語釈文を利用した普通名詞の同概念語への言い換え

藤田篤 [*1]　乾健太郎 [*2] [*3]

[*1] 九州工業大学大学院情報工学研究科

[*2] 九州工業大学情報工学部知能情報工学科

[*3] 科学技術振興事業団さきがけ研究 21「情報と知」領域

{a_fujita,inui}@pluto.ai.kyutech.ac.jp

1 はじめに

ある言語表現をできるだけ意味を保持したまま別の言語表現に変換する「言い換え」の技術は，自然言語処理の諸分野において様々な応用が考えられる重要な要素技術である [16]．たとえば，翻訳や要約などの前処理として言い換えを行う試みはすでにいくつか報告されており，日本語・手話翻訳や音声合成といった異なるタスクの前処理にも利用できる可能性がある．また，要約や推敲支援，文章読解支援のように言い換え技術が根幹をなす応用もあり，それぞれの文脈の中で言い換えの実現を目指す研究も見られるようになってきた [2, 6]．最近では，言い換え技術を応用からある程度独立した要素技術として捉え，その性質や実現方法を解明する試みもいくつか見られる [11, 16, 12]．しかしながら，日英翻訳のような言語間翻訳が長年精力的に研究されてきた経緯と比べると,「単言語内翻訳」である言い換えの研究はまだ極めて萌芽的な段階にあると言わざるをえない．

このような背景から，我々は「要素技術としての言い換え」の事例研究として，連体修飾節に着目した構文的言い換え [14] や，文中の内容語を別の語句に言い換える語彙的言い換えの研究 [5] を進めている．本稿では，語彙的言い換えの例題として普通名詞を同概念語へ言い換えるタスクを取り上げ，問題の性質と取り組むべき課題を論じる．

2 アプローチ

言い換えは，ある言語表現を同じ意味を持つ別の言語表現に変換する作業であるが，完全に同義の言い換えはほとんど存在せず，一般に何らかの意味の変化が生じる．とくに，単体で指示的意味 (denotation) を持つ内容語を言い換える場合は，この意味の差を慎重に考慮する必要がある．たとえば,「格差」と「落差」は，EDR 日本語単語辞書 [4] によると同概念に属するが，厳密には異なる指示的意味を持つため，互いに言い換え可能かどうかは文脈に依存している．たとえば，(1) の「格差」を「落差」に置き換えることはできない．

(1) 二人の賃金に 格差 をつける。

このことから言い換えは，

i) 言い換え前後の言語表現 (言い換え対) の間の意味の差 (意味差分) を計算し，
ii) その意味差分が所与の文脈に照らして無視できるかどうかを判断する

作業と考えることができる．意味差分の記述については，Edmonds が類義語 (near-synonym) の意味を記述するオントロジを開発し，機械翻訳における語選択に用いる方法を示している [3] が，この知識を安価に獲得・記述する方法については白紙の状態である．したがって，本研究では既存のシソーラス・辞書といった言語資源を利用し，先に挙げた言い換えを機械的に実現するため，(1) 意味差分の獲得方法，(2) 所与の文脈における言い換えの可否について，次のようなアプローチをとる．

2.1 共起情報と語釈文を用いた意味差分の獲得

既存の言語資源から意味差分を取り出す方法としてまず考えられるのは，語の共起情報の利用である．共起情報を利用した統計的単語クラスタリングに関する先行研究が示すように，共起情報は語の意味的類似度の推定にある程度は有効に働くため，意味差分の獲得においても有効性を期待できる．

ただし，共起情報だけでは類義語間の微妙な意味の差を計算することができないため，二つの語の語釈文を比較することによってそれらの意味差分を計算する方法を考える．たとえば岩波国語辞典 [15] は,「格差」と「落差」という類義語の各々に次のような語釈文を与えている．

「格差」価格・資格・等級の差。
「落差」落下または流下する水の、高低二か所における高さの差。転じて一般に、高低の差。

これを比較すると，語釈中の共通語である「差」の修飾語句「価格・資格・等級の」「落下または流下する水の，高低二か所における高さの」から，たとえば「差を評価する対象の違い」という知識を両者の意味差分として自動獲得できる可能性がある．

2.2 文脈における言い換えの可否

次に考えるべき課題は，意味差分と文脈における言い換えの可否の関係である．すでに多くの指摘があるように，二つの語の意味が完全に一致することは，表記のゆれなどを除けばほとんどない [3]．ただし,「コンピュータ」と「計算機」のように，形式性などの connotation の違いを無視すれば，ほぼ同義と考えられる対 (言い換え対) は少なくないと考えられる．本研究では，語彙的言い換えを目的としているため，connotation の違いには重点をおかず，ある一定のプロセスに従えば denotation が一致していると判定できる場合は，言い換え可能とみなす．ここで，二つの語の間の意味の重なり方というのは，必然的に図 1 の 5 種類のいずれかに分類できる．

図 1 の各パタンと言い換え可否の対応について，次のような仮説を立てることができる．

- (A) の関係にあれば言い換え可能である
- (E) の関係にあれば言い換え不可である
- (B)，(C)，(D) の関係にあれば，言い換えの可否は文脈に依存する
  - 可能な場合は，文脈上の S，T の意味が図の重なる部分 (共通部分) を指している
  - 不可の場合は，文脈上の S，T の意味が図の重ならない部分 (差分) を指している

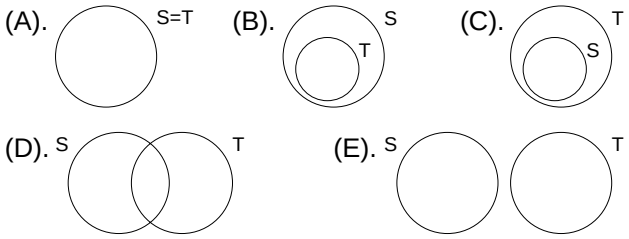

図 1: 言い換え対の意味の重なり方

3 語釈文の比較

言い換え対間の意味の重なり方を推定するため，語釈文の比較，差分抽出を自動的に行う方法について述べる．

3.1 語釈文からの要素抽出

ある見出し語に対する語釈文，例えば「経緯」について岩波国語辞典 [15] では次のように示されている．

「経緯」縦糸とぬき糸（＝横糸）。また、たてよこの方向。 [1] 物事のいきさつ。細かい事情。▽「経」は織物の縦糸、「緯」は横糸の意。[2] 経度と緯度。

このように，一般に国語辞典の語釈文は，階層化された複数の要素から構成されている場合が多く，また語釈文特有の表現もしばしば含まれており，これに対処する必要がある．たとえば，上記語釈文中，「[1]」,「[2]」の大区分で示される語の多義性 (ambiguity) については，共起制約を用いる際に自然に淘汰できると期待できる．しかし，句点で区切られた各要素の曖昧性 (vagueness) については，辞書の作成者の区分の意図を推定することは難しい．そこで今回は，語釈文を構成する複数の要素間の関係までは扱わず，語釈文の表記について以下の 2 点に注意して，語釈文を複数の要素のリストに展開した．

- 「α の転」,「α の略」,「α の謙譲語」，などの connotation の違いを説明する記述は考慮せず，単に「α」を取り出す．
- 「または」,「とくに」などの要素先頭の接続詞や副詞,「α・β」,「α や β」などの並列記述に注目して要素を展開する．

3.2 語釈要素の比較

計算機によって意味を取り扱うことを目的として，語釈文に含まれる種々の特徴を分類し，単語間の相違点を自動的に取り出す研究は，すでに文献 [17] などでなされている．今回は我々も，単に文間の類似度を算出するのではなく，言い換えの可否の判断材料として差分を取り出すために，文献 [17] で用いられた語釈文の抽象化手法を参考にし，以下の手順で語釈文の照合を行う．

i) 語釈文を構成する各語釈要素を構文解析し，自立語をノード，付属語をアークとするグラフを作成する．
ii) 語釈要素の全ての組合せに対して，図 2 に示すようにグラフを照合し，< 照合パタン，照合範囲，共通特徴 f_c，個別特徴 f_s，f_t> の 5 つ組からなるサマリを作成する．得られたサマリのリストを語釈文の照合のサマリとする．照合パタンは，以下の 6 種類とした．
   (a) 語釈要素が完全一致
   (b) 一方の語釈要素が他方の語である
   (c) 一方の語釈要素が他方の語を限定
   (d) 一方の語釈要素が他方の語釈要素を限定
   (e) 語釈要素のうち主辞と修飾語が一致
   (f) 語釈要素の主辞のみ一致

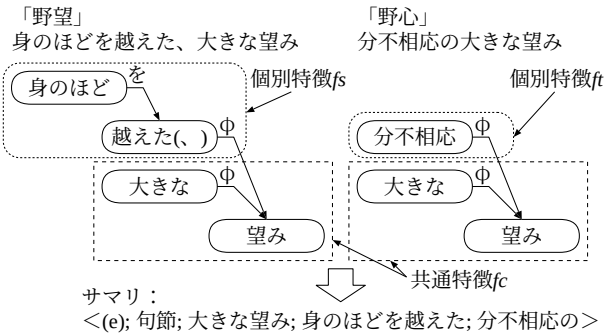

図 2: 語釈要素の照合とサマリ

4 言い換え可否の判定実験

京大コーパス [13] 中の普通名詞を EDR 日本語単語辞書中の同概念語へ言い換え，語釈文から得られる意味差分と所与の文脈との関係を整理した．

4.1 言い換え事例の収集

京大コーパスの全文 (38,383 文) から，以下の条件のすべてを満たす普通名詞を対象名詞として取り出した．

- 日本語基本語彙 6000 語 [9, 10] 外，かつ，文字単語親密度 [1] が 6.00 を超えない普通名詞．基本語，理解容易度の高い語は，そのままでも明確な意味を指示できるため，言い換えにより不用意に意味が変わる可能性があると考え，対象外とする．
- EDR 日本語単語辞書中で語義が一意に決まるもの．語の多義性解消は語彙的言い換えに限らず欠かせない処理だが，今回は扱わない．
- EDR 日本語単語辞書中で日本語概念説明が与えられている，ある程度具体的な語．
- 接辞または他の名詞を伴い複合語を形成する語以外．複合語の意味解析は扱わない．

ここで，取り出された 17,741 箇所 (4,941 語) の普通名詞のうち，8,579 箇所 (2,472 語) の語が同概念語をもっており，のべ 47,800 の同概念語が得られた．

次に，統語的な制約の初歩として，係り受けで出現する自立語の対とそれらの間の関係の 3 つ組の用例を共起制約として用いた．まず，EDR 日本語共起辞書 [4] 中の，修飾語，非修飾語，修飾関係の 3 つ組からなる共起用例 935,860 組で係り受けを保証される語への言い換え事例のみを取り出したところ，322 箇所 (161 語) に対する 434 事例が得られた．同様に，文献 [8] で収集された用例格フレームから，用言，格要素，格の 3 つ組からなる共起用例 2,089,142 組を抽出し，これを EDR の共起事例と独立に用いたところ，371 箇所 (215 語) に対する 639 事例が得られた．今回は，これらの和集合，575 箇所 (284 語) に対する 917 事例を対象とした．

4.2 言い換え事例の評価

生成した事例のうち，表記のゆれや略語の復元という素性を持つ言い換え対については，言い換え自体の効果が得られないと思われるため評価対象から除外し，残りの 809 事例 (479 箇所，229 語) を評価対象とした．

同概念語と共起制約を用いることによって生成された言い換え事例は，統語的・意味的制約をある程度満たしていると考えられる．ここで，人手により統語的・意味的適格性の評価を行った．この際，語釈文などのテキスト外の知識は参照せず，意味の差分については人間の主観で判断した．2 人の評価者が独立に評価した結果，

共通して言い換え可能，不可と判定された正負例が各々 419 事例，327 事例となった．2 人の評価が食い違ったものが 20 事例，これ以外の 43 事例は，主観による適格性の評価が困難な事例であったため，今回は評価を保留した．一方，3 節で述べた手順で，岩波国語辞典 [15] および角川新類語辞典 [7] の語釈文を各々比較し，得られたサマリのリスト中で最も一致度の高いものを言い換え対の語釈文の照合結果とした．両評価の対応を表 1 に示す．

4.3 負例となる原因の考察

表 1 から，語釈要素が一部分でも共通している事例は，正例となる率が高いことが分かる．すなわち，語釈文の照合が言い換え可否判定の手がかりの一つとなり得るといえよう．しかし，語釈文がある程度一致していても負例となる場合があるので，その原因について考察する．

4.3.1 一方のみが個別特徴を持つ場合

まず，個別特徴が一方にのみ出現する，照合パタン (c)，(d) について考察する．これらは，語釈要素が他方の語または他方の語釈文を限定するパタンであり，限定される側の語が共通特徴 f_c となる．すなわち，S，T いずれかにとって f_c は S，T それ自体であると同時に S，T の共通特徴であり，f_c を限定する語句は S あるいは T の個別特徴である．(c)，(d) は，語対の意味差分が語釈文に顕在化された照合パタンであるといえる．ここで，下の 3 点に着目して負例となる原因を整理した．

- 共通特徴 f_c が曖昧性 (vagueness) を持つか．
- 個別特徴 f_s，f_t の意味を文脈がカバーできるか．
- 文脈上 S と T が異なる意味を持ってしまうか．

まず，f_c の曖昧性に着目する．3 節で述べたように，語釈文には多義性を表す大区分と，微妙な意味の差を表す小区分がある．この小区分内に f_c が，他の語釈要素を持つ場合，f_c とこれらの関係が重要になる．これら他の語釈要素と f_c とが異なるものを指し得る場合，我々はこれを vagueness と呼ぶ．例えば，「因子」の岩波国語辞典の語釈分は，次のような微妙に異なる複数の意味を持つ．

「因子」ある結果をひき起こすもとになる要素。現象の要因を構成している作用素または力。ファクター。

ある S を「因子」に言い換えた場合，「因子」は，”要素”，”作用素”，”ファクター”のいずれを指すのかが不定になる．すなわち，限定力を失うと考えられる．ここで次に，他方の語の持つ個別特徴 f_s，f_t の意味を文脈情報から補えるかどうかが重要となる．たとえば次の言い換え事例をみてみよう．

(2) チーム力に 格差 ができることは否めない。
→ . . . チーム力に 差 が . . .
語釈文の照合のサマリ (岩波) :
<(c); 語; 差; { 価格, 資格, 等級 } の; ϕ>

(3) 同病院の医師らはこの働きを利用し、遺伝子 を抑えることで、がんの成長を抑制できるとみている。
→ *. . . 因子 を抑えることで、. . .
語釈文の照合のサマリ (岩波) :
<(c); 語; 因子; 生物の遺伝形質を規定する; ϕ>

(2) では，文脈から「チーム力の差」という意味を取り出すことによって，f_c の「差」に対して「格差」との意味差分である「価格，資格，等級の」に相当する情報を補える．しかし，(3) のように，文脈が f_c の持つ意味を必ずしも f_s と同じように特定するとは限らない．(3) では，「遺伝子」を「因子」に言い換えることによって f_s の「生物の遺伝形質を規定する」という限定要素がなくなるため，何によって「がんの成長を抑制できる」のかが不明になるのである．

しかし，(4) に見られるように，文脈から補える情報が必ずしも f_c を限定する f_s に一致するとは限らない．

(4) 出足を止められると足がそろいがちで、突き押しも腰が入らず 威力 を失う。
→ *. . . 突き押しも腰が入らず 力 を失う。
語釈文の照合のサマリ (角川) :
<(c); 語; 力; 他を威圧する; ϕ>

(4) では，「突き押しの程度」から f_s「他を威圧する」力がなくなることが表現されていたが，「威力」を「力」に言い換えることにより，「力士」から「力が抜ける」という意味になってしまう．

上記 3 点の組合せについて考察した結果，図 3 の分類木のようなものを考えれば正負例を弁別することができる可能性があることが分かった．図 3 は，(c)，(d) の中の，S が，T または T の語釈要素を限定している事例に対する正負例の分類木である．

- **$T(=f_c)$ の指示対象が vague であるか** :
  - vague でない.
    - **文脈が f_s と違う意味で f_c を限定しないか** :
      - 違う意味で限定するのであれば負例.
      - 同じ意味で限定するのであれば正例.
  - vague である.
    - **文脈は f_s の限定力を補えるか** :
      - 補えれば文脈上 f_s は無視できる.
        - **文脈が f_s と違う意味で f_c を限定しないか** :
          - 違う意味で限定するのであれば負例.
          - 同じ意味で限定するのであれば正例.
      - 補えなければ文脈上 f_s は削除不可なので負例.

図 3: 事例より得られた分類木

今回の実験では，S を収集する際の制約から，T が，S または S の語釈要素を限定している事例はほとんど得られなかったが，同様にして分類できると考えられる．

4.3.2 個別特徴を持たない場合

次に語釈要素が完全一致する，照合パタン (a)，(b) について考察する．該当する負例には以下のようなものがある．これらの照合パタンは (c)，(d) とまったく異なるが，負例の原因については同じ観点で整理できることがわかってきた．

(5) 日中両国政府は . . . 合意し，北京で 書簡 を交換した．→ *. . . 合意し，北京で 消息 を交換した．
語釈文の照合のサマリ (岩波) :
<(b); 語; { 消息, 手紙, 書状 }; ϕ; ϕ>

(c)，(d) の場合，個別特徴と共通特徴が語釈要素中に顕在化していたため，共通特徴 f_c 側の語の語釈文全体をもって言い換え語以外の語釈要素の vagueness を考えることができた．しかし，(a)，(b) の意味差分は明示的には語釈に現われないため，(c)，(d) のようにして，(5) のような負例の原因を説明することができない．そこで例えば，f_c も含めた語釈要素を個別特徴 (「書簡」「消息」「書状」「たより」) とみなして「手紙」のような共通特徴を取り出すか，あるいは「書簡」「消息」などから個別特徴として「公式的な」「形式自由な」といった情報を，さらに語釈要素の語釈文から取り出すなどすることにより，(c)，(d) と同じ枠組で計算できると考えられる．

表 1: 人手による言い換え可否の判定と語釈文の照合による判定の照合

照合パタン		岩波国語辞典				角川新類語辞典			
		可能	不可	合計	正解率 (%)	可能	不可	合計	正解率 (%)
(a)	語釈要素が完全一致	24	18	42	57.1	10	3	13	76.9
(b)	一方の語釈要素が他方の語である	46	19	65	70.8	11	9	20	55.0
(c)	一方の語釈要素が他方の語を限定	24	9	33	72.7	33	6	39	84.6
(d)	一方の語釈要素が他方の語釈要素を限定	8	0	8	100.0	2	0	2	100.0
(e)	語釈要素のうち主辞と修飾語が一致	13	5	18	72.2	11	1	12	91.7
(f)	語釈要素の主辞のみ一致	28	19	47	59.6	26	15	41	63.4
(x)	全要素間に共通特徴なし	114	133	247	46.2	144	150	294	49.0
(y)	一方の語釈文がなく，共通特徴なし	132	109	241	54.8	140	134	274	51.1
(z)	両方の語釈文がない	30	15	45	66.7	42	9	51	82.4
(a) から (f) の合計		143	70	213	67.1	93	34	127	73.2

4.3.3 両者が個別特徴を持つ場合

最後に，語釈文の一部が共通であり，S，T がともに個別特徴を持つ，照合パタン (e)，(f) について考察する．以下のような事例が該当する．

(6) 寄付行為の適用が 社交 の範囲にまで拡大されて以来、同容疑での逮捕者は初めて。
→ *...、寄付行為の適用が 外交 の範囲にまで ...

語釈文の照合のサマリ (岩波)：
<(f); 語; 交際; 社会上の; 外部との >

(6) では，f_t が「外部との」であるため,「日本における公職選挙法の適用範囲」を指す f_s「社会上の」との意味差分が生じる．さらに，S，T 自体が vague であり，f_s，f_t は「対外的な」または「会社間 (銀行，保険会社などの場合) の」という共通の意味も取り得る．

(7) 上杉謙信など「かっけ」に 野望 を砕かれた歴史上の人物も多い。
→ ...野心 を砕かれた歴史上の人物も多い。

語釈文の照合のサマリ (岩波)：
<(e); 句節; 大きな望み; 身のほどを越える; 分不相応の >

(7) のように，表層的な差分として得られた f_s と f_t が言い換えの関係にあるなど，(e)，(f) で負例が生じる原因は，(c)，(d) での原因がさらに複合的になっている．しかし，基本的には (a)，(b) 同様，共通の枠組で処理を考えることができる．

以上述べたように，語釈文の比較によって得られる f_c が vague であるか，f_s，f_t が文脈上どのような役割を果たすかを評価することにより，S と T の指示的意味の比較において，意味の重なりの部分を指しているかどうかの推定，しいては負例の棄却を実現できると考えられる．

5 おわりに

本稿では，語彙的言い換えの例題として，普通名詞の同概念語への言い換えの可否を，共起制約，語釈文の比較によって判定するタスクに取り組み，ある程度良質の言い換えを生成できることを確認した．また，本稿では紙面の都合で述べられなかったが，単語親密度 [1] を用い，テキスト簡単化に対する目的適合性の評価を並行して行った．今後の課題としては，語釈要素間の関係として捉え意味差分を抽出すること，語釈文から得られる意味差分と所与の文脈の照合を実装することがあげられる．

謝辞

本研究を進めるにあたり，京都大学の河原大輔氏，黒橋禎夫氏に格フレームの用例データを頂きました．また，東京工業大学の奥村学氏に貴重なコメントを頂きました．九州工業大学の乾裕子氏には言い換えの可否判定をお手伝い頂きました．ここに厚く御礼申し上げます．

参 考 文 献

[1] 天野成昭, 近藤公久. 日本語の語彙特性 1: 単語親密度. 三省堂, 1999.
[2] Dras, M. Reluctant paraphrase: Textual Restructuring under an Optimization Model. *Proc. of PACLING'97*, pp. 98–104, 1997.
[3] Edmonds, P. Semantic Representations of Near-Synonyms for Automatic Lexical Choice. Ph.D. thesis, published as technical report CSRI-399, Department of Computer Science, University of Toronto, 1999.
[4] EDR. 電子化辞書仕様説明書 第 2 版. Technical Report, 日本電子化辞書研究所, 1995.
[5] 藤田篤, 乾健太郎, 乾裕子. 名詞言い換えコーパスの作成環境. 電子情報通信学会思考と言語研究会, TL2000-32, 2000.
[6] 乾健太郎. テキスト簡単化による聾者向け読解支援 ー現状と展望ー. 電子情報通信学会福祉情報工学研究会, WIT2000-34, 2000.
[7] 角川書店. 角川類語新辞典. 1981.
[8] 河原大輔, 黒橋禎夫. 用言と直前の格要素の組を単位とする格フレームの自動獲得. 情報処理学会自然言語処理研究会, NL-140-18, pp. 127–134, 2000.
[9] 国立国語研究所. 分類語彙表. 大日本図書, 1964.
[10] 国立国語研究所. 日本語教育のための基本語彙調査. 秀英出版, 1984.
[11] 近藤恵子, 佐藤理史, 奥村学. 「サ変名詞＋する」から動詞相当句への言い換え. 情報処理学会論文誌 Vol. 40, No. 11, pp. 4064–4074, 1999.
[12] 近藤恵子, 佐藤理史, 奥村学. 格変換による単文の言い換え. 情報処理学会自然言語処理研究会, NL-135-16, pp. 119–126, 2000.
[13] 黒橋禎夫, 長尾 眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会第 3 回年次大会発表論文集, pp. 115–118, 1997.
[14] 野上優, 乾健太郎. 結束性を考慮した連体修飾節の言い換え. 言語処理学会第 7 回年次大会発表論文集, 2001.
[15] RWC. RWC テキストデータベース第 2 版, 岩波国語辞典タグ付き/形態素解析データ第 5 版. RWC, 1998.
[16] 佐藤理史. 論文表題を言い換える. 情報処理学会論文誌, Vol. 40, No. 7, pp. 2937–2945, 1999.
[17] 土屋雅稔, 黒橋禎夫. MDL 原理に基づく辞書定義文の圧縮と共通性の発見. 情報処理学会自然言語処理研究会, NL-140-7, pp. 47-54, 2000.