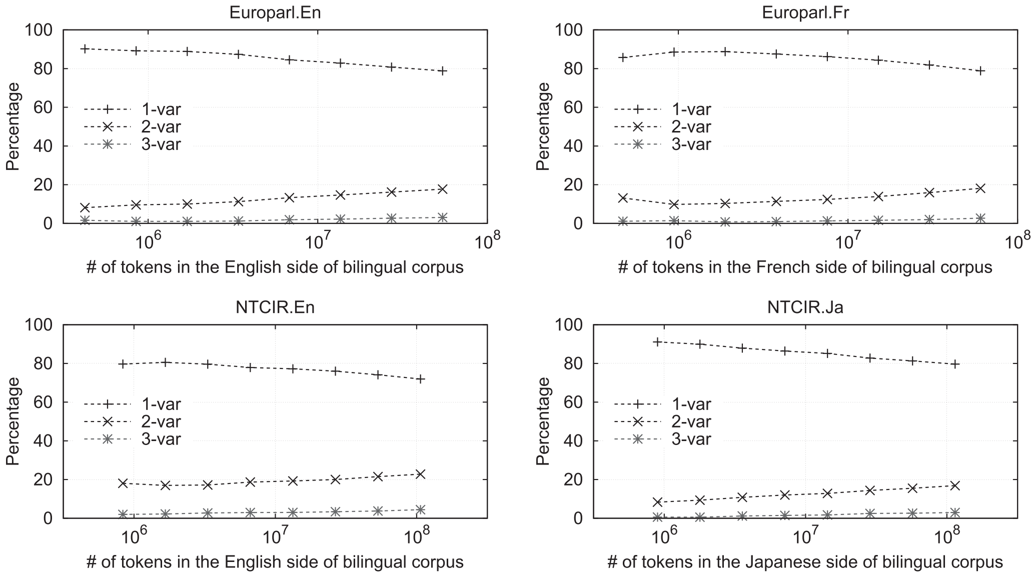

Fig. 6. Percentage of the 1-variable, 2-variable, and 3-variable paraphrase patterns.

NTCIR setting, we extracted more patterns for English than for Japanese. This could be a result of the fact that we missed some patterns by abandoning affix patterns that were supported by only short stems, such as those presented in (13). Accordingly, some seed paraphrase pairs failed to be generalized, and some partial patterns failed to be merged into a single general pattern. We are planning to address this shortcoming in our future work through an improved treatment of Japanese data, e.g., by using phonetic transcriptions of ideographical characters as well.

As shown in Figure 6, most of the acquired paraphrase patterns contained only one variable. However, the percentages of paraphrase patterns with more than one variable tended to grow with the size of bilingual data. The maximum number of variables was five in both languages for the Europarl setting and seven in both languages for the NTCIR setting, but many of them merely contain identical word sequences.

Unlike affix patterns, paraphrase patterns in the two English settings had only a small intersection. For instance, the 273 thousand and 464 thousand patterns that were obtained from the entire Europarl and NTCIR data, respectively, had only 5,226 patterns in common. With smaller quantities of bilingual data, the NTCIR setting resulted in more paraphrase patterns than the Europarl setting and the patterns in the NTCIR setting always achieved conspicuously higher coverage than those in the Europarl setting. We observed that the NTCIR data is richer in the kinds of expressions that tend to induce a lot of variation, such as technical terms and nominalizations. Our analysis of the number of RHS phrases per LHS phrase is reported in Section 4.1.4.

4.1.4 New Paraphrase Pairs Harvested from Monolingual Data. Finally, new paraphrase pairs, S_{Hvst}, were harvested from the entire monolingual data by using the induced paraphrase patterns. We only considered single words as potential slot-fillers of our paraphrase patterns and filtered the collected pairs of phrases on the basis of their contextual similarity, as for S_{Seed} in Section 4.1.2, with $th_{s2} = \epsilon$. We also excluded paraphrase pairs already present in S_{Seed}.

As depicted by the lines labeled “Pairs” in Figure 7, we managed to harvest remarkably large numbers of paraphrase pairs, irrespective of the amount of bilingual data. For instance, in the

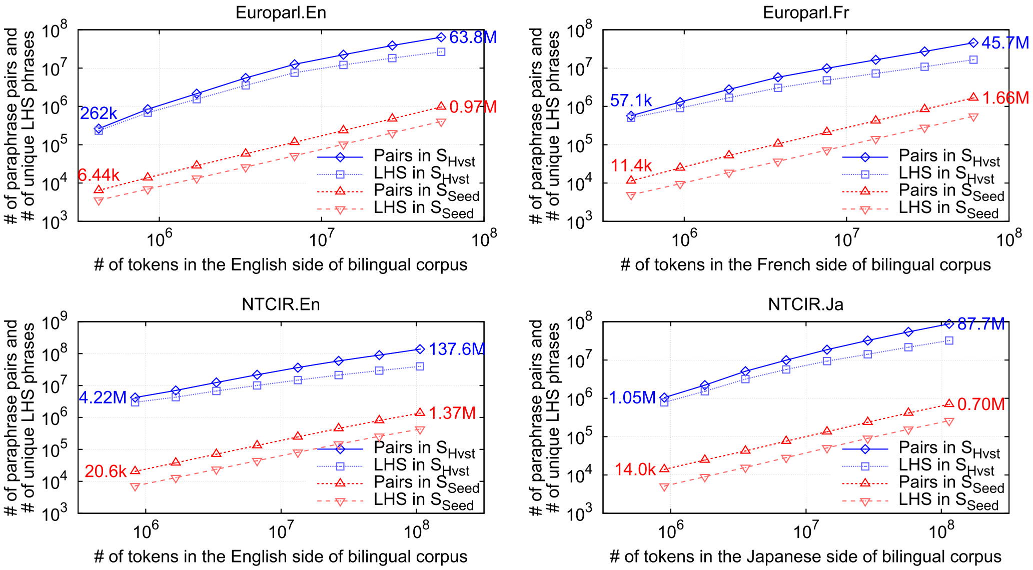

Fig. 7. Number of seed and newly harvested paraphrase pairs and unique LHS phrases covered.

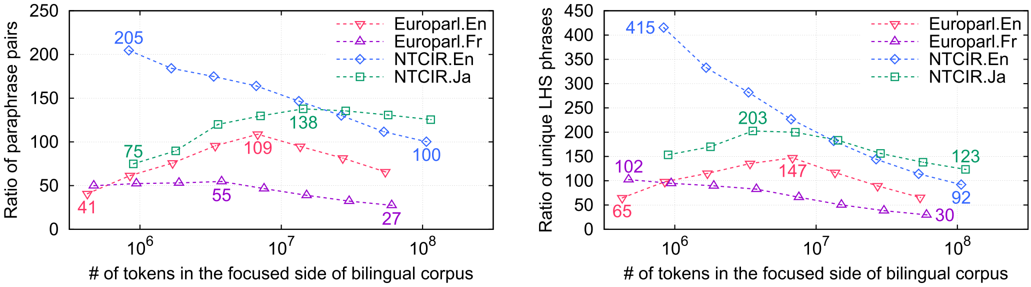

Fig. 8. Leverage ratio (left: paraphrase pairs; right: unique LHS phrases).

Europarl English setting with the entire bilingual data, we acquired 971 thousand pairs of seed paraphrases, S_{Seed}, and 63.8 million new paraphrase pairs, S_{Hvst}. As the seed set acquired early in the process can be pooled with the newly harvested set, we can say that our method expanded S_{Seed} by a factor of 67. The lines labeled "LHS" show the number of unique LHS phrases, i.e., phrases that have at least one paraphrase. These results show that our method is making a substantial contribution to the discovery of paraphrases that were missing in S_{Seed}.

Figure 8 highlights that our method has achieved a remarkably large **leverage ratio** of S_{Hvst} to S_{Seed}, with regard to both the numbers of paraphrase pairs and of unique LHS phrases. Except for the NTCIR English setting, the ratio peaked at when the middle-sized bilingual data were used. When more than 1/8 of the entire bilingual data were used, the ratio was monotonically decreasing in all of the settings. This trend reflects the ratio of the monolingual and bilingual data. In other words, when sufficiently large bilingual data are available, the need for harnessing monolingual data decreases. However, when the amount of bilingual data becomes too small, the extracted set of seed paraphrase pairs might not comprise a sufficiently large variety of productive patterns to gain a lot of leverage. Consequently, the leverage ratio may not be as high as the ratio of the monolingual

Table 8. Sample Paraphrase Patterns in the NTCIR English Setting and the Number of Corresponding Paraphrase Pairs

Paraphrase pattern	Sample paraphrase pair	1/128 data		Full data	
		S_{Seed}	S_{Hvst}	S_{Seed}	S_{Hvst}
(“X:ϵ Y:ϵ”, “Y:ϵ X:s”)	(“signal data”, “data signals”)	2	206,998	306	224,375
(“X:ϵ Y:ϵ”, “Y:ϵ of X:s”)	(“pulse number”, “number of pulses”)	1	63,134	74	67,674
(“X:ing Y:ϵ”, “Y:ϵ X:ϵ”)	(“reading data”, “data read”)	1	41,828	73	45,730
(“for X:ing”, “to X:e”)	(“for driving”, “to drive”)	1	1,876	65	1,841

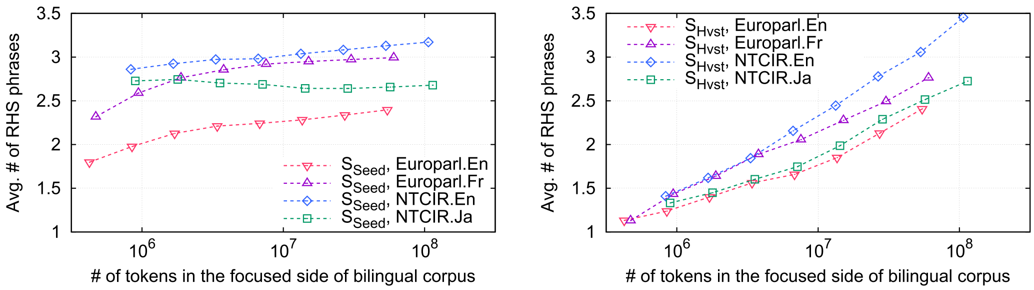

Fig. 9. Average yield (left: seed paraphrase pairs; right: new paraphrase pairs).

to bilingual data. The NTCIR English setting showed an exceptional trend. The ratio monotonically decreased with the scaling up of the bilingual data. We confirmed that some productive patterns, such as those exemplified in Table 8, were acquired even from small numbers of seed paraphrases. The leverage ratio of S_{Hvst} to S_{Seed} can be increased further by scaling up the monolingual data, as we investigate in Section 4.2.

Another striking difference between S_{Seed} and S_{Hvst} is the **yield**, i.e., the number of RHS phrases associated with an LHS phrase.[13] As displayed in Figure 9, the average yield for S_{Hvst} increased rapidly with the scaling up of the bilingual data, while that of S_{Seed} grew relatively slowly. The pivoting method based on bilingual data cannot produce very many RHS phrases per unique LHS phrase as a result of its reliance on word/phrase alignment, conditional probability, and surface forms of words. In contrast, our method does not limit the number of RHS phrases. Instead, it assesses each potential RHS phrase on the basis of its contextual similarity to the corresponding LHS phrase. Our method is unable to achieve high yield for S_{Hvst} whenever only a small number of paraphrase patterns can be induced from the given seed paraphrase pairs (see also Figure 5). Just like the leverage ratio, the average yield can be increased by harnessing larger monolingual data; this is discussed in Section 4.2.

As mentioned in Section 4.1.3, the two English settings derived only a small number of common paraphrase patterns. Nevertheless, they actually contributed to obtaining a large number of paraphrase pairs. For instance, 29.7% (19.0/63.8 million) of the paraphrase pairs obtained in the Europarl English setting can also be acquired through the application of the paraphrase patterns induced in the NTCIR English setting to the News Crawl monolingual data. In the opposite direction, 19.3% (26.6/137.6 million) of the paraphrase pairs from the NTCIR English setting can be acquired. Although these ratios appear low, the numbers of acquirable paraphrase pairs are

[13] Unlike the work in Szpektor et al. [54], we did not extract only correct pairs to calculate the yield.

Table 9. The Monolingual Subcorpora

Setting	Label	# of sentences	# of tokens
Europarl	Small	7,229,454	164,642,114
	Large	51,995,709	1,203,454,589
NTCIR	Small	4,509,076	159,277,043
	Large	39,864,775	1,359,686,076

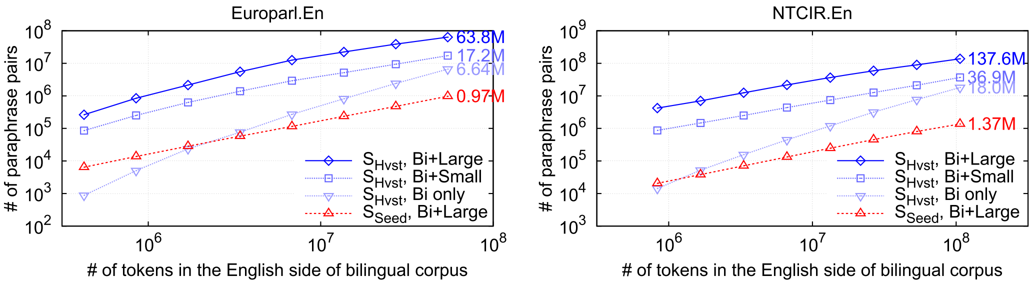

Fig. 10. Number of new paraphrase pairs (cf. Figure 7).

significantly larger than those in S_{Seed} (see Figure 4), suggesting that our generalization approach can render the S_{Seed} from a given domain useful in different domains.

4.2 Experiment 2: Comparison of Different Sizes of Monolingual Data

We also investigated how the leverage ratio of S_{Hvst} to S_{Seed} and the average yield of S_{Hvst} are affected by the size of monolingual data, comparing three different sizes of monolingual data in the two English settings. The first data, “Bi+Large”, is identical to that of the first experiment in Section 4.1. The second, “Bi+Small”, is a subset of the entire monolingual data, as shown in Table 9. The third data, “Bi only”, includes only the English side of the bilingual data without any additional monolingual data. In our method, the monolingual data is used not only as the source of new paraphrase pairs but also as the resource used for computing contextual similarity of paraphrase pairs. Nevertheless, with the same threshold $th_{s1} = \epsilon$, more than 90% of paraphrase pairs in S_{Seed} of the “Bi+Large” setting were retained by the “Bi only” setting. Consequently, almost the same sets of paraphrase patterns were acquired with the three different sizes of monolingual data, suggesting that the size of bilingual data is more important than that of monolingual data, most likely because of its relationship with the diversity of paraphrase patterns.

Figure 10 depicts the number of newly harvested paraphrase pairs, S_{Hvst}, and Figure 11 shows the leverage ratio of each S_{Hvst} to its corresponding S_{Seed}. Note that the vertical axis of Figure 11 is displayed in log scale, unlike Figure 8. The obvious observation is that the larger the monolingual data is, the more paraphrase pairs get acquired and the higher the leverage ratio is. The results of the “Bi only” setting demonstrate that our method is able to find a large number of paraphrase pairs even when the data is limited to whatever was used to acquire S_{Seed}. This follows from our exploitation of generalizations that underlie the paraphrase pairs in S_{Seed}. However, only some hundreds of new paraphrase pairs were harvested when the minimum size of data was used, i.e., the leftmost points of the “Bi only” settings. This confirms the value of additional monolingual data. On the leverage ratio, the two settings with additional monolingual data exhibited the same trend with regard to the size of bilingual data. In contrast, the lack of additional monolingual data in the “Bi only” setting resulted in a monotonic increase.

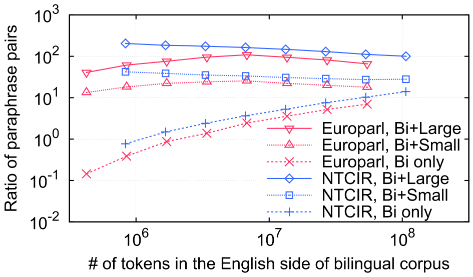

Fig. 11. Leverage ratio (cf. Figure 8).

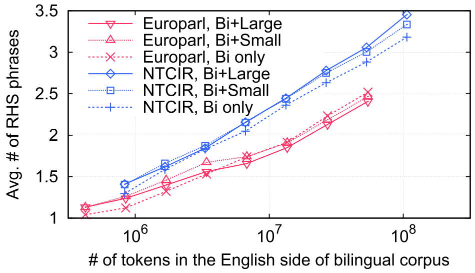

Fig. 12. Average yield of new paraphrase pairs (cf. Figure 9).

Figure 12 compares the average yield of the three settings with different sizes of monolingual data. As the sets of paraphrase patterns were almost the same in each setting, the yield of a given LHS depends entirely on whether the RHS phrase of each corresponding pattern is found in the given monolingual data. Nevertheless, the results indicate that the average yield is not significantly increased by scaling up the monolingual data. The main lesson is that the diversity of phrase constructions basically depends on the paraphrase patterns and, ultimately, on the size of underlying bilingual data.

Although the above experiments are just simulations, they demonstrate the potential applicability of our method to relatively low-resource conditions. The leftmost points of the Europarl setting in Figure 10 correspond to bilingual data with 400 thousand tokens or 16 thousand sentence pairs according to Table 4. Such quantities of bilingual data are available, for instance, through OPUS,[14] for many languages including some that are generally considered as low resourced. In cases in which less bilingual data happen to be available, it would be possible to crowdsource the missing portion, as in Tatoeba.[15] As for monolingual data, one can rely on the Web. For instance, researchers have exploited Wikipedia, which provides more than 10 thousand documents for more than 130 languages.[16] In our future work, we will examine the usefulness of our method for acquiring paraphrases in severely low-resourced languages.

4.3 Experiment 3: Expanding Existing Paraphrase Lexicons

To demonstrate the versatility of our method, we also applied it to two types of pre-compiled paraphrase lexicons in English. One is the set of paraphrase pairs acquired from a Web-harvested monolingual comparable corpus made up of definition sentences (henceforth "Web-Def") [58]. We prepared subsets of different sizes as our S_{Seed}, varying the threshold value on the score of each extracted pair: ϵ (entire set), 1, 2, 3, 4, 5, 6, and 7. The other is the English edition of PPDB-1.0 [22] created from a large-scale parallel corpus using the bilingual pivoting method enhanced with syntactic constraints. We used six different sizes of packages, merging lexical, one-to-many, many-to-one, and phrasal datasets.[17] For the monolingual data, we used the entire News Crawl data used in the first experiment (Section 4.1) for both types of lexicons.

For comparison, we also created subsets of S_{Seed} in the Europarl and NTCIR English settings with all of the bilingual and monolingual data from the first experiment, pruning them with different threshold values for th_p: 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7. Then, we expanded them using

[14] http://opus.lingfil.uu.se/.
[15] https://tatoeba.org/eng/.
[16] See the Japanese article through https://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics.
[17] http://paraphrase.org/#/download.

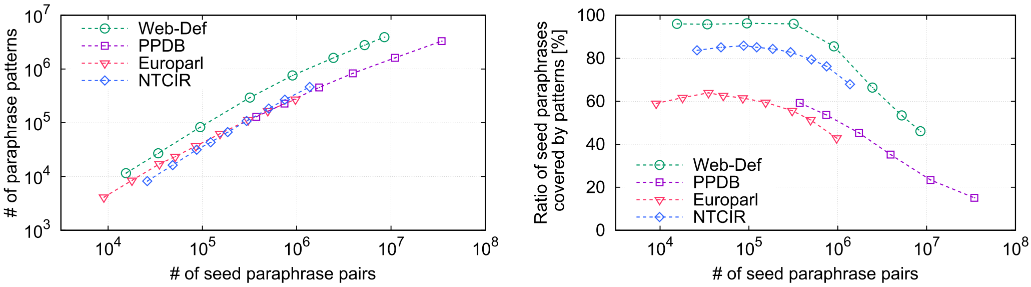

Fig. 13. Number and coverage of paraphrase patterns (cf. Figure 5).

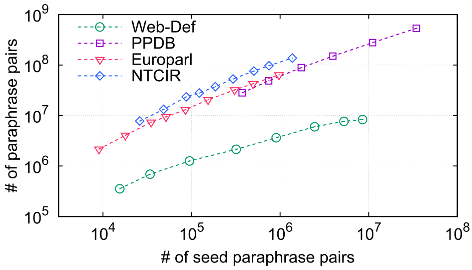

Fig. 14. Number of new paraphrase pairs (cf. Figure 7).

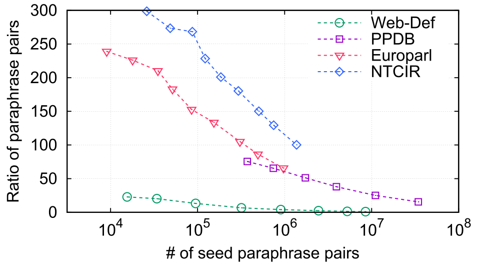

Fig. 15. Leverage ratio (cf. Figure 8).

the corresponding monolingual data, including the English side of the bilingual data, as in the first experiment.

Figure 13 summarizes the number of induced paraphrase patterns and their coverage for S_{Seed}. For all four datasets, the number of paraphrase patterns exhibited the same trend as in the learning-curve experiment (cf. Figure 5). In contrast, we observed higher coverage when only reliable parts of seed paraphrase lexicons were used. This indicates that paraphrase pairs that are scored high with preexisting methods are more susceptible to the kind of generalizations that our method exploits. The number and coverage of paraphrase patterns induced from Web-Def were especially higher than those obtained from the other datasets. There are two reasons for this. One is that the score of each pair has been computed taking into account the existence of shared words across the phrases of the pair. Another reason is that S_{Seed} of Web-Def includes relatively long phrases. While the PPDB contained phrases of length up to six tokens, depending on the threshold value, 47.4–64.9% of paraphrase pairs in S_{Seed} of Web-Def contained phrases comprising more than six tokens. A single pattern with short phrases is often derived from many paraphrase pairs; in contrast, patterns for long phrases tend to correspond to fewer pairs of phrases, reflecting a lower degree of generality. Consequently, we obtained more patterns of lesser generality.

Figure 14 displays the number of newly harvested paraphrase pairs. Again, we observe the same trend as in the learning-curve experiment. However, as is evident in Figure 15, the paraphrase patterns for Web-Def, despite their large number, harvested significantly fewer paraphrase pairs than the other datasets. As it turned out, most of the long patterns made no contribution whatsoever. For instance, when the entire S_{Seed} was used (rightmost points in the charts), 99.5% (8.27/8.31 million) of S_{Hvst} were extracted by patterns that have up to six tokens on each side. However, the S_{Seed} of Web-Def failed to yield short and highly productive patterns. For instance, among the 100 most productive patterns induced from the largest package (XXXL) of the PPDB, only seven could be

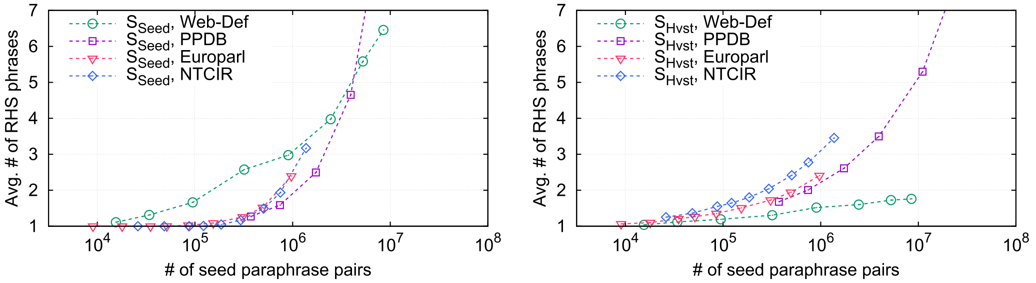

Fig. 16. Average yield (left: seed paraphrase pairs; right: new paraphrase pairs; cf. Figure 9).

induced from the entire S_{Seed} of Web-Def, while obvious pairs, such as (“X_1:ing X_2:ϵ”, “to X_1:ϵ X_2:ϵ”) and (“X_1:ϵ X_2:ed”, “X_2:ing X_1:ϵ”), were missing. Consequently, our method achieved a low leverage ratio for Web-Def only. On the other hand, the paraphrase pairs harvested with the PPDB showed the same tendencies as those of our Europarl setting. The 48.5 million paraphrase pairs in the S_{Hvst} acquired through the second smallest package (M) was larger than the 34.4 million pairs in the S_{Seed} of the largest package (XXXL), but, interestingly, they shared only 210 thousand pairs. In other words, 99.6% of the former were unseen even when all the available parallel data was used. This highlights the benefits of our generalization-and-instantiation approach in leveraging large-scale monolingual data.

Finally, we compared the average yield of S_{Seed} and S_{Hvst} as in Figure 16 and observed a general trend: the increase according to the size of S_{Seed}. However, when we used the larger packages from the PPDB, we got suspiciously high average yields: 11.6 and 34.1 for the respective S_{Seed} of the two largest packages, i.e., XXL and XXXL, and 8.9 for the S_{Hvst} of the largest one, XXXL. We found that the yield for the phrase “to bring” was 3,571 in the largest S_{Seed}, and 39 of them had derived paraphrase patterns, including many inappropriate ones, such as (“to X_1:ϵ”, “X_1:ing an”) and (“to X_1:ϵ”, “in X_1:ing the”). Another phrase, “supporting the,” had 1,174 RHS phrases in the S_{Hvst} of which 349 were obtained with different paraphrase patterns, most of which turned out to be nonsensical. In this case, the effect of our method was to amplify the noise present in a large but unreliable source. The quality of seed paraphrases is clearly a crucial success factor.

5 QUALITY ASSESSMENT THROUGH EVALUATING PARAPHRASE SUBSTITUTIONS

While the previous section focused on the quantitative aspect of expanding paraphrase lexicons, this section turns to the qualitative aspect. Automatically extracted paraphrase lexicons are generally too large to be evaluated exhaustively. Thus, researchers have typically resorted to manually evaluating samples. Since the work in Szpektor et al. [53], substitution-based evaluation has been adopted as a means of assessing the equivalence of a given paraphrase pair. Specifically, good quality pairs, such as (“looks like”, “resembles”), have the property that substituting one of the phrases of the pair with the other in a given sentence “s” should yield an equivalent sentence “t” as in (14).[18]

(14) s. The roof *looks like* a prehistoric lizard’s spine.
t. The roof *resembles* a prehistoric lizard’s spine.

We generated pairs of sentences, such as those of (14), using the paraphrase lexicons to be evaluated and then asked human evaluators to assess the quality of the paraphrased sentences.

[18] Throughout this section, original and paraphrased sentences are labeled “s” and “t”, respectively.

5.1 Evaluation Criteria and Procedure

5.1.1 Grammaticality and Meaning Equivalence. Precise guidelines are indispensable to successfully harness human judgments. While the term "paraphrase" primarily refers to **meaning equivalence**, **grammaticality** cannot be ignored in the paraphrase generation task. We use the following definitions.

Grammaticality. Whether the paraphrased sentence is syntactically correct.[19]
Meaning equivalence. Whether the meaning of the original sentence is properly preserved in the paraphrased sentence.

The following examples are presented in order to help sharpen the distinction between these two concepts. The paraphrased sentence in (15)[20] has no grammatical problem. However, the substitution of "global economy" with "environmental issues" yields a sentence that is not equivalent in meaning, even though in both cases one is talking about some kind of social issue.

(15) s. The leaders discussed the *global economy*.
t. ≠The leaders discussed the *environmental issues*.

In contrast, the phrase substitution in (16) does affect the grammaticality while the meaning appears to be unaffected. One could easily correct this grammaticality problem without referring to the original sentence. We aim to evaluate meaning equivalence separately from grammaticality. If no information gets added or lost, the two sentences will be marked as equivalent regardless of grammaticality issues.

(16) s. I like to be *30 years* old.
t. *I like to be *age of 30* old.

The example in (17) requires a more careful judgment. Clearly, the two phrases of the pair ("a movement against racism", "an anti-racism movement") are semantically equivalent. However, the substitution of these phrases within the sentence below has the effect of radically altering the meaning of the sentence by removing the coordination "racism and fascism." However, grammaticality remains unaffected in this case.

(17) s. They expressed support for *a movement against racism* and fascism.
t. ≠They expressed support for *an anti-racism movement* and fascism.

5.1.2 Classification-Based Evaluation. In Callison-Burch [8], evaluators are asked to rate grammaticality and meaning equivalence along two five-point scales. Malakasiotis and Androutsopoulos [42] used a variant of this method with four-point scales. In both experiments, evaluators are provided with some guidelines about the meaning of each score. However, in practice, numerical scores tend to be difficult for evaluators to use consistently. In our preliminary experiment, we made use of the scoring approach of Callison-Burch [8], and observed that some evaluators gave different scores to examples that had very similar types of errors. This is no doubt why some recent work, such as [36], used only a coarse-grained version of the evaluation scales. We would like to minimize intra-evaluator inconsistency as much as possible while retaining a fine-enough granularity for the results to prove useful in error analysis.

[19] Following the Chomskyan tradition, we showed "Colorless green ideas sleep furiously." [11] as a grammatical but semantically deviant example.
[20] The symbols "*" and "≠" in front of paraphrased sentences indicate ungrammaticality and nonequivalent meaning, respectively.

Table 10. Classification Labels

Criterion	Coarse	Fine-grained
Grammaticality	Positive	Perfect, Awkward
	Negative	Minor Problem, Major Problem, Irredeemable
Meaning equivalence	Positive	Equivalent, Missing Info., Additional Info., Ignorable Change
	Negative	Significantly Different, Completely Different

Table 11. Classification Results by the Authors

Example	Grammaticality	Meaning equivalence
(14)	Perfect	Equivalent
(15)	Perfect	Significantly Different
(16)	Irredeemable	Equivalent
(17)	Perfect	Significantly Different

Instead of numerical scoring, we asked our evaluators to classify each example into one of several predefined classes. Table 10 shows the lists of classes for each of two different granularity levels. To guide evaluators on the classification task, we provided them with decision trees branching on a set of basic questions (see Appendices A and B). Thus, assigning a class label to an example amounts to providing an answer to each of a series of basic questions. Table 11 shows classification results for the examples in (14) to (17) as determined by the authors.

5.1.3 Unitwise Two-Phased Evaluation. We packaged into a single **example unit** several paraphrase examples of each source phrase so that the different paraphrases could be presented together to each evaluator. This was found to help evaluators produce more consistent judgments. Also, to minimize the potential confusion between grammaticality and meaning equivalence, we controlled the evaluation process as follows.

Step 1. Grammaticality. In the first step, when an example unit is selected by an evaluator, only the paraphrased sentences are shown. The evaluator judges their respective grammaticality without seeing the original sentence. If the judge finds that a given sentence is ungrammatical but in a manner that is not caused by the substituted phrase, the judge ignores it, assuming that the ungrammaticality is inherited from the original sentence.

Step 2. Meaning equivalence. Once the grammatical assessment is completed, the original sentence is shown alongside its paraphrases. The evaluator is now asked to judge to what extent the meaning of each paraphrased sentence preserves that of the original sentence.

This unitwise evaluation also has the effect of reducing human effort by avoiding repeated reading and understanding the original sentences.

5.1.4 Evaluation Tool. We designed and implemented our own Web-based evaluation tool to provide direct support for our unitwise two-phase classification task. Unlike existing crowdsourcing platforms, such as Amazon Mechanical Turk,[21] ours allows evaluators to postpone and revise their judgments on specific examples. Throughout the evaluation process, we encouraged evaluators to reconsider past examples in order to make their judgments as consistent as possible.

[21] https://www.mturk.com/.

Table 12. Number of Generated Paraphrases and Example Units

Setting	# of unique sentences		# of paraphrases	# of example units	
	all	targeted length		all	3+ examples
Europarl English	9,000	5,850	1,013,511	88,555	31,149
Europarl French	8,995	5,639	1,391,162	97,903	34,706
NTCIR English	4,288	2,730	2,727,399	135,726	47,116
NTCIR Japanese	4,285	2,701	864,434	100,585	18,273

We avoided pointing out particular examples of disagreement among evaluators for fear that this would discourage evaluators from reconsidering other units and sentences that needed it.

5.2 Generating Examples for Evaluation

We evaluated the two paraphrase lexicons, S_{Seed} and S_{Hvst}, acquired using the entire bilingual and monolingual data in each of the four settings in Section 4.1.

As for the original sentences for the two Europarl settings, we used the "newstest" data in WMT 2011–2013, similarly to the work in Callison-Burch [8]. To reduce the human labor for the evaluation, we selected only sentences of moderate length: 10–30 words, which we expected to provide sufficient context around the substituted phrases. Test sentences for evaluating the paraphrase lexicons in the two NTCIR settings were drawn from NTCIR 9 and 10 Patent MT tasks.[22] As the average length of sentences is larger in patent documents than in news articles, we used larger values on the length constraints: 20–50 words.

Paraphrased sentences were generated simply by substituting a matching phrase of any pair from the paraphrase lexicon by its counterpart(s). Table 12 presents the statistics of generated paraphrase examples in each of the four settings. For instance, in the Europarl English setting, we obtained 88,555 example units containing 1,013,511 paraphrases. Then, for each example unit, the paraphrased sentences were ranked on the basis of a 5-gram language model trained on the corresponding monolingual data using KenLM[23] with modified Kneser–Ney smoothing. We used a naïve LM-based method to rank the candidates rather than a classifier, as in Zhao et al. [59], because our focus was the evaluation of the resources and we did not have any labeled training data. Finally, the example units for evaluation were randomly sampled from those containing at least three candidates, shown in the rightmost column in Table 12, and the 3 best candidates were selected. Regarding the Europarl English setting as the primary target, we sampled 200 example units that contain paraphrases for 200 different phrases. For the other three settings, we performed a small-scale evaluation using 40 example units for 40 different phrases.

5.3 Quality of Paraphrase Lexicons in the Europarl English Setting

For the Europarl English setting, we separately collected evaluations from three anonymous native English speakers. Table 13 summarizes the inter-evaluator agreement ratio, Cohen's κ [12] and Fleiss's κ [18], for which "fine-grained" and "coarse-grained" refer to the results based on the corresponding granularity of evaluation. The values for the coarse-grained results were "substantial" for grammaticality and "moderate" for meaning equivalence [37]. We also observed that the

[22] http://ntcir.nii.ac.jp/PatentMT-2/.
[23] https://kheafield.com/code/kenlm/.

Table 13. Range of Cohen's κ of Pairwise Agreement and Fleiss's κ for All Three Evaluators

Criterion	Fine-grained		Coarse-grained	
	Cohen	Fleiss	Cohen	Fleiss
Grammaticality	0.51 - 0.56	0.53	0.64 - 0.79	0.72
Meaning equivalence	0.27 - 0.35	0.29	0.48 - 0.53	0.50

Table 14. Precision of Paraphrase Substitution Using the Paraphrase Lexicons in the Europarl English Setting

Aggregation Method	Lexicon	N	Grammaticality		Meaning equivalence		Both	
			#	%	#	%	#	%
Individual judgments	S_{Seed}	198	169	0.85	172	0.87	147	0.74
	S_{Hvst}	1,602	1,200	0.75	1,230	0.77	938	0.59
	Total	1,800	1,369	0.76	1,402	0.78	1,085	0.60
Majority voting	S_{Seed}	66	56	0.85	60	0.91	50	0.76
	S_{Hvst}	534	396	0.74	416	0.78	314	0.59
	Total	600	452	0.75	476	0.79	364	0.61

values of Cohen's κ varied with evaluator pairs and that meaning equivalence is significantly more difficult to judge than grammaticality, even with decision tree procedure.

Our paraphrase lexicons were rated according to the proportion of examples that were assigned a label corresponding to the positive class. Table 14 summarizes the results based on individual judgments and majority voting. For the latter, an example is regarded as correct *iff* a majority of evaluators (two or three in our case) classified it into one of the positive classes. Owing to the various filters that we used, the paraphrases drawn from S_{Seed} were of substantially high quality despite the low chance of being the 3 best candidates. The paraphrases in S_{Hvst} had lower scores in both grammaticality and meaning equivalence than those in S_{Seed}. However, their precision was reasonably high, considering that no use was made of any rich language-specific resources. It is not possible to provide direct and fair comparisons with previous work because of the differences in data and human evaluators. Yet, it is worth mentioning that while using parser-oriented syntactic constraints in bilingual pivoting, Callison-Burch [8] achieved precision of no more than 0.68, 0.61, and 0.55, respectively, for grammaticality, meaning equivalence, and their combination.

Our expanded lexicon, S_{Hvst}, led to more grammatical errors and meaning differences than the seed lexicon, S_{Seed}. A manual error analysis revealed that one of the major sources of grammatical errors was the presence of alternative syntactic categories, such as those exemplified in (18).

(18) s. The safety issue was *considered sufficiently* serious for all affected parties to be informed.
t. *The safety issue was *sufficient consideration* serious for all affected parties to be informed.

Differences in grammatical number and in determiners constituted another major source of grammaticality issues.

(19) s. Federal Security Service now spread a big network of fake sites and there are tons of *potential buyers* of military weapons.

t. *Federal Security Service now spread a big network of fake sites and there are tons of _a potential buyer_ of military weapons.

Such pairs were already present in S_{Seed}, owing to incorrect translations in the bilingual corpus and/or errors introduced through word alignment. The problem was just amplified by our method. The usefulness of such morphological variants of the same word would depend on the downstream task [20]. For instance, they could be useful for paraphrase recognition tasks, including question answering and multi-document summarization. However, as they are not perfect paraphrases, substituting them in a given context can degrade grammaticality, although there are cases in which morphological variants are mutually substitutable. For instance, (“was showing”, “has shown”) are mutually substitutable in many contexts.

As for meaning errors, we found erroneous paraphrase pairs in S_{Hvst} originating from errors in S_{Seed}, such as the example in (20).

(20) s. The newspapers reported in September that he had bought the remains of the _north wing_ from the demolition contractor.
t. $^{\neq}$The newspapers reported in September that he had bought the remains of the _south wing_ from the demolition contractor.

In this case, an erroneous translation of “South America” by “Amérique du Nord” in the bilingual corpus led to the extraction of the incorrect seed paraphrase pair, (“north america”, “south america”) whose generalization then led to the extraction of the pair (“north wing”, “south wing”). The number of such errors can be reduced by using a higher threshold value for the paraphrase probability, i.e., th_p, while filtering based on contextual similarity will not discard this type of errors as described in Section 2.1.

On the other hand, a semantically equivalent pair of phrases might cause a meaning error owing to their incompatibility in the given context, as shown in (21).

(21) s. Are you not _turning your back_ on those who voted for you in doing this?
t. $^{\neq}$Are you not _to turn your back_ on those who voted for you in doing this?

Our expanded lexicon, S_{Hvst}, includes many paraphrases that were missing in the seed set, S_{Seed}. Phrases that were already covered by S_{Seed} might not need any further expansion. Effective means of using these different types of paraphrase lexicons will be addressed in our future work.

5.4 Quality of Paraphrase Lexicons in Other Settings

For the other three settings besides Europarl English, we conducted the same evaluation experiment in order to confirm their quality and to identify their characteristics. For each setting, we asked one native speaker[24] to evaluate the 40 example units (Section 5.2), following our criteria and procedure (Section 5.1). Then, we calculated precision of each paraphrase lexicon in the same manner, according to the proportion of examples that were assigned a label corresponding to the positive class (Table 10).

As presented in Table 15, the quality of paraphrase lexicons varies with the settings. In the Europarl French setting, the paraphrased sentences preserved the meaning of the original sentences at a substantially high rate, but more than half were ungrammatical. The most prominent grammatical errors were related to inflectional morphology, like the example in (19). This problem can

[24] Examples in the NTCIR Japanese and Europarl French settings were evaluated by the first and second authors, respectively.

Table 15. Precision of Paraphrase Substitution (cf. Table 14)

Setting	Lexicon	N	Grammaticality		Meaning equivalence		Both	
			#	%	#	%	#	%
Europarl French	S_{Seed}	12	4	0.33	11	0.92	4	0.33
	S_{Hvst}	108	52	0.48	92	0.85	45	0.42
	Total	120	56	0.47	103	0.86	49	0.41
NTCIR English	S_{Seed}	21	20	0.95	19	0.90	19	0.90
	S_{Hvst}	99	64	0.64	57	0.58	48	0.48
	Total	120	84	0.70	76	0.63	67	0.56
NTCIR Japanese	S_{Seed}	34	32	0.94	29	0.85	29	0.85
	S_{Hvst}	86	59	0.69	60	0.70	47	0.55
	Total	120	91	0.76	89	0.74	76	0.63

be attributed to the difference in morphological richness between the two languages: inflectional variations in French tended to be conflated during the bilingual pivoting process regarding English as the pivot language.

In the two NTCIR settings, S_{Seed} had a level of quality comparable with that in the Europarl English setting. In contrast, S_{Hvst} had visibly lower quality in terms of both grammaticality and meaning equivalence. In both languages, sentences from this domain tend to contain relatively long noun phrases, and paraphrasing only parts of them hurts grammaticality and/or alters the meaning. For instance, the phrase substitution in (22) causes an agreement error between the determiner "a" and plural noun and inconsistent reference to the elements numbered 61 and 62; while their first occurrences in the paraphrased sentence, (22t), refer to the entire wire, their second occurrences refer only to the core parts of the wire.

(22) s. Each press-contacting blade 71 cuts a sheath of the wire 61, 62, and is electrically connected to a *conductor of the wire* 61, 62.
t. Each press-contacting blade 71 cuts a sheath of the wire 61, 62, and is electrically connected to a *wire conductors* 61, 62.

Most of the grammatical errors in the NTCIR Japanese setting were related to verb conjugation and/or selection of verbal suffixes and case markers. Many of the meaning errors in this setting resulted from the presence of incorrect translations in the bilingual corpus. As in example (20), we consequently observed, for instance, ("複数␣の␣生産␣設備" (plural production facilities), "1つ␣の␣生産␣設備" (single production facility)) and ("カラム␣選択␣線" (column select lines), "行␣選択␣線" (row select lines)).

6 CONCLUSION

In this article, we have presented a paraphrase acquisition method that improves coverage while maintaining accuracy. It works by expanding a smaller paraphrase lexicon through (i) the automatic induction of paraphrase patterns from the smaller lexicon and (ii) the extraction of new paraphrase pairs through the automatic instantiation of those induced patterns from a large-scale monolingual data. To the best of our knowledge, this is the first attempt to exploit lexical variants for acquiring paraphrases in a fully empirical way. Other than raw monolingual data, our method requires only minimal language-dependent resources: tokenizers and lists of stop words. We have demonstrated its strong quantitative impact and have found the quality of the resulting paraphrase pairs to be relatively high.

Our planned future work is three-fold.

Sophistication of the method. We would like to extend our method along the following three axes.

- Paraphrase lexicons created using different methods and sources have different properties (Section 5). We intend to design an overall model that will facilitate the effective use of such heterogeneous lexicons.
- We are also interested in the potential usefulness of linguistic tools, such as POS tagger and parser, to enhance our method (Section 4.1.3).
- Finally, we plan to apply bootstrapping-based methods for relation extraction (Section 2.1) to each collected pattern as a way to collect more paraphrase patterns and paraphrase pairs.

Application to other languages. We are also interested in testing the applicability of our method to various languages. For some languages, extensions will be needed in order to handle affixation processes other than prefixation and suffixation, orthographic character alternations, such as "e" and "é" in French, and other writing systems such as those for Arabic, Hindi, and Thai.

Integration into downstream applications. Paraphrasing is a fundamental linguistic phenomenon that affects a wide range of NLP tasks. We would like to determine to what extent our paraphrase lexicons can improve the performance of such tasks as machine translation, question answering, text summarization, and text simplification.

APPENDIXES

Appendices A and B show decision trees and basic questions presented to human evaluators. They were used for evaluating grammaticality and meaning equivalence, respectively.

A GRAMMATICALITY: IS THE PARAPHRASE GRAMMATICAL?

Given a (paraphrased) sentence, answer the following questions without seeing the corresponding original sentence (see Figure 17).

Q1. Is it grammatical?

Yes. Answer Q2.
- apart from whether it is true or not: e.g., "I saw a unicorn yesterday."
- apart from whether it is nonsensical or not:
  e.g., "Colorless green ideas sleep furiously."

No. Answer Q3.

Q2. Is it perfectly grammatical or something awkward?

Perfect. Label it "Perfect."

Awkward. Label it "Awkward."
- Strange collocation:
  e.g., "Individual members are equipped with *strong computer* systems."
- Fail to form a contrast: e.g., "Eleven **men** and *three* **workers** were arrested."
- Stylistically inconsistent: e.g., "In each category, this award totals *10* ***m*** Swedish krona (approximately 25 **million** CZK)."

Q3. Are the grammatical errors correctable?

Yes. Answer Q4.

No. Label it "Irredeemable."

Q4. Are the grammatical errors corrected with only one edit, such as the following?
- Deletion of unnecessary word: e.g., "*thirty years* ***old*** **old**"
- Correction, deletion, or addition of determiner:
  e.g., "**a** ***ambitious*** *level* of advantage"
- Correction of hyphenation error:
  e.g., "The Bank of England replies to concerns by lending 10 billion pounds for ***5-weeks***."
- Correction of tense mismatch (present and past, etc.):
  e.g., "The *commission* ***report*** that BSkyB's stake **thwarted** competition and **allowed** it unfair influence over ITV."
- Correction of agreement error (subject and verb, determiner and plurality of noun, etc.):
  e.g., "The *commercial* ***results*** of the US **feeds** optimism."
- etc.

Yes. Label it "Minor Problem."

No, more than that. Label it "Major Problem."

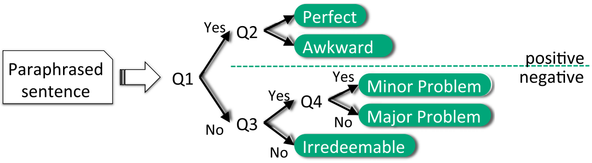

Fig. 17. Decision tree for evaluating grammaticality.

B MEANING: DOES THE PARAPHRASE PRESERVE THE MEANING OF THE ORIGINAL SENTENCE?

Given a pair of a paraphrased sentence and its original sentence, answer the following questions (see Figure 18).

Q1. Do the two phrases share some meaning in this context?

Yes. Answer Q2.

No. Label it “Completely Different.”

Q2. Does the paraphrase convey a meaning significantly different from the original sentence in this context?

Yes. Label it “Significantly Different.”

- Different: e.g., “He waited for *two years*.” ⇒ “He waited for *three years*.”
- Different:

  e.g., “Gaudi designed a *central heating* system in the house.”

  ⇒ “Gaudi designed a *first heating* system in the house.”
- Narrowing the area is critical:

  e.g., “The leaders discussed the *global economy*.”

  ⇒ “The leaders discussed the *economic issues in Europe*.”
- Broadening the area is critical:

  e.g., “The leaders discussed the *economic issues in Europe*.”

  ⇒ “The leaders discussed the *global economy*.”
- etc.

No, nothing is changed or there are only ignorable changes. Answer Q3.

Q3. Are the meaning that two sentences convey perfectly equivalent?

Yes. Label it “Equivalent.”

No. Answer Q4.

Q4. Is there a slight difference between two sentences?

Loss. Label it “Missing Info.”

e.g., “The baby boom crested *around 1957*.”

⇒ “The baby boom crested *in the late 1950s*.”

Addition. Label it “Additional Info.”

e.g., “*Twelve million* people were affected in the crash.”

⇒ “*12.00 million* people were affected in the crash.”

Something else including both loss and addition. Label it “Ignorable Change.”

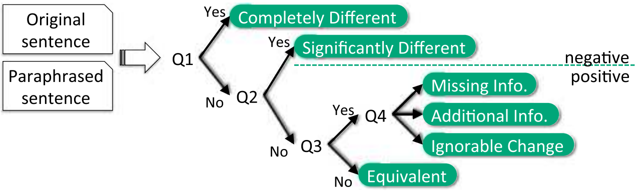

Fig. 18. Decision tree for evaluating equivalence of meaning.

ACKNOWLEDGMENTS

This work was partly supported by JSPS Postdoctoral Fellowship for Research Abroad (FYs 2011–2012) and JSPS KAKENHI Grant-in-Aid for Young Scientists, grant no. (B) 25730139.

We are deeply grateful to anonymous reviewers for their valuable comments on the earlier version of this article.

REFERENCES

[1] Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research* 38, 135–187.

[2] R. Baayen, R. Piepenbrock, and L. Gulikers. 1995. CELEX2 LDC96L14. Philadelphia: Linguistic Data Consortium.

[3] 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)*. 597–604.

[4] Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 25–32.

[5] Regina Barzilay and Lillian Lee. 2002. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 164–171.

[6] 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)*. 50–57.

[7] 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)*. 161–170.

[8] 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)*. 196–205.

[9] 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 Natural Language Semantics (GEMS)*. 33–42.

[10] 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)*. 190–200.

[11] Noam Chomsky. 1957. *Syntactic Structures*. Mouton Publishers, The Hague, The Netherlands.

[12] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1, 37–46.

[13] Stijn De Saeger, Kentaro Torisawa, Masaaki Tsuchida, Jun'ichi Kazama, Chikara Hashimoto, Ichiro Yamada, Jong Hoon Oh, István Varga, and Yulan Yan. 2011. Relation acquisition using word classes and partial patterns. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 825–835.

[14] 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)*. 85–91.

[15] 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)*. 350–356.

[16] 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)*. 420–429.

[17] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

[18] Joseph F. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5, 378–382.

[19] 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)*. 151–158.

[20] Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. 4276–4282.

[21] 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)*. 1168–1179.

[22] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 758–764.

[23] Éric Gaussier. 1999. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing*. 24–30.

[24] Maayan Geffet and Ido Dagan. 2004. Feature vector quality and distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*. 247–253.

[25] Nizar Habash and Bonnie Jean Dorr. 2003. A categorial variation database for English. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*. 96–102.

[26] Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama. 2006. Selection of effective contextual information for automatic synonym acquisition. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and the 21st International Conference on Computational Linguistics (COLING-ACL)*. 353–360.

[27] Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama. 2008. Effective use of indirect dependency for distributional similarity. *Journal of Natural Language Processing* 15, 4, 19–42.

[28] Zellig Harris. 1954. Distributional structure. *Word* 10, 23, 146–162.

[29] Zellig Harris. 1957. Co-occurrence and transformation in linguistic structure. *Language* 33, 3, 283–340.

[30] 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)*. 1087–1097.

[31] Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Stijn De Saeger, Masaki Murata, and Jun'ichi Kazama. 2009. Large-scale verb entailment acquisition from the web. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1172–1181.

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

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

[34] Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press, New York, NY.

[35] 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 Association for Computational Linguistics (HLT-NAACL)*. 48–54.

[36] 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)*. 145–153.

[37] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1, 159–174.

[38] Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*. 25–32.

[39] Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL)*. 768–774.

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

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

[42] Prodromos Malakasiotis and Ion Androutsopoulos. 2011. A generate and rank approach to sentence paraphrasing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 96–106.

[43] Yuval Marton. 2013. Distributional phrasal paraphrase generation for statistical machine translation. *ACM Transactions on Intelligent Systems and Technology* 4, 3, Article 39, 32 pages.

[44] 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)*. 237–249.

[45] 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)*. 656–666.

[46] Igor Mel'čuk and Alain Polguère. 1987. A formal lexicon in meaning-text theory (or how to do Lexica with words). *Computational Linguistics* 13, 3–4, 261–275.

[47] 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)*. 119–130.

[48] 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)*. 102–109.

[49] Patric Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and the 21st International Conference on Computational Linguistics (COLING-ACL)*. 113–120.

[50] Hideki Shima. 2015. *Paraphrase Pattern Acquisition by Diversifiable Bootstrapping*. Ph.D. Dissertation. Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

[51] 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)*.

[52] 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.

[53] 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 for Computational Linguistics (ACL)*. 456–463.

[54] Idan Szpektor, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 41–48.

[55] Kentaro Torisawa. 2002. An unsupervised learning method for associative relationships between verb phrases. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*. 1009–1015.

[56] Julie Weeds, David Weir, and Bill Keller. 2005. The distributional similarity of sub-parses. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. 7–12.

[57] 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 (EWNLG)*. 122–125.

[58] Yulan Yan, Chikara Hashimoto, Kentaro Torisawa, Takao Kawai, Jun'ichi Kazama, and Stijn De Saeger. 2013. Minimally supervised method for multilingual paraphrase extraction from definition sentences on the web. In *Proceedings of Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 63–73.

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

Received November 2016; revised October 2017; accepted October 2017