言い換え可能な括弧表現の抽出法

岡崎　直観　　石塚　満

東京大学大学院情報理工学系研究科

1　はじめに

我々は，異なる語，句，文法構造を用いて，同じ内容を伝達する複数の文を生成できる．このような自然言語の柔軟な記述力に対応するための技術として，言い換えの研究が注目を浴びている [乾 04]．言い換えには様々な方式があるが，語彙的な言い換えには，WordNetやＥＤＲ電子化辞書などの語彙資源が欠かせない．語彙的な言い換え知識を自動的に獲得する研究は，パラレル・コーパスを用いるもの [BL03, PKM03] と，ノンパラレル・コーパスを用いるもの [Lin98, 山本 02, 酒井 05, HOT06] に大別できる．言い換え知識を獲得する手がかりとなるパラレル・コーパスとは，例えば，フランス語の同一の記事を英語に翻訳した複数の記事集合や，同じ出来事を伝える複数の情報源の記事集合などである．ノンパラレル・コーパスを用いる研究は，単語の分布仮説（distributional hypothesis）[Har54] に基づき，単語間の意味的な距離を計測し，類義語を獲得するものが大勢を占める．

本研究は，言い換え可能な語彙対を，ノンパラレル・コーパスに含まれる括弧表現から獲得する手法を提案する．日本語の括弧表現「X（Y）」を分析すると,「朝鮮民主主義人民共和国（北朝鮮）」のように，文脈にほとんど依存することなく，相互に言い換えが可能な語彙対と,「人民日報（中国）」のように，言い換えが必ずしも成立しない語彙対が混ざって得られる．本研究のゴールは，括弧表現「X（Y）」が与えられたときに，表現「X」と「Y」が相互に言い換え可能であるかどうかを推定することである．

2　括弧表現とその関連研究

括弧表現に着目した語彙的言い換え知識の獲得は，主に英語を対象とした略語抽出として研究が進められてきた [Ada04, OA06]．略語抽出は,「European Union (EU)」と記述される括弧表現に着目し，略語の短縮形と完全形（定義）の対を獲得するタスクである．英語文書を対象とする場合，括弧表現「X（Y）」に対して,「X」と「Y」に略語の関係が成立するかどうかは,「Yに含まれる文字がすべてXに含まれる」という，文字一致のヒューリスティックで検証するのが主流である．

残念なことに，英語向けの略語抽出手法を日本語文書に適用することはできない．理由としては，日本語の括弧表現の用法が多様であること，言い換え可能な語彙対の多くに，文字一致のヒューリスティックが適用できないことが挙げられる．久光ら [久光 97] は，括弧表現「X（Y）」の共起の強さに関する統計的指標と,「X」や「Y」の文字種に関する簡単なルールを組み合わせて，言い換え可能な括弧表現を抽出する手法を提案した．笹野ら [笹野 06] は，照応解析に必要な語彙的言い換え知識を獲得するために，括弧表現の要素「X」と「Y」の頻度，文字種に関するルールを設計した．村山ら [村山 06] は，日本語の頭文字が生成される過程を Noisy-Channel モデルで定式化した．

我々は，括弧表現が実際にどのような用法で用いられるか調べるために，毎日新聞と読売新聞の 1998–1999 年の記事に含まれる括弧表現を，言い換え可能性という観点から 7 種類に分類した．表 1 は，括弧表現の種類別に，言い換えの可能性（換言），文字の一致（一致），括弧表現を示している．

「頭文字」は，完全形に含まれる文字を間引いて短縮形を作成したものである．このグループの括弧表現は，短縮形と完全形が相互に言い換え可能である．短縮形に含まれるすべての文字が，完全形の中に現れるので，英語の略語と同様に，文字一致のヒューリスティックを用い，言い換えを認識できる．これに対し,「外来語の頭文字」は，完全形から短縮形を生成させるときに，英語などの外来語への翻訳を伴うため，文字一致

括弧表現のタイプ	換言	一致	例
頭文字	○	○	東京大学（東大），首都圏中央連絡自動車道（圏央道）
外来語の頭文字	○	×	欧州連合（ＥＵ），夜間離着陸訓練（ＮＬＰ），ワールドカップ（W杯）
その他の換言	○	×	朝鮮民主主義人民共和国（北朝鮮），日米防衛指針（ガイドライン） 特定非営利活動促進法（ＮＰＯ法），簡易型携帯電話（ＰＨＳ） ２０００年問題（Ｙ２Ｋ），モラルハザード（倫理観の欠如） 犯罪で得た資金の洗浄（マネーロンダリング）
属性（読み）	△	×	毅然（きぜん），Ｏ（オー）１５７
属性（場所）	×	×	つくば学園都市（茨城県つくば市），東大医科学研究所病院（東京都港区）
属性（所属）	×	×	前田（広），インディペンデント（英国），ミヒャエル・シューマッハー（独）
属性（年齢）	×	×	岡崎直観（２７）
補足・注釈	×	×	参院議員秘書（元），西ドイツ（当時），平成金融再生機構（仮称）
補完	×	×	真摯に（批判を）受け止めている．
その他	×	×	... (中略)，…（笑い），…（おわり）

表 1: 日本語の新聞記事で見つかる括弧表現の例

のヒューリスティックでは言い換えを認識できない．

「その他の換言」は，頭文字であるとは認められないものの，別の用語へ言い換えるタイプである．例えば,「朝鮮民主主義人民共和国」は，その正式名称よりも「北朝鮮」という別称でよく用いられるが，正式名称には別称の文字「北」が由来する原因が見当たらない.「ＰＨＳ」は「簡易型携帯電話」という正式名称があり，外来語の頭文字に似た構成をしている．しかしながら,「ＰＨＳ」は「Personal Handy-phone System」の頭文字で，厳密に解釈すると日本語と英語の正式名称の間には翻訳の関係が成立しない．

「属性」は，括弧表現「X（Y）」が,「X」の暗黙の属性Zの属性値として「Y」を与えるものである．例えば,「インディペンデント（英国）」は「インディペンデントの国籍は英国」と解釈できるし,「つくば学園都市（茨城県つくば市）」は「つくば学園都市の所在地は茨城県つくば市」と解釈できる．このタイプの括弧表現は，表現「X」と「Y」に言い換えの関係が成立しない．暗黙の属性Zとしては，読み，場所，所属，年齢，構成員，曜日，順位，情報源など，多様なものが文脈に応じて用いられる．紙面の都合で「補足・注釈」「補完」「その他」に関する説明は省略するが，これらのタイプも言い換えを導入しない．

本研究のゴールは，括弧表現を言い換え可能なものと不可能なものに分類することであった．表 1 の分類に従うと,「頭文字」「外来語の頭文字」「その他の換言」に属する括弧表現を抽出すれば良さそうである．我々は，1998–1999 年の毎日新聞・読売新聞記事（全 596,098 記事）に含まれる括弧表現 「X（Y）」に対し，表現「X」と「Y」の共起頻度が 8 よりも大きい語彙対 7,887 件を抽出した．評価コーパスを構築するため，抽出した語彙対のすべてを，手作業で「頭文字」「外来語の頭文字」「その他の換言」「非換言」に分類した．表 2 は，各タイプの語彙対の分布を示したものである．言い換えが可能な語彙対は「外来語の頭文字」「その他の換言」に集中していることが分かる．

括弧表現のタイプ	事例数	(%)
頭文字	90	(1.2)
外来語の頭文字	717	(9.1)
その他の換言	623	(7.9)
非換言	6,457	(81.9)
計	7,887	(100.0)

表 2: 括弧表現の用例の分布

表 3 は，抽出された括弧表現「X（Y）」のうち，共起頻度の高いもの 10 件を抜粋したものである．タイプは，その括弧表現が頭文字（A），外来語の頭文字（T），その他の換言（O），非換言（F）のどのタイプに属するか示している．この表の中に含まれる括弧表現は，7/10 が言い換え可能であるが，残りの 3/10 は共起頻度が高いにも関わらず，新聞社の国籍を示す属性の用法で用いられている．

3 提案手法

本節では，文書において言い換えを導入するパターンに着目し，表現「X」と「Y」に言い換えの関係が成立するかどうかを調べる指標を提案する．前節で述

#	「Y」	「X」	頻度	タイプ
1	北朝鮮	朝鮮民主主義人民共和国	4160	O
2	W杯	ワールドカップ	2891	T
3	ＥＵ	欧州連合	2638	T
4	ＮＡＴＯ	北大西洋条約機構	2593	T
5	ＩＭＦ	国際通貨基金	2473	T
6	中国	人民日報	1561	F
7	ＩＯＣ	国際オリンピック委員会	1550	T
8	ＷＴＯ	世界貿易機関	1504	T
9	独	ディ・ウェルト	1484	F
10	エジプト	アルアハラム	1350	F

表 3: 新聞記事で頻繁に共起する括弧表現

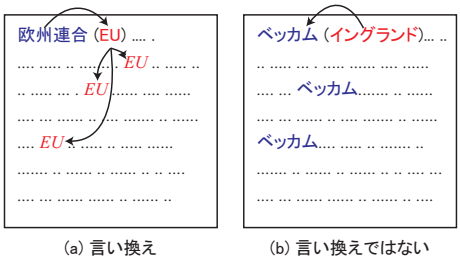

図 1: 括弧表現による言い換え

べたように，括弧表現「X（Y）」の表現「X」と「Y」の間に言い換えの関係が成立する要因は複雑であるため，複数の指標を教師あり機械学習で統合する．

言い換え発生率　文書の著者が括弧表現「X（Y）」で言い換え「X → Y」を導入する状況を考える．「X（Y）」と併記する理由は，表現「Y」を単独で記述したとしても，読者が「Y」の定義を正しく認識できるようにすることである．例えば，「夜間離着陸訓練（ＮＬＰ）」という括弧表現があれば，その文書における表現「ＮＬＰ」は「夜間離着陸訓練」を指し，特に断りが無ければ「自然言語処理」という意味で解釈しない．

同時に，もし括弧表現「X（Y）」が言い換え「X → Y」を導入するためのものであれば，その括弧表現の後では表現「X」よりも「Y」が好んで用いられると推測される．この状況を図示したものが図 1 である．文書 (a) は，「欧州連合 → ＥＵ」という言い換えを定義し，括弧表現以降では「ＥＵ」という表現を多く用いているのに対し，文書 (b) では固有名詞「ベッカム」の国籍の属性値として「イングランド」を挙げており，括弧表現以降でも「ベッカム」が多く用いられている．

そこで，「X（Y）」というパターンが出てくる文書を集め，以下の 2 つの条件を同時に満たす文書は，「X → Y」の語彙的言い換えを導入したと認定する．

1. 「X（Y）」のパターンが出てくる前の文において，表現Yが出現しない．
2. 「X（Y）」のパターンが出てきた後の文において，表現Xよりも表現Yの出現頻度が高い

式 1 は，表現「X」と「Y」の言い換え発生率を与える．

$$\mathrm{PR}(X,Y) = \frac{d_{\mathrm{para}}(X,Y)}{d(X,Y)}. \qquad (1)$$

ただし，$d_{\mathrm{para}}(X,Y)$ は上述の条件を満たす文書の数，$d(X,Y)$ は括弧表現「X（Y）」を含む文書の総数である．言い換え発生率 $\mathrm{PR}(X,Y)$ は，表現「X」と「Y」に対して，0（言い換えの発生なし）から 1（すべての括弧表現が言い換えを導入している）までの値を返す関数である．

その他の指標　本研究では，括弧表現による言い換えを様々な要因から捉えるため，以下の指標も導入する．

- 共起頻度 $\mathrm{FREQ}(X,Y)$: 「X（Y）」の出現頻度
- χ^2 による共起度 $\mathrm{CHI2}(X,Y)$: 「X（Y）」の共起度を χ^2 値で測ったもの [久光 97]
- 文字の一致 $\mathrm{MATCH}(X,Y)$: 「Y」の中にあるすべての文字が「X」にも含まれる場合に 1 を返し，それ以外の場合に 0 を返す関数
- コンテキストの類似性 $\mathrm{SKEW}(X,Y)$: 「X」と係り受け関係を持つ単語の頻度分布を P とし，「Y」と係り受け関係を持つ単語の頻度分布を Q としたときに，確率分布 P と Q の距離を Skew Divergence ($\alpha = 0.99$) で測ったもの [Lee01, 山本 02]

$$\mathrm{SKEW}_\alpha(P||Q) = \mathrm{KL}(P||\alpha Q + (1-\alpha)P), \qquad (2)$$

$$\mathrm{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \qquad (3)$$

複数の指標の統合　括弧表現「X（Y）」の言い換え可能性を言い当てる問題は，事例「X（Y）」を言い換え可能（正例）と言い換え不可能（負例）の 2 値に分類する問題に帰着される．そこで，これまで述べてきた複数の指標を Support Vector Machine (SVM) の素性とし，評価コーパスを訓練例として教師有り学習を

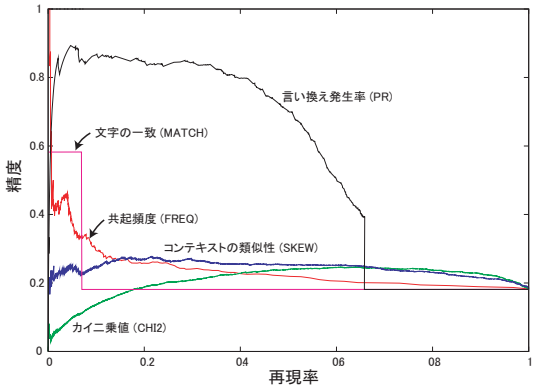

図 2: 各指標単独での再現率–精度

行い，言い換え可能性に関する分類器を構築した．括弧表現「X（Y）」に対する素性は，これまでに述べた 5 つの指標の値に，逆向きの括弧表現「Y（X）」に関する各指標の値を追加した 10 個である．

4 評価

図 2 は，各指標を単独で用い，適当な閾値を用いて括弧表現「X（Y）」を言い換え可能・不可能に分類したとき，精度と再現率の変化をプロットしたものである．言い換え発生率（PR）の指標がもっとも良い識別性能を示していることが分かる．もっとも高い F 尺度が得られたときの評価は，精度 69.8%，再現率 50.8%であった．文字の一致（MATCH）の指標は，日本語の括弧表現ではうまく働かず，もっとも性能が良い場合でも，精度 58.2%，再現率 6.9%だった．カイ二乗 (CHI2) の指標は，予想に反して性能が悪いが，「ポタポビッチ（カザフスタン）」，「ビドヘルツル（オーストリア）」など，固有名詞とその国籍を示す括弧表現の共起が強く，カイ二乗値だけで言い換えの有無を判定するのは難しい．

SVM で言い換えに関する分類器を構築し，評価コーパスをそのまま訓練コーパスとして用い，10 分割交差検定を行ったときの性能は，正解率 90.2%，精度 80.9%，再現率 59.7%であった．各指標を単独で用いた場合と比較すると，指標を統合した効果が見られる．再現率を括弧表現のタイプ別に調べると，頭文字（94.4%），外来語の頭文字（74.3%），その他の換言（46.0%）であり，翻訳辞書を用いなくても外来語を認識していることが分かる．今後は，頭文字以外の言い換え表現のモデル化を精緻化し，再現率の向上を図る予定である．

参考文献

[Ada04] Eytan Adar. SaRAD: A simple and robust abbreviation dictionary. *Bioinformatics*, Vol. 20, No. 4, pp. 527–533, 2004.

[BL03] Regina Barzilay and Lillian Lee. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the NAACL-HLT 2003*, pp. 16–23, 2003.

[Har54] Zellig S. Harris. Distributional structure. *Word*, Vol. 10, pp. 146–162, 1954.

[HOT06] Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama. Selection of effective contextual information for automatic synonym acquisition. In *Proceedings of the COLING-ACL 2006*, pp. 353–360, Sydney, Australia, 2006.

[Lee01] Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics 2001*, pp. 65–72, 2001.

[Lin98] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the COLING-ACL 1998*, pp. 768–774, Montreal, Quebec, Canada, 1998.

[OA06] Naoaki Okazaki and Sophia Ananiadou. Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, Vol. 22, No. 24, pp. 3089–3095, 2006.

[PKM03] Bo Pang, Kevin Knight, and Daniel Marcu. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. In *Proceedings of the NAACL-HLT 2003*, pp. 102–109, 2003.

[乾 04] 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.

[久光 97] 久光徹, 丹羽芳樹. 統計量とルールを組み合わせて有用な括弧表現を抽出する手法. 研究報告 - 自然言語処理, Vol. 1997, No. 109, pp. 113–118, 1997.

[笹野 06] 笹野遼平, 河原大輔, 黒橋禎夫. 自動獲得した知識に基づく統合的な照応解析. 言語処理学会第 12 回年次大会, pp. 480–483, 2006.

[山本 02] 山本和英. テキストからの語彙的換言知識の獲得. 言語処理学会第 8 回年次大会, pp. 639–642, 2002.

[酒井 05] 酒井浩之, 増山繁. 略語とその原型語との対応関係のコーパスからの自動獲得手法の改良. 自然言語処理, Vol. 12, No. 4, pp. 207–231, 2005.

[村山 06] 村山紀文, 奥村学. Noisy-channel model を用いた略語自動推定. 言語処理学会第 12 回年次大会, pp. 763–766, 2006.