テキストからの語彙的換言知識の獲得

山本 和英
kazuhide.yamamoto@atr.co.jp
ATR 音声言語コミュニケーション研究所

概要

非換言テキストコーパスを用いた内容語に関する換言知識の自動獲得手法を提案する。局所的な文脈の類似性を利用することによって換言可能性を計算し、外部知識とヒューリスティックによる２段階の候補絞り込みを行なうことで獲得精度を向上させる。検証実験を行ない、手法の特徴と問題点を述べる。

1 研究の動機

近年、換言処理の重要性が徐々に認識されてきており、換言処理の研究も活発になってきた。換言処理には様々な種類の表現変換が含まれ、対象とする現象によって必要な情報や処理は異なる。本研究では局所的な換言を対象にし、特に内容語の換言知識構築を試みる。機能語の換言や文構造変換のような換言現象とは異なり、内容語の換言現象は大量に見ることができるため、換言知識の獲得の自動化または効率的な作成が望まれる。

換言知識を自動獲得する研究はすでに始まっている (例えば [Bar01, Kat99]) が、これらの多くは換言コーパスを使用した研究である。ただし現状で整備された換言コーパスは [Shi01, Zha01] などを除いてほとんど存在しないため、各研究においては、同一ニュースが複数のマスコミから配信されるニュース記事や、複数の翻訳版がある著名な物語のような特殊な状況を利用して擬似的な換言コーパスとしている。しかし、(擬似的なものも含め) 換言コーパスは今後とも多種で大量に得られる可能性はほとんどないため、我々はテキストコーパスからの換言知識抽出の可能性 (と限界) も検討する必要がある。我々は以前に一部の連体修飾表現を対象にして換言知識の抽出を試みた [Kat00] が、対象を他の表現にも広げなければならない。

以上の動機により、本稿では内容語換言知識を獲得することを目指す。この際、現在最も一般的に入手可能な新聞記事コーパス１年分を言語資源とする。また、換言知識獲得の際に、本研究では方向性を持たせた換言知識の獲得を行なう。すなわち、従来の研究で得られる換言表現対は片方向のみのものが多く、得られた換言知識が逆方向に適用可能かどうかについてはあまり議論されていない。そこでこの点にも注意して換言知識獲得を目指した。

2 換言可能性と同義性

換言可能とは、何らかの意味で異なっている表現を置換可能だと判断することであると考える。すなわち、二つの表現が「同じ (もしくは似ている)」と判断することでは必ずしもない。もちろん、意味の類似性と表現の置換可能性はかなり高い相関を持つのは明らかであり、例えば黒橋らの研究 [Kur99] は同義性に着目した換言処理の試みである。一方、換言という観点ではより重要なのは同義性／類義性ではなく、置換可能性であると考える。例えば、同義表現であっても、古語や幼児語などへの換言が可能かどうかは文体、文脈など当該表現の使われ方に依存するはずで、同義であるからといって換言可能とは限らない。

その一方で、同義性にこだわらず置換可能なものは換言知識と考える。例えば、上位概念語の一部なども本研究の収集対象に含める。同義でないこれらの語は、元の語に対して何らかの情報が欠落または変化しているため、厳密な意味では換言可能とは考えにくい。しかし、現実に我々は同義性のない換言をごく自然に行なっており、これらは (1) 文脈や状況、知識などから補完可能である (2) 欠落や変化した情報は文中において大きな意味を持たないため問題とならない、という理由で換言可能となるのではないか。このような、同義性のない換言も有用であると考えたため、本研究では換言可能性をできるだけ広く捉え、収集を目指した。

3 手法と実装

換言知識抽出のための具体的な提案内容を述べる。実際には、以下の作業はすべて Perl を用いて実装した。

3.1 文脈の収集

まず、本研究での「文脈」を定義する。本研究では、内容語に関する文脈は文法的直接依存関係であると考えた。すなわち、ある内容語 c に対して、c が直接係る語、および c に直接係る語を、内容語 c の文脈と定義する。

この定義に従い、依存関係をコーパス (毎日新聞 1995 年全文) から収集した。まず、形態素解析器 JUMAN[1] と構文解析器 KNP[2] を用い、内容語 c_1 がある関係 r によって内容語 c_2 に係っているすべての事例を収集した。なお、以下ではこれを (c_1, r, c_2) からなる三項で記述する。実際に収集した三項の分類とその例は以下の通りである。

- (名詞, 関係, 名詞) : 「今度の法律」「テロ法案」
- (形容詞, 関係, 名詞) : 「新たな法律」
- (名詞, 関係, 動詞) : 「衆議院が可決する」
- (動詞, 関係, 名詞) : 「空爆する米軍」

ここで、三項中の「関係」とは、格関係や助詞「の」で結ばれる関係のように助詞による修飾関係もあるが、動詞や形容詞による連体修飾関係や複合名詞の場合のように内容語間に助詞 (機能語) が存在しない場合がある。この場合は構成素境界 [Fur99] という考え方を導入し、架空の機能語 (例えば $\langle nn \rangle$) が挿入されていると考え、助詞と同様に扱った。

3.2 二部グラフの作成

次に、三項の集合を二部グラフに変換する。まず、三項を内容語と関係子 (内容語に対する修飾語、もしくは被修飾語) の二つ組として、それぞれの内容語に対して変換する。ここで、関係子とは内容語 c と方向性を持った関係 r によって、$r \rightarrow c$ (r によって c に依存している) もしくは $r \leftarrow c$ (r によって c が依存している) と規定される。例えば、ある三項関係が (c_1, r, c_2) とすると、c_1 と c_2 の各内容語に対して、$(c_1, r \rightarrow c_2)$ と $(c_2, r \leftarrow c_1)$ の二つ組を抽出する。

これをすべての三項に対して行なうことで、内容語集合と関係子集合を二つの節点集合とし、前述のすべての二つ組を枝集合とする (重み付き) 二部グラフを作成する。ここで、各枝の重みは当該の事例が出現した頻度である。

3.3 換言可能性の算出

次に、換言可能性の定式化を行なう。本研究では、二つの内容語 c_i と c_j に対する換言可能性 P を以下のように定義する。

$$P(c_i, c_j) = \frac{\sum_{m \in M(c_i) \cap M(c_j)} p(m, c_i)}{\sum_{m \in M(c_i)} p(m, c_i)} \quad (1)$$

$$p(m, c_i) = \frac{f(m, c_i)}{\sum_c f(m, c)} \quad (2)$$

$$M(c_i) = \{m | f(m, c_i) > 1\} \quad (3)$$

ここで、$f(m, c)$: 内容語 c が 関係子 m との依存関係で出現した頻度である。

本研究では換言可能性を以下のように考えた。内容語 c_i と内容語 c_j の換言可能性は、それぞれの内容語の依存関係の振舞いが類似するほど換言可能性が高いとした。すなわち c_i と依存関係を持つ関係子集合のうち c_j とも依存関係を持つ関係子がどの程度存在するかによって定義した。ここで、各関係子は異なる重要性を持つと考え、式 (2) に定義する重みを付加した。また、低頻度の事例は偶然性が高く、内容語と関係子を本質的に結び付ける手がかりとはならないと考え、式 (3) に示すように頻度 2 以上のもののみを関係子集合とした。

式 (1) の定義より明らかなように、$0 \leq P(c_i, c_j) \leq 1$ である。また $P(c_i, c_j) \neq P(c_j, c_i)$ であり、換言可能性は換言の方向性によって異なることがわかる。

3.4 換言知識の絞り込み

以上の定式化によってすべての内容語対に対して式 (1) によって換言可能性が計算できる。しかし、ここまでに述べた手法は内容語のコーパスにおける語の振舞い (文脈) のみを情報源として換言可能性を計算しているため、換言可能性が高いと判断される単語対の中に、換言とはなり得ない、例えば以下に示す関係も原理的に含まれてしまう。

1. 数詞 : 「三 → 四」など
2. 固有名詞 : 「東京 → 大阪」など
3. 対義語、対照語 : 「右 → 左」など

しかし、以上の関係は別の言語情報を使用することで排除可能なので、ここで絞り込みを行なった。固有名詞辞書としては日本語語彙大系[3] の固有名詞辞書 (169682 項目) を、対義語、対照語辞書としては、学研国語大辞典と角川類語新辞典 (両者の合計で 11981 項目) を使った。

3.5 換言語数によるヒューリスティック

最後に、ヒューリスティックを用いた換言知識の絞り込みを行ない、最終的な換言知識を確定する。

前節までに絞り込まれた内容語対を観察したところ、次のような傾向が見られた。すなわち、一部の語は、多くの語へ換言可能、あるいは多くの語から換言可能と判断された。このような語は、主に頻出語や意味の広い語であり、非常に多くの語に対して依存関係

[1] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[2] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[3] http://www.kecl.ntt.co.jp/icl/mtg/resources/GoiTaikei/

表 1: 換言知識の評価

	単一換言先	単一換言元	合計
抽出総数	668	1149	1684
うち正解	422	780	1117
正解率	63.2%	67.9%	66.3%

表 2: 換言可能性の高い知識

換言知識	P	P (←)
逸話 → 話	1	.0015
したてなげ → うわてなげ	1	.2539*
かぎり → 限り	1	.1877*
図式 → 構図	.9982	.2671*
パニック → 混乱	.9978	.0496
勝ち → 勝ち[4]	.9802	.5176*
ホッケー → 野球	.9752	.0286
結党 → 結成	.9672	.0449
違和感 → 痛み	.9667	.0352
激変 → 変化	.9582	.0177

を持つことが可能である。このため、実際は換言可能でないにも関わらず多くの語に対して換言可能性 P が高い値となってしまうのではないかと考えた。

そこで、獲得精度の向上を目的として、以下の絞り込みを行なった。ある語 c_i について、式 (1) の換言可能性が $P(c_i, c_j) > Th$ を満たす語 c_j が 1 語のみである場合 (単一換言先)、最終的に c_i は c_j へ換言可能であると判定する。同様に、ある語 c_j について、式 (1) の換言可能性が $P(c_i, c_j) > Th$ を満たす語 c_i が 1 語しかなかった場合 (単一換言元) においても、最終的に c_i は c_j へ換言可能であると判定する。なお、以下の検証実験においては $Th = 0.1$ とした。

4 検証実験

4.1 内容語の換言知識の獲得

表 1に評価結果を示す。表における「単一換言先」「単一換言元」は 3.5節における二つの場合を示す。本手法による換言可能性の計算と絞り込みの結果得られた全換言知識について、実際に換言可能かどうかを検討したところ、66% 程度の正解率で換言可能な知識であることがわかった。

次に、表 2に、得られた換言知識の中で換言可能性の高かった上位を示す。表中の P (←) は逆方向の換言可能性を参考として示し、そのうち * は双方向で換言知識として採用されたことを意味する。表では「したてなげ → うわてなげ」のような換言誤りも存在するが、概ね換言可能な知識を抽出できていることがわかる。同時に、逆方向の換言可能性が極端に低い例もあり、例えば多くの場合に「逸話」を「話」に換言できても逆方向に換言することは困難である、という我々の直感が算出値に反映されていることが確認できる。

双方向に換言可能と判定された場合は狭義の換言知識とも考えられるが、本実験では 114 対がこれに該当した。このうち、正解 (双方向で換言可能な事例) は 75 対 (65.8%) であった。これらの一部を付録に示す。

誤りと判断されたものの多くは、何らかの意味的な類似性はあるものの換言が不可能である事例であり、これ以外は「投票率 → 気温」[5]など少数である。観察の結果、誤りの原因は (1) 同類語「土曜 → 日曜」「チェロ → ピアノ」(2) 下位語「建造物 → マンション」(3) 反対語の排除もれ「アップ → 低下」(4) 固有名詞の排除もれ「イチロー → オマリー」(5) その他「休日 → ボランティア」と分類でき、(5) 以外では (1) が最も多いようである。

原因 (1) を排除するには (曜日名、楽器名などと) 一般化するための語彙集を入手することが必要である。一般的には、シソーラスがこの種の言語資源として使用されるが、これをそのまま使用するだけでは換言知識そのものも排除されてしまう。例えば、仮にシソーラスから楽器リストが入手できたとしても、排除したい換言知識「チェロ → ピアノ」と獲得したい換言知識、例えば「バイオリン → ヴァイオリン」の弁別ができない。よって弁別性を上げるためにはさらに別の情報を得る必要があり、容易ではない。

4.2 関係子の換言知識の獲得

3.2節において、二部グラフは主従関係を逆にしても二部グラフである。すなわち、これまでは内容語と内容語の換言可能性を議論してきたが、全く同一の二部グラフから関係子と関係子の換言可能性を計算することも可能である。本研究ではこれについても検証した。実験条件は前節と同一である。

抽出実験を行なった結果、432 個の換言知識が得られ、うち正解は 312 個 (72.2%) であった。表 3に、絞り込み後のうち換言可能性の高かった換言知識を示す。得られた換言知識は、前節での抽出結果に助詞を付加しただけの知識も多く存在する。しかし、助詞「の」の削除や挿入 (「⟨nn⟩ 処理 → の 処理」) 助詞の置換 (「が 長い → の 長い」)、格変換 (「が 当たる → を 当てる」) 漢語と和語の置換 (「を 禁止 → を 禁ずる」) 動詞型連体修飾から形容詞型連体修飾への換言 [Kat00](「代表する ⟨vn⟩→ 代表的な ⟨an⟩」) 以上の混合 (「から 向かう → を 出発」) など、様々な種類のものが存在し、興味深い。

[4]サ変名詞「勝ち」から普通名詞「勝ち」への換言。JUMAN では両品詞が存在するため同表記でも別単語扱いされる。

[5]数値による修飾表現の類似性が高いため換言可能性が高くなった可能性がある。

表 3: 換言可能性の高い知識

換言知識	P
依頼 が → 注文 が	1
理事 を → 教授 を	.9940
支部 で → 地裁 で	.9334
高裁 で → 地裁 で	.8734
⟨nn⟩ 短期 →⟨nn⟩ 短大	.8723
⟨nn⟩ ヤ →⟨nn⟩ 巨	.8553
毎週 ⟨nn⟩→⟨nn⟩ 夜	.8123
市議 ⟨nn⟩→ 県議 ⟨nn⟩	.8063
十数 ⟨nn⟩→ 数 ⟨nn⟩	.7961
県議 ⟨nn⟩→ 市議 ⟨nn⟩	.7859

5 関連研究

非換言コーパスからの換言知識獲得を Jacquemin et al.[Jac97] が行なっている。ここでは複数単語 (例えば gene expression) を対象にして派生的表現 (genic expressions) や語順入れ換え (expression of this gene) などに伴う多様な表現の収集を目指している。

本研究では語彙単位での換言知識の獲得を目的とした。この際、コーパスにおける文脈の類似性を用いたが、この考え方は語の意味距離計算、すなわち 2 語の意味的な類似性を計算する際に用いられる考え方と基本的に同一である[6]。このような研究は例えば [Nag96, Kan00] などがあるが、本研究とは以下の点で重要な相違がある。まず、類似度には対象性があるが、換言可能性は換言方向によって異なる (3.3節)。次に、換言知識獲得問題は類似度計算よりも絞り込みなどによってより条件を厳しくしなければならない (3.4, 3.5 節)。また係り受け情報として [Nag96] では (EDR 辞書における) 深層格を、[Kan00] では名詞に係る連体修飾成分のみを使用しているのに対し、本稿ではコーパスから機械的に抽出可能である表層的な係り受け関係をできるだけ抽出している。ただし多様な情報を使用したことによる功罪は検証できていない。

6 結論と今後の課題

内容語及び内容語に対する修飾／被修飾語に関する換言知識をテキストコーパスから自動収集する手法を提案した。文脈の類似性を利用した換言可能性計算と二段階の候補絞り込みを行なうことで、1700 個程度の内容語換言知識を獲得し、その正解率は 66% であった。同時に、内容語に係る修飾／被修飾語の換言知識獲得も行なった結果、400 語程度の知識が得られ、その精度は 72% であることを示した。

[6]同様に、シソーラスの自動構築や未知語の意味推定などの意味処理においても利用される考え方である。

重要な課題は正解率を維持したままの網羅性の向上である。すなわち、換言知識絞り込みで排除されたものの中に、まだ多くの有用な換言知識が残されている。しかし、絞り込みの条件を緩くすると正解率が急激に低くなることが予想され、収集の効率は大幅に低下する。今後は、他の絞り込み条件を見つけるなどの方法でより網羅的な収集をしなければならない。

本研究は通信・放送機構の研究委託により実施したものである。

参考文献

[Bar01] BARZILAY, R. and MCKEOWN, K. R.: Extracting Paraphrases from a Parallel Corpus, In *Proc. of ACL-2001*, pp. pp.50–57 (2001).

[Fur99] 古瀬蔵, 山本和英, 山田節夫：構成素境界解析を用いた多言語話し言葉翻訳, 自然言語処理, Vol. 6, No. 5, pp. 63–91 (1999).

[Jac97] JACQUEMIN, C., KLAVANS, J. L., and TZOUKERMANN, E.: Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax, In *Proc. of ACL-EACL'97*, pp. 24–31 (1997).

[Kan00] KANZAKI, K., MA, Q., and ISAHARA, H.: Similarities and Differences among Semantic Behaviors of Japanese Adnominal Constituents, In *Proc. of ANLP/NAACL 2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing System*, pp. 59–68 (2000).

[Kat99] 加藤直人, 浦谷則好：局所的要約知識の自動獲得手法, 自然言語処理, Vol. 6, No. 7, pp. 73–92 (1999).

[Kat00] 片岡明, 増山繁, 山本和英：動詞型連体修飾表現の"N1 の N2"への言い換え, 自然言語処理, Vol. 7, No. 4, pp. 79–98 (2000).

[Kur99] 黒橋禎夫, 酒井康行：国語辞典を用いた名詞句「A の B」の意味解析, 情報処理学会研究報告 NL129–16, 情報処理学会 (1999).

[Nag96] 永松健司, 田中英彦：コーパスから抽出した係り受け共起情報に基づく類似度と文書検索における評価, 研究報告 NL116-11, 情報処理学会 (1996).

[Shi01] SHIRAI, S., YAMAMOTO, K., and BOND, F.: Japanese-English Paraphrase Corpus, In *Proc. of NLPRS2001 Workshop on Language Resources in Asia*, pp. 23–30 (2001).

[Zha01] ZHANG, Y., YAMAMOTO, K., and SAKAMOTO, M.: Paraphrasing Utterances by Reordering Words Using Semi-Automatically Acquired Patterns, In *Proc. of NLPRS2001*, pp. 195–202 (2001).

付録：双方向に換言可能と判定された内容語の例

(GDP, 国内総生産) (まなざし, 視線) (キレ, 切れ) (コンクール, コンテスト) (スキャンダル, 不祥事) (スト, ストライキ) (スポークスマン, 高官) (データ, 情報) (トップ, 首位) (ドア, 扉) (リレー, 継投) (慰霊, 追悼) (会社, 企業) (刊行, 出版) (棄却, 却下) (気持ち, 思い) (規制, 制限) (救出, 救助) (橋げた, 高架) (欠点, 弱点) (構図, 図式) (告訴, 告発) (参政権, 選挙権) (質疑, 討論) (守り, 守備) (修復, 補修) (侵攻, 進攻) (深夜, 未明) (診察, 診療) (選手権, 大会) (対決, 対立) (俳優, 役者) (賠償, 補償) (発覚, 表面化) (発射, 発砲) (批判, 非難) (疲れ, 疲労) (被告, 容疑者) (平年, 例年) (利子, 利息) (了解, 了承) (料, 料金)