語彙的対応関係の一般化に基づく言い換え知識の拡張

藤田 篤　　Pierre Isabelle
情報通信研究機構　National Research Council Canada

1　はじめに

言い換えを頑健かつ精度よく自動生成するには，同義関係にある語や句に関する大規模な知識 (言い換え知識) が不可欠である (cf. 深層学習による言い換え認識 [17])．これまでに，そのような資源の自動構築に関する研究が盛んに行われてきた．しかし依然として，精度とカバレージの両立という課題が残っている．

本稿では，従来手法で自動的に獲得された言い換え知識を拡張する手法を提案する．提案手法では，まず，言い換え知識中の個々の言い換え表現対における語彙的な対応関係に着目して言い換えパターンを獲得する．例えば (1a) から，両辺に共通の語 “regulation” および派生関係にある語の対 (“amendment”, “amending”) を一般化して，(1b) のパターンを得る．

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

そして，獲得した言い換えパターンを用いて，単言語コーパスから新たな言い換え表現対を収集する．この方法により，例えば (2) に示すような，元々の対とは表層的にまったく異なる語で構成される対も得られる．

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

(1b) のようなパターンによって言い換えの一般性を表す既存の試み [11, 15, 13, 5] では，パターンの記述に人手を要し，また変数に合致する様々な関係の語対を頑健に捉えることもできていない．これに対して提案手法は，言い換えおよび語の対応関係のパターンを与えられた言い換え知識から，対応関係にある語対を単言語コーパスから，大規模かつ経験的に獲得する．

2　先行研究

2.1　言い換え知識の自動獲得

既存の (経験的な) 言い換え知識獲得手法は，単言語コーパスに基づく手法，単言語パラレル/コンパラブルコーパスに基づく手法，異言語パラレルコーパス (対訳コーパス) に基づく手法，の 3 種類に大別できる．

単言語コーパスに基づく多くの手法 (e.g., [14]) は，分布仮説 [10] に基づいて，使用される文脈が類似する (文脈類似度が高い) 表現の対を言い換え表現対として獲得する．単言語コーパスは，カバレージの面で最も有望な知識源である．しかし，反義関係や上位-下位関係など，同義以外の関係を持つ表現の対も高い文脈類似度を持ちうるため，この手法の精度は概して低い．

単言語パラレル/コンパラブルコーパスがあれば，対応する文の対における同義の部分を同定することによって，精度よく言い換え表現対を獲得できる．これまでに，同じ文書に対する複数の人間訳 [2]，同じ事柄について述べている複数の新聞社の記事 [16]，同じ概念/用語に対する複数の定義文 [12] などが用いられてきた．しかし，この種のコーパスからは，パラレル/コンパラブルでないコーパスほどのカバレージは得られない．

対訳コーパスから翻訳テーブルを学習し，異なる言語において共通の訳を持つ表現を言い換えとして獲得できる [1]．個々の言い換え表現対の尤度も，$p(e_2|e_1) = \sum_{f \in F} p(e_2|f)p(f|e_1)$ という確率で表せる．ここで，F は e_1 と e_2 に共通の訳の集合である．対訳データは，日々大量に生産され，また機械翻訳のコミュニティにおいて大規模に蓄積されているため，この手法は同コミュニティにおける標準的な言い換え知識獲得手法として認識されている．しかし，いかに大きな対訳コーパスであっても単言語コーパスに比べると極めて小さく，潜在的な言い換え知識のカバレージも低い．

2.2　関係のある語群の自動処理

本稿では，次の 3 種類の語群に着目する．

派生語: 表記および意味の一部を共有する異なる語の群．例: {“develop”, “developer”, “development”, . . .}．この群の語は品詞が異なる場合もある．

活用形/屈折形: 活用や屈折に由来する同じ語の異なる出現形．例: {“amend”, “amends”, “amending”, . . .}．

異表記: 同じ語の同じ活用形/屈折形の異なる表記．例: {“color”, “colour”}，{“authorize”, “authorise”}．

英語については，派生語群を包括的に収録した資源 Catvar [9] が存在する．WordNet [4] もこのような情報を収録しており，英語以外の言語の資源も存在する．これらの言語資源は高品質ではあるが，構築・保守にかかる人的なコストは無視できない．文献 [8, 5] は，辞書の見出し語のリストと品詞情報を用い，語の接辞に着目して派生語群を抽出した．このアプローチも一定の精度を達成しうるが，人間によって編纂された辞書を前提としており，やはりカバレージに限界がある．

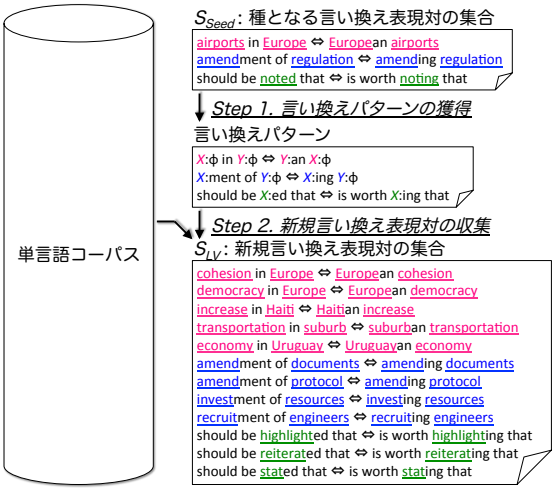

図 1: 提案手法の概要.

3 提案手法

与えられた言い換え表現対の集合を，単言語コーパスを用いて図 1 に示す手順で拡張する．

3.1 言い換えパターンの獲得

まず，与えられた言い換え表現対の集合 (S_{Seed}) から，(1b) のような言い換えパターンの集合を獲得する．

(“amendment”, “amending”) のような対応関係のある語対を，(“X:ment”, “X:ing”) のような接辞の対応関係のパターンで表す [8, 5]．現時点では，上の例のような語末の対応関係と (“reliable”, “unreliable”) などに見られる語頭の対応関係を扱うが，(“directly”, “indirect”) などの語頭と語末の両方が異なる語対は扱っていない．

高品質な S_{Seed} を前提とし (cf. 語のリスト [8, 5])，次の仮定に基づいて接辞パターンの候補を獲得する．

> 言い換え表現対の各辺にあり，同じ語幹を持つ語の対は，特定の (意味的な) 関係を持つ．

ただし，個々の語の語幹を厳密には同定しない．代わりに，1 文字以上を共有する (内容) 語の対をすべて抽出し，そこから接辞パターンの候補を生成する．例えば，(3) の言い換え表現対から表 1 の接辞パターンを得る．

(3) is aimed at achieving ⇔ aims to achieve

品質保持のため，抽出された接辞パターンの候補を次の基準 [8] に基づいてフィルタリングする．

> 長さ k 以上の語幹 n 種類以上に対して観察された接辞パターンのみを残す．

n は当該接辞パターンの生産性を表す．一方 k は，接辞派生は多くの言語における一般的な現象であり，真の接辞パターンならば，長い語 (語幹) に対しても適用される (観察される) ということを表す．文献 [8] にならい，我々も $k = 5$，$n = 2$ とする．この条件で表 1 の接辞パターンをフィルタリングした結果を表 2 に示す．

表 1: 例 (3) から抽出した語対と接辞パターンの候補.

語 1	語 2	接辞 1	接辞 2	語幹
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

表 2: 表 1 の接辞パターンのフィルタリング結果: 接辞 1 と接辞 2 の順序は正規化済．語幹の種類数については実験結果からの引用.

接辞 1	接辞 2	語幹の種類数 長さ ≥5	語幹の種類数 長さ <5	結果
X:chieve	X:imed	0	1	捨てる
X:chieving	X:ims	0	1	捨てる
X:ed	X:s	69	22	残す
X:ing	X:e	330	70	残す

最後に，上で得た接辞パターンを用いて，S_{Seed} から言い換えパターンを生成する．例えば，(3) の言い換え表現対から，次のパターンを得る．

(4) is X:ed at Y:ing ⇔ X:s to Y:e

なお，接辞パターンに合致する語対の組み合わせをすべて考慮する．

3.2 新規言い換え表現対の収集

次に，(1b) や (4) のような言い換えパターンを用いて，単言語コーパスから新たな言い換え表現対の集合 S_{LV} を収集する．ここでは，各パターンの両辺に合致する表現のみを収集する．

接辞パターンだけでは対応関係にある語対の意味的関係は保証できない．例えば，(5a) から得た言い換えパターン (5b) において，(“X:ϕ”, “X:an”) は特定の (意味的な) 関係を仮定している．

(5) a. people of Europe ⇔ European population
b. people of X:ϕ ⇔ X:an population

単言語コーパスからは，同じ関係を持つ (“Haiti”, “Haitian”) や (“suburb”, “suburban”) などだけでなく，(“uncle”, “unclean”) や (“Shakespeare”, “Shakespearean”) なども抽出されうる．ただし，“people of” や “population” など，言い換えパターンにおける他の語が適切な語対のみを得るための制約になると期待できる．

最後に，新たに獲得した個々の言い換え表現対について，文脈類似度を単言語コーパスを用いて計算し，置換可能な文脈を全く持たない対を除外する[1]．

4 言い換え知識の拡張実験

次の 2 種類の設定で，文献 [1] の手法で対訳コーパスから英語の言い換え表現対を獲得し，それを提案手法によって拡張する実験を行った．

[1] コサイン類似度や Jaccard 係数など，比較する表現間に共通する文脈を参照する計算式を用いる限り，同じ結果が得られる．

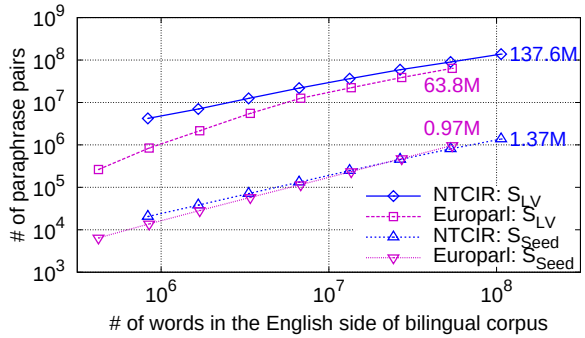

図 2: S_{LV} と S_{Seed} 中の言い換え表現対の数.

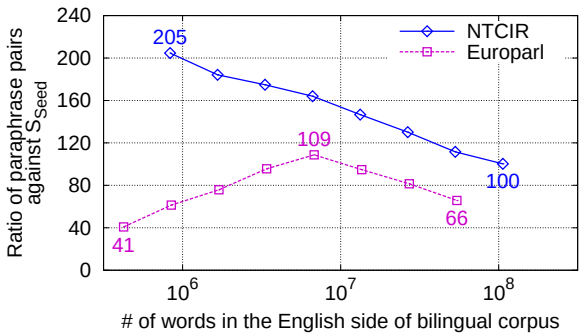

図 3: S_{LV} と S_{Seed} 中の言い換え表現対の数の比.

Europarl: Europarl コーパス[2]の英仏対 200 万文 (英語 5,570 万語，仏語 6,190 万語) を対訳コーパスとして用いた．単言語コーパスとしては，上記コーパスの英語側と WMT の News Crawl コーパス[3]の 2011-2013 年分 (5,200 万文，12.0 億語) を用いた．

NTCIR: 特許分野の日英対訳コーパス[4] (320 万文，英語 1.07 億語，日本語 1.16 億形態素) を用いた．単言語コーパスとしては，上記コーパスの英語側と NTCIR の単言語文書 2006-2007 年分 (3,990 万文，13.6 億語) を用いた．

4.1 種となる言い換え表現対の獲得

次の言語資源およびツールを用いて，対訳コーパスから，種となる言い換え表現対の集合 S_{Seed} を得た．

トーカナイザ: 英語および仏語については Moses[5]中のツールを，日本語については MeCab[6]を用いた．

統計的機械翻訳ツール: SyMGIZA++[7]を用いて IBM モデル 2 の単語アラインメントを行い，Moses 中のツールを用いて grow-diag-final ヒューリスティックスに従い翻訳フレーズ対を抽出した．

ストップリスト: 翻訳フレーズ対および言い換え候補表現対に対する種々のフィルタリングに使用した．英語および仏語については，CLEF プロジェクトの 571 語，463 語のリスト[8]を用いた．日本語については，160 種類の形態素出現形を人手で列挙した．

S_{Seed} からは，対訳に基づく言い換え確率が低い言い換え表現対 ($p(e_2|e_1) < 0.01$) を除外した．また，S_{LV} に対する処理と同様に，単言語コーパスにおいて置換可能な文脈を全く持たない対も除外した．

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

4.2 拡張結果

S_{Seed} および新たに獲得した S_{LV} の言い換え表現対の数を図 2 に，それらの数の比を図 3 に示す．対訳コーパスの規模によらず，S_{Seed} よりも顕著に多くの言い換え表現対が新たに獲得できたこと分かる．すべてのデータを用いた場合，各ドメインで 6,380 万対，1 億 3,760 万対，すなわち S_{Seed} の約 66 倍，約 100 倍の新たな対が得られた．

対訳コーパスが大きいほど，より大きな S_{Seed} が得られ，単言語コーパスは相対的に小さくなる．そのため我々は，言い換え表現対の数の比は単調減少すると予測していた．NTCIR の設定ではこの予測通りの結果が得られたが，Europarl の設定では対訳コーパスの規模が中程度の場合に比が最大となった．対訳コーパスが極端に小さく，提案手法が着目しているような言い換えの一般性が十分に観察できない場合には，提案手法の効果が制限されることが明らかになった．

5 言い換え生成を通じての質の評価

自動獲得した言い換え知識を用いて言い換え文を生成し，その言い換え文の適否を人間が評価することによって，言い換え知識の質を間接的に評価した．

5.1 評価方法と設定

自動生成した言い換え文が文法的か否か，および言い換え文が原文と同じ意味を有するか否か，の 2 点 [3] を評価した．具体的には，人間の評価者から一貫した評価結果を得るために，同じ原文から生成された複数の言い換え文を横並びに見ながら，最初に文法性を，その後で意味の等価性を，各々決定木に従って評価する，という方法 [6] を採用した．

評価用の言い換え文は，Europarl の設定で得た言い換え知識と単言語コーパス，および新聞記事を用いて作成した．前節で得た S_{Seed} および S_{LV} を用いた語句の単純な置換によって，WMT 2011-2013 の評価用データ “newstest” 中の 10-30 語の長さの文 (5,850 文)

表 3: 評価者各対の Cohen の κ.

評価基準	粗い分類	細かい分類
文法性	0.64 - 0.79	0.51 - 0.56
意味の等価性	0.48 - 0.53	0.27 - 0.35

表 4: 自動生成した言い換え文の精度.

	文数	文法性	意味の等価性	両方
S_{Seed}	66	0.85	0.91	0.76
S_{LV}	534	0.74	0.78	0.59
合計	600	0.75	0.79	0.61

から，88,555 箇所に対する 1,013,511 件の言い換え文を得た．各言い換え箇所について，単言語コーパスから改良 Kneser-Ney スムージングで推定した 5-gram 言語モデルに照らして，スコアが最も高い言い換え文 3 文を選択し，最後に，無作為に，ただし重複のないように，語句 200 件 (言い換え文数 600) を選択した．

5.2 評価結果

3 名の英語母語話者が独立に評価を行った．各評価者対の評価の一致率 (κ 値) は表 3 に示す通りである．粗い分類 [6] の一致の程度は，文法性については “substantial”，意味の等価性については “moderate” であった．

粗い分類に基づく多数派の評価を個々の言い換え文の評価とした．このときの言い換え文の精度を表 4 に示す．S_{Seed} 中の言い換え表現対は，上位 3 位に入るような言い換え文を生成できる可能性が低いものの，十分に高い精度を達成した．一方，S_{LV} の精度は，文法性，意味の等価性，それら両方，の 3 つの評価基準のいずれにおいても S_{Seed} よりも低かった．ただし，提案手法は，言語依存の高価な言語資源をほとんど使用することなく，構文解析器などを用いた従来手法と同等以上[9]の精度を達成している．

S_{LV} を用いて生成した言い換え文中の誤りの多くは，(6) に示すような，文法カテゴリの変化が原因であった．

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

その他にも，(7) のような，数や冠詞の違いが原因で誤りとなる言い換え文が少数見られた．

(7) . . . there are tons of *potential buyers* ($\Rightarrow$ *a potential buyer*) of military weapons.

同じ言い換えパターンで表される言い換え表現対が S_{Seed} に含まれていたことが元々の原因であるとはいえ，このような表現対は，厳密には言い換えではない．ただし，質問応答や複数文書要約などにおける言い換え認識には有用であると考えられる [7].

[9]データおよび評価者が異なるため優劣の議論には資さないが，文献 [3] では，CCG 解析器を用いて，文法性，意味の等価性，それら両方について，各々0.68，0.61，0.55 という精度を得ている．

6 おわりに

本稿では，従来手法で自動的に獲得された言い換え知識を，言い換え表現対に見られる語彙的対応関係に着目し，単言語コーパスを用いて拡張する手法を提案した．そして，この手法が高いカバレージと，許容可能なレベルの精度を達成しうることを示した．

今後は，英語以外の言語の，あるいは他の手法で獲得された言い換え知識についても同様の実験を実施し，提案手法の有用性を検証する予定である．

参考文献

[1] C. Bannard and C. Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL*, pp. 597–604, 2005.

[2] R. Barzilay and K. R. McKeown. Extracting paraphrases from a parallel corpus. In *Proc. of ACL*, pp. 50–57, 2001.

[3] C. Callison-Burch. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proc. of EMNLP*, pp. 196–205, 2008.

[4] C. Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[5] A. Fujita, S. Kato, N. Kato, and S. Sato. A compositional approach toward dynamic phrasal thesaurus. In *Proc. of WTEP*, pp. 151–158, 2007.

[6] A. Fujita. A consideration on the methodology for evaluating large-scale paraphrase lexicons. In *IPSJ SIG Notes, NL-214-21*, pp. 1–8, 2013.

[7] J. Ganitkevitch and C. Callison-Burch. The multilingual paraphrase database. In *Proc. of LREC*, pp. 4276–4282, 2014.

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

[9] N. Habash and B. J. Dorr. A categorial variation database for English. In *Proc. of HLT-NAACL*, pp. 96–102, 2003.

[10] Z. Harris. Distributional structure. *Word*, Vol. 10, No. 23, pp. 146–162, 1954.

[11] Z. Harris. Co-occurrence and transformation in linguistic structure. *Language*, Vol. 33, No. 3, pp. 283–340, 1957.

[12] C. Hashimoto, K. Torisawa, S. De Saeger, J. Kazama, and S. Kurohashi. Extracting paraphrases from definition sentences on the Web. In *Proc. of ACL*, pp. 1087–1097, 2011.

[13] C. Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *Proc. of ACL*, pp. 341–348, 1999.

[14] D. Lin and P. Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, Vol. 7, No. 4, pp. 343–360, 2001.

[15] I. Mel’čuk and A. Polguère. A formal lexicon in Meaning-Text Theory (or how to do lexica with words). *Computational Linguistics*, Vol. 13, No. 3-4, pp. 261–275, 1987.

[16] Y. Shinyama, S. Sekine, K. Sudo, and R. Grishman. Automatic paraphrase acquisition from news articles. In *Proc. of HLT*, 2002.

[17] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, , and C. D. Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proc. of NIPS*, 2011.