Expanding Paraphrase Lexicons by Exploiting Lexical Variants

Atsushi Fujita[†] **Pierre Isabelle**[‡]
[†]National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Souraku-gun, Kyoto, 619-0289, Japan
atsushi.fujita@nict.go.jp
[‡]National Research Council Canada
1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada
Pierre.Isabelle@nrc.ca

Abstract

This study tackles the problem of paraphrase acquisition: achieving high coverage as well as accuracy. Our method first induces paraphrase patterns from given seed paraphrases, exploiting the generality of paraphrases exhibited by pairs of lexical variants, e.g., "amendment" and "amending," in a fully empirical way. It then searches monolingual corpora for new paraphrases that match the patterns. This can extract paraphrases comprising words that are completely different from those of the given seeds. In experiments, our method expanded seed sets by factors of 42 to 206, gaining 84% to 208% more coverage than a previous method that generalizes only identical word forms. Human evaluation through a paraphrase substitution test demonstrated that the newly acquired paraphrases retained reasonable quality, given substantially high-quality seeds.

1 Introduction

One of the characteristics of human languages is that the same semantic content can be expressed using several different linguistic expressions, i.e., paraphrases. Dealing with paraphrases is an important issue in a broad range of natural language processing (NLP) tasks (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010).

To adequately and robustly deal with paraphrases, a large-scale knowledge base containing words and phrases having approximately the same meaning is indispensable. Thus, the task of automatically creating such large-scale **paraphrase lexicons** has been drawing the attention of many researchers (see Section 2 for details). The challenge is to ensure substantial coverage along with high accuracy despite the natural tension between these factors. Among the different types of language resources, monolingual corpora[1] offer the largest coverage, but the quality of the extracted candidates is generally rather low. The difficulty lies in the manner of distinguishing paraphrases from expressions that stand in different semantic relations, e.g., antonyms and sibling words, using only the statistics estimated from such corpora. In contrast, highly accurate paraphrases can be extracted from parallel or comparable corpora, but their coverage is limited owing to the limited availability of such corpora for most languages.

This study aims to improve coverage while maintaining accuracy. To that end, we propose a method that exploits the generality exhibited by pairs of **lexical variants**. Given a seed set of paraphrase pairs, our method first induces paraphrase patterns by generalizing not only identical word forms (Fujita et al., 2012) but also pairs of lexical variants. For instance, from a seed pair (1a), a pattern (1b) is acquired, where the pair of lexical variants ("amendment", "amending") and the shared word form "regulation" are generalized.

(1) a. amendment of regulation
$\Leftrightarrow$ amending regulation
b. X:ment of Y:ϕ $\Leftrightarrow$ X:ing Y:ϕ

With such patterns, new paraphrase pairs that would have been missed using only the surface forms are extracted from a monolingual corpus. Obtainable pairs can include those comprising words that are

[1]The term "monolingual corpora" in this study refers to monolingual non-parallel corpora, unless otherwise explicitly noted. As reviewed in Section 2.1.2, monolingual parallel corpora have also been used as a source of paraphrases.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 630–640,
Denver, Colorado, May 31 – June 5, 2015.

completely different from those of the seed paraphrases, e.g., (2a) and (2b).

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

While the generality underlying paraphrases has been exploited either by handcrafted rules (Harris, 1957; Mel'čuk and Polguère, 1987; Jacquemin, 1999; Fujita et al., 2007) or by data-driven techniques (Ganitkevitch et al., 2011; Fujita et al., 2012), we still lack a robust and accurate way of identifying various types of lexical variants. Our method tackles this issue using affix patterns that are also acquired from high-quality seed paraphrases in a fully empirical way. Consequently, our method has the potential to apply to many languages.

2 Previous Work

2.1 Creating Paraphrase Lexicons

Researchers have been intensively studying methods for automatically creating paraphrase lexicons using various types of corpora. There are two major streams: one that uses monolingual corpora and one that uses parallel or comparable corpora.

2.1.1 Monolingual Corpora

A monolingual corpus is the most promising resource when targeting increased coverage, thanks to the availability of Web-scale monolingual data. Techniques that use such corpora mostly extract pairs of expressions by exploiting the **contextual similarity** associated with the Distributional Hypothesis (Harris, 1954). A given expression is represented with its co-occurring expressions such as adjacent word n-grams (Paşca and Dienes, 2005; Bhagat and Ravichandran, 2008; Marton, 2013), nominal elements (Lin and Pantel, 2001; Szpektor et al., 2004; De Saeger et al., 2011), and modifiers and modified words (Hagiwara et al., 2006). The similarity of a pair of expressions is calculated by comparing the distributions of their contexts.

Despite the quantitative advantage, this approach tends to result in low accuracy. This is because contextual information alone often fails to differentiate paraphrases from expressions that have other semantic relations, e.g., antonyms and sibling words.

2.1.2 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 (2001) and Pang et al. (2003) collected multiple human translations of the same source text. Multiple verbalizations of mathematical proofs were also used (Barzilay and Lee, 2002). This triangulating method provides 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 in that way (Shinyama et al., 2002; Barzilay and Elhadad, 2003; Dolan et al., 2004; Wubben et al., 2009). Chen and Dolan (2011) created such corpora by collecting multiple descriptions of short movies through crowdsourcing. Web-harvested definition sentences of the same term often contain paraphrases (Hashimoto et al., 2011; Yan et al., 2013).

Bilingual parallel corpora have been recognized as sources of paraphrases since (Bannard and Callison-Burch, 2005). First, a translation table is created using techniques developed for statistical machine translation. Then, pairs of expressions in the same language that share the same translations are extracted. For instance, a pair ("under control", "in check") will be extracted if they are both linked with the German translation "unter controlle." Each paraphrase pair (e_1, e_2) is assigned probabilities, $p(e_2|e_1)$ and $p(e_1|e_2)$, estimated by marginalizing over all the translations F shared by e_1 and e_2, i.e., $p(e_2|e_1) = \sum_{f \in F} p(e_2|f)p(f|e_1)$.

This **bilingual pivoting** approach inspired further techniques such as the use of syntactic information as the basis of constraints (Callison-Burch, 2008; Zhao et al., 2009), learning patterns using synchronous grammar (Ganitkevitch et al., 2011), uncovering missing links by combining multiple translation tables and other lexical resources (Kok and Brockett, 2010), and re-ranking candidate pairs on the basis of contextual similarity (Chan et al., 2011). Ganitkevitch and Callison-Burch (2014) compiled paraphrase lexicons for various languages on this approach.

Parallel/comparable corpora are useful sources of highly accurate paraphrases. However, for most languages, only small paraphrase lexicons can be created due to the limited availability of such corpora.

2.1.3 Combination of Multiple Corpora

Unlike the above methods, which used only a single type of corpus as sources of paraphrases, Fujita et al. (2012) used both bilingual parallel and monolingual corpora as sources. In that method, paraphrase pairs, e.g., (3a), are first acquired from a bilingual parallel corpus using the bilingual pivoting method and several heuristic filters for drastic noise reduction. Second, each paraphrase pair is generalized into a paraphrase pattern[2], e.g., (3b). Finally, new pairs, e.g., (3c), are extracted from a monolingual corpus using the patterns.

(3) a. amendment of regulation
⇔ amending regulation
b. amendment of X ⇔ amending X
c. amendment of documents
⇔ amending documents

Using that method, they were able to expand the seed lexicon by a large multiple (15 to 40 times), and the new paraphrase pairs were of reasonably good quality. However, they introduced variables only for identical word forms shared by both sides of each pair and left corresponding pairs of lexical variants, e.g., (“amendment”, “amending”) in (3a), untouched.

2.2 Dealing with Lexical Variants

In this study, the term **lexical variants** covers, at least, the following three types of word groups.

Lexical derivations: different words that share the same stem and a large part of their meaning, e.g., {“develop”, “developer”, “development”, . . .}. Words in such a group can have different parts-of-speech.

Morphological variants: different surface forms of the same word, e.g., {“amend”, “amends”, “amending”, . . .}. These are derived based on processes such as inflection and conjugation.

Orthographic variants: different spellings of the same inflectional/conjugation form of the same word, e.g., {“color”, “colour”} and {“authorize”, “authorise”}.

Several syntactic and semantic theories, such as transformational grammar (Harris, 1957) and Meaning-Text Theory (Mel’čuk and Polguère, 1987), propose a representation of paraphrases that involve alternations of lexical variants. Jacquemin (1999) and Fujita et al. (2007) addressed this type of paraphrase using manually described syntactic transformation patterns in combination with dictionaries of lexical variants.

Catvar (Habash and Dorr, 2003) is a comprehensive lexical derivation database for English. WordNet (Fellbaum, 1998) also contains information of that kind and is currently available for various languages. Despite its high accuracy, manual creation of rich lexical resources requires a large human effort. Gaussier (1999) and Fujita et al. (2007) extracted groups of lexical derivations from a list of headwords of dictionaries through mining affix patterns. This approach significantly reduces human effort, maintaining reasonable accuracy, but the coverage is still limited because of the reliance on manually compiled dictionaries.

3 Proposed Method

This study is the first attempt to exploit various types of lexical variants for acquiring paraphrases in a completely empirical way.

Given a seed paraphrase lexicon (S_{Seed}) our method (henceforth ***LEXVAR***) expands it in two steps (see also Figure 1).

Step 1. Learning paraphrase patterns: From S_{Seed}, we learn a set of paraphrase patterns, generalizing various types of lexical variants in addition to identical word forms.

Step 2. Harvesting new paraphrase pairs: Using the learned paraphrase patterns, we harvest a set of new paraphrase pairs (S_{LV}) from monolingual corpora.

LEXVAR subsumes Fujita et al. (2012)’s method explained in Section 2.1.3 (henceforth ***IDENT***), and its output S_{LV} always subsumes *IDENT*’s output (S_{ID}). As *LEXVAR* and *IDENT* have the effect of expanding pre-existing paraphrase lexicons, they can be used as a complement to the other methods for acquiring paraphrases, provided they produce a

[2]If a constituency parser is available for the language of interest, one can learn syntax-based patterns during the bilingual pivoting process (Ganitkevitch et al., 2011).

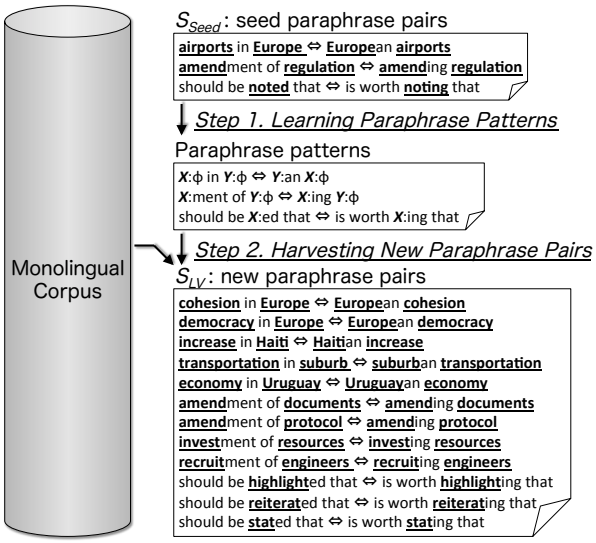

Figure 1: Overview of our proposed method.

sufficient number of high-quality pairs to make lexical generalization possible.

3.0 Step 0. Acquiring Seed Paraphrase Pairs

Our method requires as input a seed paraphrase lexicon (S_{Seed}) that has high quality and preferably exhibits various lexical correspondences that our method will exploit. For this purpose, paraphrases acquired from bilingual or monolingual parallel corpora are preferable (see Section 2.1.2).

In this study, we take the bilingual pivoting method as an example for the sake of reproducibility. However, the method also outputs a large number of non-paraphrases. To obtain further clean seeds, we apply several filters as described in (Fujita et al., 2012) and discard pairs that have low paraphrase probability, i.e., $p(e_2|e_1) < 0.01$, following the convention in (Du et al., 2010; Max, 2010; Denkowski and Lavie, 2010; Fujita et al., 2012).

Previous work (Chan et al., 2011; Fujita et al., 2012; Ganitkevitch et al., 2013) has proved that the information obtained from monolingual data can be used for assessing bilingually originated paraphrases. Thus, pairs that have low contextual similarity are also filtered out. Among various recipes for computing contextual similarity, we use a simple one: cosine measure of two context vectors comprising adjacent word 1–4 grams of all of the phrase appearances in given monolingual data. For a fair comparison with previous work, we eliminate only pairs that have no shared context, i.e., $Sim(e_1, e_2) = 0$.

3.1 Step 1. Learning Paraphrase Patterns

Given a set of seed paraphrases (S_{Seed}) we first induce a set of paraphrase patterns. From a seed paraphrase (4a), for instance, while *IDENT* learns (4b), *LEXVAR* generates (4c) by exploiting the generality exhibited by corresponding pairs of lexical variants, i.e., ("amendment", "amending").

(4) a. amendment of regulation
⇔ amending regulation
b. amendment of X ⇔ amending X
c. X:ment of Y:ϕ ⇔ X:ing Y:ϕ

The central issue at this stage is to robustly and accurately identify various types of lexical variants. We examine a data-driven approach, targeting for increased coverage, but manually created resources such as dictionaries can also be used.

3.1.1 Collecting Affix Patterns

As exemplified by ("X:ment", "X:ing") in (4c), we represent pairs of lexical variants with **affix patterns**. While Gaussier (1999) considered only suffix patterns, we also deal with prefix patterns such as those exhibited by ("reliable", "unreliable") and ("exist", "coexist") observed in the following paraphrase pairs.

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

However, we currently do not consider prefix/suffix combinations, such as ("directly", "indirect") and ("believed", "unbelievable"), and other types of affixes than prefixes and suffixes.

Reliable affix patterns are collected from S_{Seed} (cf., headwords of manually compiled dictionaries (Gaussier, 1999; Fujita et al., 2007)). First, candidates of affix patterns are extracted from S_{Seed} on the following assumption.

> A pair of words will share a definite semantic relation if the words appear on opposite sides of a paraphrase pair and have the same stem.

We do not rely on any language resources to identify the stems of words. Instead, we regard word pairs that share at least one character as candidate

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 1: Candidate pairs of lexical variants and corresponding affix patterns extracted from (6).

Affix_1	Affix_2	# of unique stems length $\geq$5	# of unique stems length $<$5	Result
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

Table 2: Filtering affix patterns (# of unique stems taken from our experimental result of Europarl setting).

pairs of lexical variants and extract the longest common prefix/suffix as their corresponding affix patterns. From a paraphrase pair (6), for instance, we separately extract four pairs of words and their corresponding affix patterns, as shown in Table 1.

(6) is aimed at achieving $\Leftrightarrow$ aims to achieve

Our candidate affix patterns are then filtered using the following criterion (Gaussier, 1999).

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

This criterion relies on two parameters, n and k. The parameter n assesses whether a pattern is sufficiently productive. The other (k) is more linguistically motivated: a genuine pattern is more likely to be used for long stems, as affixation is a general operation for producing lexical derivations in many languages. In particular, we set $k = 5$ and $n = 2$, as proposed in (Gaussier, 1999). Table 2 presents examples of filtering affix patterns eliminated and retained with this setting.

3.1.2 Generating Paraphrase Patterns

Using the affix patterns acquired in the previous step, paraphrase patterns are generated from the seed paraphrase pairs in S_{Seed}. In this step, we exhaustively consider all the combinations of word forms and lexical variants that match one of the affix patterns. From the paraphrase pair (6), the following pattern is generated.

(7) is X:ed at Y:ing $\Leftrightarrow$ X:s to Y:e

Thanks to the above filtering mechanism, spurious patterns, such as (8), are not generated.

(8) is X:imed at Y:chieving
$\Leftrightarrow$ Y:ims to X:chieve

3.2 Step 2. Harvesting New Paraphrase Pairs

Given a set of paraphrase patterns, e.g., (4c) and (7), new paraphrase pairs are harvested from monolingual corpora. In this process, each paraphrase pattern is used as a template such that the expressions that match both sides of the patterns are collected.

Unlike *IDENT*'s patterns, e.g., (4b), *LEXVAR* also collects corresponding pairs of lexical variants designated by each pattern. However, affix pattern alone cannot guarantee the semantic relation between a corresponding pair of words that each paraphrase pattern implicitly requires. For instance, the pattern (9b) is learned from (9a), where a definite relation is assumed between the two elements of ("X:ϕ", "X:an").

(9) a. countries of Europe
$\Leftrightarrow$ European nations
b. countries of X:ϕ $\Leftrightarrow$ X:an nations

Word pairs inappropriate for this pattern, e.g., ("uncle", "unclean") and ("beg", "began"), would be extracted alongside appropriate ones, e.g., ("Haiti", "Haitian") and ("suburb", "suburban"). Nonetheless, we suppose that the other surface parts of each paraphrase pair, e.g., "countries of" and "nations" in (9b), can effectively constrain instances, guaranteeing the existence of each entire phrase of the pair.

Pattern matching alone can generate pairs that are not suitable as paraphrases in any context. Thus, we assess the reliability of each pair by calculating contextual similarity between two phrases in the same manner as cleaning S_{Seed}: a pair of phrases is eliminated, if the phrases are used in completely dissimilar contexts.

3.3 Limitation

While *LEXVAR* exploits a kind of generality of paraphrases exhibited by pairs of lexical variants, it does not exploit paraphrase pairs comprising completely different surface forms such as those pairs in (10).

(10) a. look like $\Leftrightarrow$ resemble
b. burst into tears $\Leftrightarrow$ cry

To create further large paraphrase lexicons, we need to acquire these idiosyncratic paraphrases by improving existing methods and/or exploring yet another approach.

Another limitation of *LEXVAR* is that it considers only prefixes and suffixes of words as clues of lexical correspondences. We will need extensions to deal with a wider range of lexical correspondences. For instance, depending on the targeted language, other types of affixes, such as infixes and circumfixes, should be taken into account. Gaussier (1999) pointed out that some lexical derivations involve character-level alternations, e.g., "c" and "ç." Fujita et al. (2007) demonstrated that lexical derivations in an ideographic language, i.e., Japanese, can be captured by considering both ideographs and their phonetic transcriptions.

Last but not least, as *LEXVAR* regards only corpus as source, it does not acquire paraphrases that do not appear in a given corpus.

4 Expanding Paraphrase Lexicons

To what extent can our *LEXVAR* method expand a given paraphrase lexicon? We examined this, taking English as a target language and the bilingual pivoting method as the means of acquiring S_{Seed}.

4.1 Seed Paraphrase Pairs

We conducted experiments on the following two corpora configurations.

Europarl setting: The English–French version of the Europarl Parallel Corpus[3] comprising 2.0 M sentence pairs (55.7 M words in English and 61.9 M words in French) was used as a bilingual corpus. Its English side and the 2011–2013 editions of News Crawl corpora[4] comprising 52.0 M sentences (1.20 B words) were used as a monolingual corpus.

NTCIR setting: The Japanese–English Patent Translation data[5] comprising 3.2 M sentence pairs (107 M words in English and 116 M morphemes in Japanese) was used as a bilingual parallel corpus, while its English side and the 39.9 M sentences (1.36 B words) from the 2006–2007 chapters of NTCIR unaligned patent documents were used as a monolingual corpus.

For learning curve experiments, several sizes of bilingual sub-corpora were created by sub-sampling sentence pairs for both settings.

The other language resources involved in this experiment are as follows.

Phrase table learner: SyMGIZA++[6] was used for IBM2 alignment, then grow-diag-final phrase extraction and phrase table pruning were performed using toolkits in Moses[7].

Tokenizer: The tokenizer distributed with Moses was used for both English and French texts. For Japanese data, MeCab[8] was used.

Stoplists: To perform several types of filtering proposed by Fujita et al. (2012), we used the stoplists available on the Web[9]: 571 English and 463 French words. For Japanese, we manually listed 160 morphemes.

4.2 Paraphrase Patterns

Paraphrase patterns were learned from the set of seed paraphrase pairs. Figure 2 shows the numbers of the acquired paraphrase patterns and the percentages of paraphrase pairs in the seed lexicon, S_{Seed}, covered by the patterns.

As illustrated by example (4), *LEXVAR* learns more general paraphrase patterns than *IDENT*. Applied to another seed paraphrase pair (11a), *IDENT* will generate another pattern (11b), but *LEXVAR* will not: the corresponding (4c) is already learned.

(11) a. development of tourism
$\Leftrightarrow$ developing tourism
b. development of $X \Leftrightarrow$ developing X

On the other hand, *LEXVAR* also learns patterns from seed paraphrase pairs that *IDENT* ignores, e.g., (6) and (9a). Consequently, a wider range of seed paraphrases were involved in learning patterns and more patterns were acquired.

4.3 New Paraphrase Pairs

Finally, new paraphrases were acquired from the monolingual data. At this time, only single words

[3] http://statmt.org/europarl/, release 7
[4] http://statmt.org/wmt14/translation-task.html
[5] http://ntcir.nii.ac.jp/PatentMT-2/
[6] http://psi.amu.edu.pl/en/index.php?title=SyMGIZA
[7] http://statmt.org/moses/, RELEASE-2.1.1
[8] https://code.google.com/p/mecab/, version 0.996
[9] http://members.unine.ch/jacques.savoy/clef/

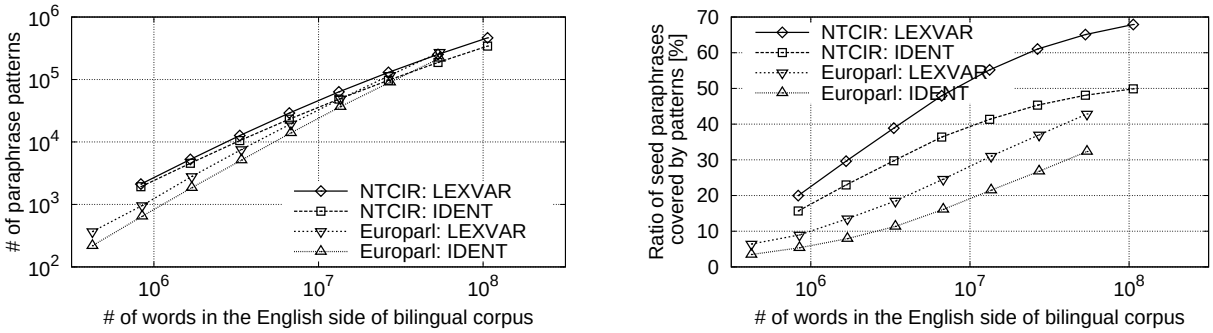

Figure 2: Statistics for the acquired paraphrase patterns: number and coverage against S_{Seed}.

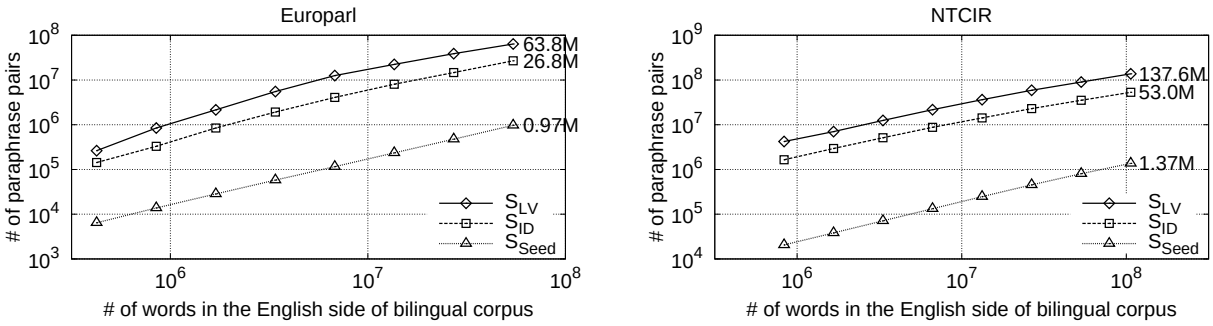

Figure 3: Number of acquired paraphrase pairs (left: Europarl, right: NTCIR).

were regarded as potential slot-fillers for the patterns. Recall that S_{LV} and S_{ID} are the sets of paraphrases generated by *LEXVAR* and *IDENT*, respectively, and $S_{LV} \supseteq S_{ID}$. Pairs that appeared in S_{Seed} and those used in completely dissimilar contexts were excluded from both S_{ID} and S_{LV}.

Figure 3 demonstrates that, irrespective of the size of the bilingual corpus, *LEXVAR* yielded far more (relative) coverage of paraphrase pairs S_{LV} than not only S_{Seed} but also S_{ID}. When the full bilingual corpora were used, S_{LV} contained 63.8 M and 137.6 M paraphrase pairs in the two respective settings, while S_{ID} contained only 26.8 M and 53.0 M pairs. The seed set S_{Seed} can be pooled with S_{LV}; thus, *LEXVAR* expanded S_{Seed} by approximately 67 and 101 times in the two respective settings. Figure 4 illustrates the ratio of the expanded parts of the paraphrase lexicons S_{LV} and S_{ID} against the seed set S_{Seed}. The ratio of S_{LV} against S_{Seed} ranged over 41–109 and 100–205 in the two respective settings. This figure also emphasizes the visible advantage of S_{LV} over S_{ID}: 84%–208% and 139%–159% more coverage.

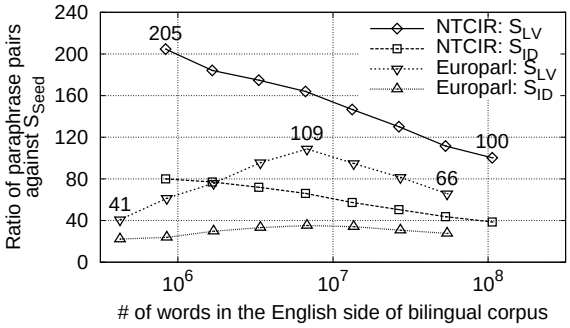

Figure 4: Ratio of S_{LV} and S_{ID} to S_{Seed}.

We expected that the more the bilingual data there are, the lower the leverage ratio is, because when a larger bilingual corpus is used, more seed paraphrases can be acquired, and the relative size of the monolingual data compared to the bilingual is lower. While the leverage ratio in the NTCIR setting follows this, the ratio in the Europarl setting does not: it peaks at approximately the middle of the scale. We found that from a very small bilingual corpus, we do not necessarily obtain seed paraphrases that exhibit

the generality exploited by *LEXVAR* and *IDENT*. In this case, the leverage ratio cannot be extremely high despite the large difference in the corpora sizes.

LEXVAR also largely contributed to discovering paraphrases for phrases that were not paraphrased using only S_{Seed} and S_{ID}. The ratio of the numbers of unique left-hand side phrases in S_{LV} to those in S_{Seed} ranged over 65–147 and 92–415 in the two respective settings, gaining 76%–210% and 145%–175% more coverage than S_{ID}.

5 Quality Assessment

The quality of the created paraphrase lexicons was manually evaluated through a paraphrase substitution test: we generated pairs of paraphrase sentences using the paraphrase lexicons and asked human evaluators to assess their quality.

5.1 Criteria and Procedure

Generating paraphrased sentences by substituting words and phrases involves two different tasks: generating new sentences and ensuring that the meaning is preserved. It is therefore straightforward to separately evaluate the **grammaticality** and **meaning equivalence** of each paraphrased sentence (Callison-Burch, 2008).

Grammaticality: whether the paraphrased sentence is grammatical

Meaning equivalence: whether the meaning of the original sentence is properly preserved by the paraphrased sentence

We adopted the detailed criteria and procedure described in (Fujita, 2013), as they resulted in a reasonably high inter-evaluator agreement ratio. The evaluation protocol is characterized by the following three features introduced for reducing human labor and making results consistent.

Unit-wise: Several paraphrase examples for the same source are packaged into an **example unit** and provided at the same time.

Two-phased: Evaluators are first asked to assess only the grammaticality of each paraphrased sentence without seeing the original sentence. Then, by comparing each pair of original and paraphrased sentences, they assess to what extent the paraphrased sentence retains the meaning of its counterpart.

Classification-based: Evaluators are asked to classify each example into one of the predetermined categories, guided by the decision trees respectively designed for evaluating grammaticality and meaning equivalence.

5.2 Data

We used news sentences as in (Callison-Burch, 2008; Fujita et al., 2012): the English sentences from WMT 2011–2013 "newstest" data (9,000 unique sentences). To reduce the human labor for the evaluation, they were restricted to those with moderate length: 10–30 words, which we expected to provide sufficient but succinct context of the substituted phrases. 5,850 sentences were retained.

By substituting phrases in the above sentences using the paraphrase lexicons S_{Seed} and S_{LV} in the Europarl setting, 88,555 example units comprising 1,013,511 paraphrases were generated. For each example unit, 3-best paraphrases were then selected by a 5-gram language model trained on the monolingual data in the Europarl setting with modified Kneser–Ney smoothing using KenLM[10]. Finally, from 31,149 units that contained at least three paraphrases, we randomly sampled 200 example units for 200 unique left-hand side phrases.

5.3 Results

We collected evaluations from three native English speakers. Table 3 summarizes the inter-evaluator agreement ratio, Cohen's κ (Cohen, 1960). The values for a coarse-grained binary decision[11] were "substantial" for grammaticality and "moderate" for meaning equivalence (Landis and Koch, 1977).

The quality of the examined paraphrase lexicons is measured by the precision of the evaluated examples: an example was regarded as correct if and only if a majority of evaluators (two or three in our case) assigned a label corresponding to the positive class in the binary decision. Table 4 summarizes the results. Despite the low chance of being the 3-best candidates, thanks to various filters,

[10] https://kheafield.com/code/kenlm/

[11] We regarded "Perfect" and "Awkward" for grammaticality, and "Equivalent" and either of three categories of slight differences "Missing Info.," "Additional Info.," and "Ignorable Change" for meaning equivalence as positive. This is consistent with (Callison-Burch, 2008).

Criterion	Fine-grained	Coarse-grained
Grammar	0.51 - 0.56	0.64 - 0.79
Meaning	0.27 - 0.35	0.48 - 0.53

Table 3: Cohen's κ of pairwise agreement.

Lexicon	n	Grammar	Meaning	Both
S_{Seed}	66	0.85	0.91	0.76
S_{ID} ($\subseteq S_{LV}$)	339	0.84	0.78	0.66
S_{LV}	534	0.74	0.78	0.59
Total	600	0.75	0.79	0.61

Table 4: Precision of paraphrase substitution.

paraphrases drawn from S_{Seed} were of substantially high quality. Compared to S_{Seed}, paraphrases sampled from S_{LV} have relatively low precision in both grammaticality and meaning equivalence. However, these scores are reasonably high, considering that no use is made of rich language-specific resources[12].

However, more grammatical errors occurred than with S_{Seed} and S_{ID}. A manual error analysis revealed that the majority of these errors were caused by the differences of syntactic categories between phrases, e.g., (12).

(12) The safety issue was *considered sufficiently* ($\Rightarrow$ *sufficient consideration*) serious for all affected parties to be informed.

Differences of grammatical number and determiners were the other major error sources.

(13) Federal Security Service now spread a big network of fake sites and there are tons of *potential buyers* ($\Rightarrow$ *a potential buyer*) of military weapons.

These types of pairs originally existed in S_{Seed} but were amplified by *LEXVAR*. Ganitkevitch and Callison-Burch (2014) stated that morphological variants of the same word might be desirable depending on the downstream task. For instance, they could be useful for paraphrase recognition tasks including question answering and multi-document summarization. As they are morphological variants rather than genuine paraphrases, substituting them in a given context often degrades grammaticality.

[12] Although we cannot make a direct comparison owing to the differences of data and human evaluators, for reference, Callison-Burch (2008) achieved 0.68, 0.61, and 0.55 precision for grammaticality, meaning equivalence, and both, respectively, by introducing parser-oriented syntactic constraints in bilingual pivoting.

6 Conclusion

We proposed a method for expanding given paraphrase lexicons by first inducing paraphrase patterns and then searching monolingual corpora with these patterns for new paraphrase pairs. To the best of our knowledge, this is the first attempt to exploit various types of lexical variants for acquiring paraphrases in a completely empirical way. Our method requires minimal language-dependent resources, i.e., stoplists and tokenizers, other than raw corpora. We demonstrated the quantitative impact of our method and confirmed the potential quality of the expanded paraphrase lexicon.

Our future work is four-fold. (i) Paraphrase lexicons created by different methods and sources have different properties. Designing an overall model to harmonize such heterogeneous lexicons is an important issue. (ii) We aim to investigate an extensive collection of corpora: there are far more corpora than those we used in this experiment. We are also interested in expanding paraphrase lexicons created by a method other than bilingual pivoting; for instance, those extracted from a Web-harvested monolingual comparable corpus (Hashimoto et al., 2011; Yan et al., 2013). (iii) We will apply our method to various languages for demonstrating its applicability, extending it for a wider range of lexical variants depending on the targeted language. (iv) Paraphrases are the fundamental linguistic phenomena that affect a wide range of NLP tasks. We are therefore interested in determining to what extent our paraphrase lexicons can improve the performance of application tasks such as machine translation, text summarization, and text simplification.

Acknowledgments

We are deeply grateful to Eiichiro Sumita, Masao Utiyama, Taro Watanabe, Kentaro Torisawa, and anonymous reviewers for their valuable comments on the earlier version of this paper. This work was partly supported by JSPS Postdoctoral Fellowship for Research Abroad (FYs 2011–2012) and JSPS KAKENHI Grant-in-Aid for Young Scientists (B) 25730139.

References

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

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 597–604.

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

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

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*), pages 25–32.

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

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 196–205.

Tsz Ping Chan, Chris Callison-Burch, and Benjamin Van Durme. 2011. Reranking bilingually extracted paraphrases using monolingual distributional similarity. In *Proceedings of the Workshop on Geometrical Models of Natual Language Semantics* (*GEMS*), pages 33–42.

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

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

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*), pages 825–835.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the 5th Workshop on Statistical Machine Translation* (*WMT*) *and MetricsMATR*, pages 339–342.

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

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

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

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

Atsushi Fujita, Pierre Isabelle, and Roland Kuhn. 2012. Enlarging paraphrase collections through generalization and instantiation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL*), pages 631–642.

Atsushi Fujita. 2013. A consideration on the methodology for evaluating large-scale paraphrase lexicons. In *Information Processing Society of Japan SIG Notes, NL-214-21*, pages 1–8.

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

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*), pages 758–764.

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

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

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*), pages 96–102.

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*), pages 353–360.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Zellig Harris. 1957. Co-occurrence and transformation in linguistic structure. *Language*, 33(3):283–340.

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

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

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

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

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

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

Yuval Marton. 2013. Distributional phrasal paraphrase generation for statistical machine translation. *ACM Transactions on Intelligent Systems and Technology*, 4(3).

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

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.

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

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

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

Idan Szpektor, 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*), pages 41–48.

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*), pages 122–125.

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*), pages 63–73.

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