G-08

自動獲得した大規模な言い換え知識を用いた言い換えの分類

河合 剛巨† 橋本 力† 鳥澤 健太郎†
Takao Kawai Chikara Hashimoto Kentaro Torisawa

概要

言い換えの認識／生成は，多くの自然言語処理アプリケーションの精度向上に寄与すると期待されている技術であり，近年活発に研究されている．しかし，言い換えという言語現象，あるいは言い換えの自然な分布に関する理解はさほど進んでいないと思われる．つまり，世の中にはどのような種類の言い換えがどの程度の割合で存在しているのかが未だにほとんど不明である．その大きな原因の一つとして，世の中に自然に存在する言い換えの縮図と見なせるデータが入手困難であることが挙げられる．本研究では，Web から獲得した大規模な言い換え知識を，世の中に自然に存在する言い換えの縮図と近似的に見なすことで，言い換えという言語現象を，意味的な構成性の観点から，定量的に分類／分析することを試みる．

１．はじめに

自然言語処理の難しさは，言葉の意味と表現形式が一対一に対応しない場合があることに由来すると考えられる．つまり，一つの表現形式が複数の意味を表したり，一つの意味が複数の表現形式で表されたりするという現象のせいであると考えられる．前者は曖昧性の問題であり，後者は言い換えの問題である．曖昧性も言い換えも言葉の意味に直接関わる問題であり，それを解決する技術は多くの自然言語処理アプリケーションにとって有用であると考えられる．

言い換えの認識／生成の研究は，比較的最近になって活発化し，技術レベルも確実に進歩していると考えられるが，言い換えという言語現象そのものに関する理解，とりわけ言い換えの定量的な分析はさほど進んでいないと思われる．その大きな原因の一つとして，世の中に自然に存在する言い換えの縮図と見なせるデータが入手困難であることが挙げられる．この原因は次のように考えることができる．人は言い換える能力を持っているが，実際に書くあるいは発話するのは，通常，言い換え前あるいは言い換え後の表現のいずれか一方のみである．それが，言い換え関係にある表現対の収集を困難にしているものと考えられる．

本研究では，Web から獲得した大規模な言い換え知識を，世の中に自然に存在する言い換えの縮図と近似的に見なし，それに基づいて，言い換えという言語現象の分析を試みる．この言い換え知識は，日本語の Web 文書約 6 億から獲得した言い換え関係にある用言句の対の集合で，Web 上に存在する多くのドメインをカバーしている．詳しくは後述するが，その規模と多様性，ドメイン非依存性から，世の中の言い換えの縮図と近似的に見なすことに一定の妥当性があると考えられる．以後，この言い換え知識を言い換え用言句対と呼ぶ．言い換え関係にある句対を用いて分析するため，単語レベル，句レベル，文レベル，談話レベルといくつかある言い換えの粒度のうち，本研究の対象は句レベルに限定される．しかし，本研究で用いる言い換え用言句対は，単語数の少ないものから多いものまでバラエティーに富んでおり，単語レベルや文レベルの言い換え現象に対しても一定の知見を与えるものと考えられる．

本研究では言い換え用言句対を意味的構成性の観点から分類／分析する．具体的には，句間の同義性を，どの程度，句を構成する単語の間の同義性に還元できるか，という観点から言い換え用言句対を分類／分析する．例えば，「本を買う」と「書籍を購入する」という言い換え用言句対は，「本」と「書籍」，「買う」と「購入する」という 2 つの単語対の同義性に完全に還元できるため，構成性が高いと言える．一方，「苦しみを取り除く」と「やすらぎを与える」，あるいは「癪に障る」と「腹の虫がおさまらない」という言い換え用言句対は，単語間の同義性に全く還元できないため，構成性が低いと言える．構成性の観点に基づく言い換えの定量的な分析により，言い換えの認識／生成という問題がどの程度難しいものなのかが見えてくると期待できる．つまり，仮に世の中の言い換えのほとんど全てが構成性の非常に高い，あるいは単語間の同義性に完全に還元できるものであるなら，言い換えの認識／生成は単語間の同義性の認識／生成の問題に帰着できる可能性が高く，同義性を扱う語彙資源を着実に整備することによって，言い換えの問題のほとんどを解決できるかもしれないという見通しが立つ．逆に，世の中の言い換えのほとんど全てが構成性の非常に低いものであるなら，句，文，あるいは談話自体が備える（単語には還元できない）意味を考慮する必要があるため，言い換えの認識／生成はやはり難しい問題である，と考えることができる．

以下，本稿では，次節で関連研究について述べた後，我々が世の中に存在する言い換えの縮図と近似的に見なす言い換え用言句対について述べる．その後，本研究の分析の枠組みについて説明する．分析結果とその考察，結論は大会当日に発表する．

２．関連研究

これまでに，言語学的研究や言い換え生成の観点から言い換えにはどのような種類があるか，言い換え認識と言い換え生成の技術にはどのような手法があるかが整理されてきた[1,2]．乾ら[2]によると，言い換えには少なくとも，語彙・構文的言い換え，参照的言い換え，語用論的言い換えの３種類がある．後述するように，語彙・構文的言い換えは使用されている語彙や構文の違いによる言い換えである．参照的言い換えは「昨年」と「2010 年」などのように特定の文脈や談話の状況でのみ成り立つものであり，語用論的言い換えは特定の目的を達成することを期待して生じる表現の違いである．

また，藤田ら[3]は，先行研究で示された事例や人手による作例をもとにした言い換え現象の類型化を試みており，

†独立行政法人 情報通信研究機構 (NICT) ユニバーサルコミュニケーション研究所 情報分析研究室

新聞を用いて半自動生成した特定の言い換えパターンの事例収集も行っている[4].

しかし，言い換えという言語現象自体に関する理解はさほど進んでいないと思われる．特に，自然言語処理にとって極めて重要な意味的構成性の観点からの定量的な分類／分析はほとんどなされていない．本研究では，Web から獲得した大規模でバラエティーに富む，ドメイン非依存の言い換え用言句対を用いて，この問題に対して定量的な知見を与えることを目的とする．

言い換えは，互いに意味的に含意し合うテキストのペアであると見なすことができる．つまり，言い換え認識の問題はテキストペアの間の両方向の含意関係認識の問題と見なすことができる．このように言い換え認識とテキスト含意認識（RTE; Recognising Textual Entailment）[5]は関連し合っており，言い換えの知見は RTE にとっても重要であると言える．

3. 分類／分析対象の言い換え用言句対

本節では，我々が言い換えの分析に使用する言い換え用言句対について述べる．

本研究で用いる言い換え用言句対は[6]で提案されている手法の獲得結果から得られている．この手法（以下，橋本法）は，同じ概念の定義文は，その定義内容が正しければ，使われている表現が異なっていても，互いに概ね同じ意味を表す，つまり，概ね言い換え関係にある，という仮説に基づいている．橋本法は，この仮説に基づき，同じ概念の定義文ペアから，言い換え関係にある用言句対を獲得する．例として(1)に「骨粗鬆症」の定義文ペアを挙げる．両定義文が概ね同じ意味を表し，その中に下線 1，2 で表されている言い換え用言句対が含まれているのがわかる．

（1）a. 骨粗鬆症とは，<u>骨の量が減り</u>$_1$，<u>骨がもろくなってしまう</u>$_2$ 病気だ．
　　b. 骨粗鬆症とは，<u>骨量が低下し</u>$_1$，<u>骨が折れやすくなる</u>$_2$ 病気だ．

橋本法の概略は次の通りである．(i) 日本語 Web 文書約 6 億から何らかの概念を定義している定義文を自動獲得する．定義文自動獲得手法は，「^NP とは.*」という言語パターンと，文末表現と「^NP とは.*」直後の表現を素性とする分類器を併用したもので，F 値 0.91 と高い精度を示している．(ii) 自動獲得した定義文から，定義対象の概念が同じもの同士ペアにすることで，定義文ペアを構築する．結果として約 2,960 万の定義文ペアが得られている．(iii) 各定義文から，係り受け解析を通して，用言句を抽出する．ただし用言句は，元の定義文で連続している 2 つ以上の文節（いずれも指示詞は含まない）から成り，末端の文節全てが体言を含むものに限っている．また，1 用言句につき用言は一語のみである．(iv) 定義文間で，全ての用言句を用いて用言句対を網羅的に構成する．(1)の場合，(骨の量が減り，骨量が低下し)，(骨の量が減り，骨が折れやすくなる)，(骨がもろくなってしまう，骨量が低下し)，(骨が折れやすくなる，骨が折れやすくなる)等が構成される．(v) SVM により，用言句対を言い換え関係にあるかどうかで分類し，さらに分離平面からの距離で用言句対を順位づける．分類に用いた素性は，用言句間の類似度に関するものと，用言句の文脈の間の類似度に関するものの 2 種類に大別される．これは，用言句自身が互いに類似しているか，文脈が類似しているか，あるいはその両方が成り立つ場合に言い換え用言句対になりやすい，という観察結果に基づく．

[6]によると，橋本法は，定義文ペア 10 万ペアを対象にした評価実験において，ランク上位 1,000 での適合率が約 94%と，先行研究の手法と比べて非常に高い精度を示している．

橋本法は Web 上に存在するあらゆる概念の定義文から言い換えを獲得しているため，カバーしているドメインも非常に多岐にわたる．実際，「FX」，「メタボリックシンドローム」，「ポケモン」，「カイピリーニャ」，「労働者派遣」，「捕鯨」，「オフ会」など，約 87 万にのぼる非常に多彩な概念に関する定義文から言い換え用言句対が獲得されている．

以下に橋本法で実際に獲得された言い換え用言句対の例を挙げる．

（2）a. 個人情報が流出した
　　b. 個人情報が漏洩した
（3）a. デオキシリボ核酸の塩基配列を分析する
　　b. ＤＮＡの塩基配列を分析する
（4）a. 品質や価格のみで選ぶのではない
　　b. 品質や価格だけで選ぶのではない
（5）a. 胸から下だけをお湯に浸ける
　　b. みぞおちから下だけをお湯につける

上記の(2)は用言が，(3)は体言が異なる例である．(4)は機能語が異なり，(5)は用言だけでなく体言部分にも相違がある例である．この他，モダリティが異なるものや，これらの組み合わせなど多様な言い換え用言句対が存在する．もちろん，手法上の制約はないとは言えないため，この獲得結果に偏りが全く無いとは断定できない．しかし，上記の通り，多種多様な言い換え用言句対を大量に含むことから，橋本法の獲得結果を本研究の分析に用いることには一定の妥当性があると考えることができる．

4. 分類／分析の枠組み

本研究では次の手順に基づいて，言い換え用言句対の分類／分析を行う．

I. 日本語 Web 文書約 6 億から得られる定義文ペア約 2,960 万全体に対して橋本法を適用する．

II. 橋本法の出力から言い換え用言句対をサンプリングして，言い換えとして正しいかどうかを人手でチェックする．

III. 言い換えとして正しいと判断された言い換え用言句対のみを対象に，用言句の間で同義関係にある構成要素の対（単語対あるいは文節対）を人手でチェックする．

IV. 言い換えとして正しいと判断された言い換え用言句対を，次の 4 通りに分類する．

① one-to-one
　用言句の間で，全ての構成要素が 1 対 1 に対応する．これに属する言い換えは意味的に完全に構成的であり，また，構成要素間の意味的対応が容易に判断できると考えられるため，その認識／生成の問題が構成要素間の同義性の問題に帰着できる可能性が高い．

② many-to-many

用言句の間で，全ての構成要素が相手方の用言に同義である構成要素を持つが，1 対 1 対応ではなく，多対多あるいは多対 1 対応である．これに属する言い換えは意味的な構成性が非常に高いが，構成要素間の意味的対応の判断が難しい場合があると考えられるため，one-to-one に属する言い換えに比べて，その認識／生成の問題を構成要素間の同義性の問題に帰着できる可能性は低いと考えられる．

③ partial

意味的な対応が取れる構成要素だけでなく，対応が取れない構成要素も存在する．これに属する言い換えは意味的な構成性が低いため，その認識／生成の問題を，その構成要素の同義に関する知識だけを用いて解くことは困難であると考えられる．つまり，句，文，談話自体の（構成要素に還元できない）意味を考慮する必要があるため，上記 2 種類の言い換えに比べて，認識／生成が困難であると考えられる．

④ none

意味的な対応が取れる構成要素が全く存在しない．これに属する言い換えは，構成要素の同義に関する知識を認識／生成の手がかりとして利用することができないため，句，文，談話自体の意味を考慮して，その結果のみに基づいて処理を行わなくてはならないと考えられる．つまり，自動認識／生成が最も困難な種類である．

V. 上記の各分類，one-to-one，many-to-many，partial，none の数を集計する．

以下，one-to-one の例を(6)に，many-to-many の例を(7)に，partial の例を(8)に，none の例を(9)に挙げる．下線部の添字は意味的に対応する構成要素を表す．

(6) a. 体の$_0$毒素を$_1$排出する$_1$

b. 体の$_0$毒素を$_1$出す$_1$

(7) a. ネットカフェで$_0$寝泊りを$_1$している$_2$

b. ねぐらに$_{1,2}$ネットカフェを$_0$利用している$_{1,2}$

(8) a. 赤血球に$_1$含まれる$_2$

b. 血液の$_0$赤血球に$_1$含まれる$_2$

(9) a. 苦しみを$_0$取り除く$_1$

b. やすらぎを$_2$与える$_3$

5. 言い換え用言句対に対する既存言語資源のカバレッジ

また，本研究では，これまでに構築されてきた言語資源の世の中に存在する全ての言い換えに対するカバレッジを見積もるため、本研究で用いた言い換え用言句に対する既存言語資源のカバレッジを測定する。

既存の言語資源として，下記を用いる．

- ALAGIN[1] 動詞含意関係データベース(V1.3.0) [7]
- 推移律による自動拡張版の動詞含意関係 DB [8]
- ALAGIN 日本語異表記対データベース(V1.1) [9,10]
- ALAGIN 基本的意味関係の事例ベース(V1.1) [9,11]
- ALAGIN 上位語階層データ(V1.1) [12]
- Co-STAR 上位下位関係データ[13]
- ALAGIN 文脈類似語データベース(V1.1.2) [11,14]

以下で，それぞれについて説明する．

ALAGIN 動詞含意関係データベース(V1.3.0) (以下，VEDB)には，自動獲得[7]により得られた含意関係が成立している人手チェック済の動詞のペア（正例 50,079 ペア，負例 38,787 ペア）が含まれる．えば「再婚する」は「離婚する」を含意する(この含意関係を「再婚する→離婚する」のように表す)．本稿では，正例を用いる．

推移律による自動拡張版の動詞含意関係データベースは，VEDB を推移律の適用により拡張した動詞のペア[8]である．例えば「再婚する→離婚する」と「離婚する→結婚する」から推移律の適用により「再婚する→結婚する」が得られる．本研究では，推移律を一回だけ適用して得られた 108,366 ペアを用いる．

ALAGIN 日本語異表記対データベース(V1.1)は，人手作成の異表記対[9]と，編集距離 1 の異表記対の正例と負例を集めたデータである．本稿では人手作成の正例 48067 対と，機械学習の自動獲得[10]による約 150 万対を用いる．

ALAGIN 基本的意味関係の事例ベース(V1.1)は，後述の文脈類似語データベースをもとに，文脈類似度の高い 2 語間の関係を分類し，ラベル付けした 39,561 対である[9,11]．本稿では，各分類を同義性の観点から下記２つのグループに分けて利用する．S は同義や対義の関係にあるグループで， R は S に関連語を加えたグループである．

- S (38,240 対):
  異表記対，略記対，異形同義語対，対義語対
- R (39,561 対):
  上記 S に，部分・全体語対 を追加

ALAGIN 上位語階層データ(V1.1)は，日本語 Wikipedia から自動処理で得られた上位下位関係のうち，人手クリーニングにより構築された約 69,000 名詞句の階層シソーラスである[12]．本稿は各階層で隣接するペアを用いる．

Co-STAR[13]は，Web の非構造テキストと Wikipedia の構造テキストの co-training 手法で獲得された上位下位関係のデータである．本稿では，適合率 90%相当の関係（Web 文書から約 43 万，Wikipedia から約 460 万）を用いる．

ALAGIN 文脈類似語データベース(V1.1.2)は，100 万語の名詞に対して，約 1 億ページの Web 文書上での文脈が類似している名詞を類似度とともに順に最大 500 個列挙したデータである．本研究では[14]の手法で得られた結果から単語とは認めがたいものを人手で除去したデータを用いる．

6. おわりに

本研究では，Web から獲得した大規模な言い換え用言句対を世の中に存在する言い換えの縮図と近似的に見なし，それらを意味的構成性という自然言語処理にとって極めて

[1] ALAGIN（高度言語情報融合フォーラム）
http://alaginrc.nict.go.jp/

重要な観点から分類するという試みについて述べた．分類／分析結果は発表の際に報告する．

近年言い換えの認識／生成に関する研究が活発化し，その技術は着々と進歩していると考えられるが，言い換えという言語現象自体，あるいは言い換えの自然な分布についての理解はほとんど進んでいないと考えられる．本研究は，言い換えという言語現象，あるいは言い換えの自然な分布の理解に一定の役割を果たすものであるが，その分析の方法論は改善の余地が多分にある．今後は，まず，調査の大規模化に向けて，言い換えの構成要素間の意味的な対応を自動で取るための高精度なアライメント手法を開発する．また，用言句対だけでなく，体言句対や，文レベルの言い換えも今後調査対象とする．

参考文献

[1] Shiqi Zhao, Haifeng Wang. Paraphrases and applications. The 23rd Int'l Conf. on Computational Linguistics (Coling 2010), Paraphrases and Applications–Tutorial notes, pp. 1–88, 2010.

[2] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理, Vol.11, No.5, pp.151–198, 2004.

[3] 藤田篤, 乾健太郎, 松本裕治. 言い換え知識の類型化と例文集構築の試み, 言語処理学会第 10 回年次大会発表論文集, 2004.

[4] 藤田篤, 乾健太郎, クラス指向事例収集手法による言い換えコーパスの構築, 自然言語処理 Vol.13, pp.133–150, 2006.

[5] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Proc. of the PASCAL Challenges Workshop on Recognizing Textual Entailment, 2005.

[6] Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Junichi Kazama, and Sadao Kurohashi. Extracting paraphrases from definition sentences on the web. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011), pp.1087–1097, 2011.

[7] Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Masaki Murata, and Jun'ichi Kazama. Large-scale verb entailment acquisition from the web. In Proc. of EMNLP2009, pp.1172–1181, 2009.

[8] 橋本力, 鳥澤健太郎, 黒橋禎夫, ステイン・デ・サーガ, 風間淳一, 藤田篤. 動詞含意関係データベースの自動拡張. 言語処理学会第 16 回年次大会発表論文集, (D4-4), pp.4–7, 2010.

[9] 黒田航, 風間淳一, 村田真樹, 鳥澤健太郎. Web 文書に対応できる日本語異表記対の認定基準. 言語処理学会 16 回 年次大会発表論文集, (A5-6), pp.990‑993, 2010.

[10] 小島正裕, 村田真樹, 風間淳一, 黒田航, 藤田篤, 荒牧英治, 土田正明, 渡辺靖彦, 鳥澤健太郎. 機械学習と種々の 素性を用いた編集距離の小さい日本語異表記対の抽出. 言語処理学会第 16 回年次大会発表論文集, (D4-1), pp.928‑931, 2010.

[11] 風間淳一,ステイン・デ・サーガ, 鳥澤健太郎, 村田真樹. 係り受けの確率的クラスタリングを用いた大規模類似語リストの作成. 言語処理学会第 15 回年次大会発表論文集, (C1-6), pp.84–87. 2009.

[12] 黒田航, 李在鎬, 野澤元, 村田真樹, 鳥澤健太郎, 鳥式改の上位語データの人手クリーニング. 言語処理学会 15 回年次大会発表論文集, (C1-3), pp.76-79, 2009.

[13] Jong-Hoon Oh, Ichiro Yamada, Kentaro Torisawa, and Stijn De Saeger. Co-STAR: A Co-training Style Algorithm for Hyponymy Relation Acquisition from Structured and Unstructured Text. Computational Linguistics, pp.842–850, 2010.

[14] Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, Kentaro Torisawa. A Bayesian Method for Robust Estimation of Distributional Similarities. ACL 2010.