2G4-4

日本語新聞記事からの略語抽出

Abbreviation Recognition in Japanese Newspaper Articles

岡崎　直観[*1]　　石塚　満[*1]
Naoaki Okazaki　Mitsuru Ishizuka

[*1]東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo

Human languages are rich enough to express the same meaning in different words; we may produce different sentences to convey the same information by choosing alternative words, phrases, or syntactic structures. We present a novel method for acquiring lexical paraphrases expressed by parentheses. More specifically, given a parenthetical expression *X (Y)*, our goal is to determine if the expressions *X* and *Y* are *paraphrasable*, i.e., the expression *X* can be substituted with *Y* and vice versa. This paper addresses the difficulty of this task by examining various relation types expressed by parentheses in Japanese newspaper articles. We designed a classifier based on Support Vector Machines (SVMs) that combines paraphrase likelihoods modeled by various features. The proposed method achieved 95.7% accuracy on our evaluation corpus containing 7,887 (1,430 paraphrase and 6,457 non-paraphrase) instances found in Japanese newspaper articles.

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計算機が対応するためには，WordNet やＥＤＲ電子化辞書のような語彙資源を参照し，同一の実体・概念を示す異なる表現を認識しておく必要がある．本研究は，日本語のテキストに含まれる括弧表現に着目し,「欧州連合（ＥＵ）」のような言い換え可能な語彙対を自動的に獲得する手法を提案する．

これまで，語彙的な言い換え知識の自動獲得に向けて，様々な手法が提案されてきた．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（χ^2 検定，対数尤度比など）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．山本 [山本 02] は，括弧表現の要素「X」と「Y」が言い換え可能であれば,「X」と「Y」の周辺語の分布が類似するという単語の分布仮説（distributional hypothesis）[Harris 54] に基づき，言い換え可能性を定式化した．菅野ら [菅野 05] は，括弧表現の要素「X」と「Y」の意味的な近さをシソーラスを用いてスコア付けした．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雑音あり通信路モデルで定式化した．

このように，言い換え可能な語彙対の自動獲得に向けて，括弧表現の要素「X」と「Y」に関する様々な特徴に着目した手法が提案されてきた．本研究では，それらの特徴の有効性を見積もるため，括弧表現がどのような用法で用いられるのか調べた．調査の対象は，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）」のうち，表現「X」と「Y」の共起頻度が 8 よりも大きい語彙対 7,887 件である．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 1 は，それぞれの分類に対して，言い換えの可否（換言），文字一致ヒューリスティックの成否（一致），見つかった事例数とその割合，括弧表現例をまとめたものである．

頭文字は，表現「X」に含まれる文字を間引いて「Y」を作成したものである.「Y」に含まれるすべての文字が,「X」の中に現れるので，その特徴に着目することで言い換えを認識できるが，事例数は 1.2%と少ない．外来語の頭文字は，英語などの外来語「Z」を日本語に翻訳して語「X」を導入したものの，略語「Y」は元々の外来語「Z」に由来していると解釈できるものである．たとえば「欧州連合」は,「European Union」という英語表現を日本語に翻訳したものであるが，日本語においても英語の略語「ＥＵ」をそのまま用いる．言い換え可能な括弧表現のうち，もっとも事例数が多かったのはこのタイプである．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」が由来する表層的な要因が見当たらない．

頻繁に共起する括弧表現の 81.9%は，要素「X」と「Y」の間に言い換えの関係が成立しない．このタイプの括弧表現にはいろいろな用法があるが，括弧表現「X（Y）」が,「X」の暗黙の属性Ｒの属性値として「Y」を与えるものが多く見つかる．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．暗黙の属性Ｒ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．

2. 提案手法

本研究では，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，略語抽出を行う．前節で述べたように，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．そこで，事例「X（Y）」に対して，言い換え可能（正例）と言い換え不可能（負例）の２値に分類する問題として捉え，Support Vector Machine (SVM) で分類器を構築する．表 2 は，今回用いた素性と，その値を「欧州連合（ＥＵ）」という括弧表現を例に示したものである．数値

連絡先: 岡崎直観，東京大学大学院情報理工学系研究科，113-8656 東京都文京区本郷 7-3-1，03-5841-4120，okazaki@is.s.u-tokyo.ac.jp

括弧表現のタイプ	換言	一致	事例数	(%)	例
頭文字	○	○	90	(1.2)	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	○	×	717	(9.1)	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	○	×	623	(7.9)	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）	×	×	6,457	(81.9)	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

型，真理値型の素性の値は，そのまま SVM の属性値として与え，文字列型の素性は，値がその文字列になるかどうかの真理値に変換される．

表中の「X（Y）」の頻度から，対数尤度比による共起度までは，括弧表現「X（Y）」の出現頻度に関する統計値である．文字の包含は，「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 を返す関数で定義される素性であり，頭文字による略語の解釈を試みる．

コンテキストの分布距離は，語「X」と「Y」が言い換え可能であれば，それぞれの周辺に存在する語も似たような分布を示すであろうという仮説に基づいている．「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 ($\alpha = 0.99$) で測定する [Lee 01]．

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

$$\mathrm{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \quad (2)$$

品詞コードのペアは，「X」と「Y」が言い換え可能であれば，語「X」と「Y」は同一の品詞であるという仮説に基づいている．たとえば，「欧州連合（ＥＵ）」に対して，「欧州連合」と「ＥＵ」の品詞コードは「名詞-固有名詞-組織」で一致するが，「欧州連合（本部はベルギー）」では，括弧内の表現の品詞が「名詞-固有名詞-地域-国」と異なる．同様の考え方に基づき，「X」と「Y」に対する品詞カテゴリ，固有表現タグのペアも素性として導入した．なお，係り受け解析と固有表現抽出には，南瓜 [Kudo 02] を用いた．

表 2 の言い換え発生率を説明するために，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．

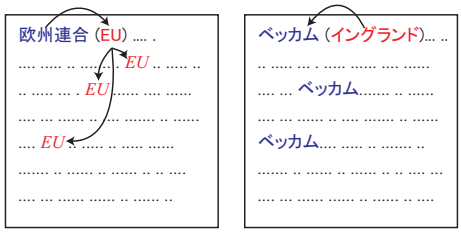

図 1: 括弧表現による言い換え

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 3 は，表現「X」と「Y」の言い換え発生率を与える．

$$\mathrm{PR}(X,Y) = \frac{d_{\mathrm{para}}(X,Y)}{d(X,Y)}. \quad (3)$$

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．

前節で分析した括弧表現のリストを，1,430 件の正例，6,457 件の負例から構成される学習コーパスとみなし，それぞれの学習インスタンスを素性で表現した．文字列型の素性は，すべての学習インスタンスを調べあげ，取りうる値すべてを列挙し，真理値型の属性値に変換した．たとえば，文字種のペアの素性の値として「kanji/alpha」と「kanji/kanji」が学習コーパス内に存在する場合は，この素性から 2 個の属性値が生成される．このようにして，表 2 で示された素性から，730 種類の属性値が得られた．さらに本研究では，言い換え「X→Y」が成立するならば，その逆の言い換え「Y→X」も成立すると考

素性	型	概要	欧州連合（ＥＵ）
「X（Y）」の頻度	数値	パターン「X（Y）」の総出現回数	2,638
「X」の頻度	数値	語「X」の単独での総出現回数	8,326
「Y」の頻度	数値	語「Y」の単独での総出現回数	3,121
χ^2 による共起度	数値	「X（Y）」の共起度を χ^2 値で測ったもの [久光 97]	2,484,521
対数尤度比による共起度	数値	「X（Y）」の共起度を対数尤度比で測ったもの [久光 97]	6.8
文字の包含	真理値	「X」が「Y」のすべての文字を含む場合に 1	0
コンテキストの分布距離	数値	「X」「Y」と係り受け関係を持つ語の分布の Skew Divergence [Lee 01]	1.35
品詞コードのペア	文字列	南瓜が「X」と「Y」それぞれに付与した品詞コードを並べたもの	固有名詞/固有名詞
品詞カテゴリのペア	文字列	南瓜が「X」と「Y」それぞれに付与した品詞カテゴリを並べたもの	名詞/名詞
固有表現タグのペア	文字列	南瓜が「X」と「Y」それぞれに付与した固有表現タグを並べたもの	ORG/ORG
文字種のペア	文字列	「X」と「Y」の字種（アルファベット，数字，ひらがな，カタカナ，漢字）のペア	kanji/alpha
言い換え発生率	数値	「X（Y）」の表記に対し，言い換えと推測されるものの割合 [岡崎 07]	0.426

表 2: 学習に用いた素性と「欧州連合（ＥＵ）」に対する素性の値の例

え,「X（Y）」の言い換え可能性を「Y（X）」の言い換え可能性からも推定する．最終的に，学習インスタンス「X（Y）」は,「X（Y）」に関する属性値と,「Y（X）」に関する属性値の両方で表現される．

3. 評価

SVM の実装として LIBSVM[*1]を用い，実験時に最もよい分類性能を示した線形カーネルで分類器を構築し，10 分割交差検定で評価を行った．括弧表現「X（Y）」が与えられた時，その表現を言い換え可能／不可能に分類するときの正解率は，95.7%であった．評価コーパスや言い換えの認定基準が異なるため，単純な比較はできないが，従来手法など [山本 02, 菅野 05] では，60%前後の精度が報告されており，提案手法による大幅な性能向上が確認できた．

次に，各素性の分類タスクにおける貢献度合いを調べるため，それぞれの素性を単独で用いて SVM 分類器を構築した時のパフォーマンスを調べた．最も高い正解率を出力した素性は，文字種（91.5%）であり，品詞コード（90.0%），固有表現タグ（89.3%），品詞カテゴリ（88.2%），言い換え発生率（87.9%）と続く．従来研究 [久光 97, 笹野 06] では，カタカナや英字などの文字種に着目して，言い換え可能性を高精度に言い当てる方法が提案されていたが，素性を単独で用いた場合，文字種が最も良いパフォーマンスを示すことは，これを裏付ける結果と言える．逆に，分類の性能が芳しくなかった素性は,「Y」の頻度（82.2%）,「X」の頻度（82.2%），文字の包含（82.2%）の順となっている．表現「X」と「Y」の単独の頻度が分類タスクに貢献しないのは，容易に想像がつくが，文字の包含（「Y」のすべての文字が「X」に含まれるかどうか）が，分類タスクに全く貢献しないのは，日本語の括弧表現特有の現象として興味深い．

4. 結論

本稿では，日本語の括弧表現「X（Y）」において，表現「X」と「Y」の間に成立する関係が複雑であることを説明した．与えられた括弧表現「X（Y）」を言い換え可能／不可能に分類するため，括弧表現の様々な特徴を SVM で統合した分類器を設計し，95.7%の正解率が得られた．今回は新聞記事をコーパスに用いたが，今後は大規模な Web 文書に本手法を適用し，用語のバリエーションを自動獲得する実験を行う予定である．また，本研究のアプローチは,「○○こと××」のような，括弧以外で表現される別称を扱うことができるので，様々な言語パターンを利用も検討していきたい．

*1 http://www.csie.ntu.edu.tw/~cjlin/libsvm

参考文献

[Harris 54] Harris, Z. S.: Distributional Structure, *Word*, Vol. 10, pp. 146–162 (1954)

[Kudo 02] Kudo, T. and Matsumoto, Y.: Japanese Dependency Analysis using Cascaded Chunking, in *Proceedings of the CoNLL 2002 (COLING 2002 Post-Conference Workshops)*, pp. 63–69 (2002)

[Lee 01] Lee, L.: On the Effectiveness of the Skew Divergence for Statistical Language Analysis, in *Artificial Intelligence and Statistics 2001*, pp. 65–72 (2001)

[岡崎 07] 岡崎 直観, 石塚 満：言い換え可能な括弧表現の抽出法, 言語処理学会第 13 回年次大会, pp. 911–914 (2007)

[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)

[菅野 05] 菅野 紘平, 横山 昌一, 西原 典孝：丸括弧解析システムの構築, 言語処理学会第 11 回年次大会, pp. 1217–1220 (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)