文脈を考慮した日本語類義表現の言い換え

岡本 紘幸　斎藤 博昭
慶應義塾大学大学院 理工学研究科 開放環境科学専攻
Email: {motch,hxs}@nak.ics.keio.ac.jp

1 はじめに

文中の内容語を他の類義語に言い換える語彙的言い換えでは，主に可読性の向上を目的として，言い換えにおける文法的整合性を保つための研究がほとんどであった．一方，文章入力時における推敲支援を目的とした語彙的言い換えに関する議論はほとんどなされておらず，これまでに研究されてきた推敲支援は文法的な複雑さを補正するもののみである [1][2].

本稿では，推敲支援を目的とする言い換えにおける類義語群の提示に焦点を当てる．入力文中のある表現について，その類義語群の間の意味差分を示すために必要な知識を獲得し，その意味的適格性から類義語群の順位付けを行う手法を提案する．

2 対象とする言い換え

まず，以下のような言い換えについて考える．

(a) 彼は激怒した.
(b) 彼は激しく憤った.
(c) 彼はひどく怒った.

これまでの言い換え研究では，(a) から (b) や (c) への言い換えについては議論がなされており，例えば「激怒」の語釈文が「激しく憤ること」であるなどの知識をもとに (a) から (b) への言い換えを実現する研究もある [4]．これに対し，(c) から (a) への言い換えについては，語釈文をそのまま用いても解決が困難な問題であり，これまで議論の対象とされていなかった．このような言い換えを実現するためには，(1) 「怒る」の類義語である「激怒」が「激しく」という指示的意味を持つといった知識をいかに獲得するか，(2)「激しく」と「ひどく」との意味的等価性をいかに判定するか，(3) 実際に言い換えを行う際の文法的制約，といった問題の解決が必要となる．本稿では，(c) のような入力文が与えられたとき，(a) のように言い換えが可能な単語を類義語群から優先的に提示するために，上記の問題のうち (1) および (2) について，以下のアプローチで解決を試みる．

(i) 語釈文と共起情報を利用して各類義語のもつ粒度の細かい意味を抽出する．
(ii) 各類義語を入力文脈との意味的適格性をもとに順位付けする．

なお，対象とする品詞は限定せず，慣用表現なども1語として考えることで言い換えの対象とした．また実際に言い換えを行う際の文法的な制約は扱わない．

3 類義語間の意味差分の自動抽出

言い換えにおいて類義語間の意味差分を扱う研究としては，言い換え対の語釈文から意味差分を抽出し，言い換えの可否判定に利用しようという試みがすでに存在する [3]．本稿ではこの方法を参考にして，類義語群の順位付けを行うため，1対の語釈文ではなく類義語群全体の語釈文から意味差分を抽出した．

3.1 語釈文の持つ情報

Edmonds は類義語間の意味差分の種類を以下の4つに大別した [5].

A. 指示的意味 (denotational): 語自体の持つニュアンスの差異.
B. 表現 (expressive): 語の持つ態度や感情の差異.
C. 文体 (stylistic): 方言や形式性などの差異.
D. 連語 (collocational): 語とよく共起する語の差異や慣用表現など.

このうち，表現および文体の差異については，その種類によって記述方法がまちまちであり，語釈文からの取得が困難である．そのため本稿では，語釈文の解析により得られた意味差分の情報を以下の2種類に分類する．

- **指示的意味:** 表現自体に含まれる表層的意味以外の意味.
- **語彙的制約:** 表現が文脈中の他の語句の選択に与える制約.

例えば,「建てる」の類義語である「再建 (さいこん)」の語釈文からは以下のように指示的意味および語彙的制約を抽出できる.

再建(さいこん) 神社・仏閣を建て直すこと.
　表層的意味: 建てる
　指示的意味: 直す
　語彙的制約: 神社，仏閣

本稿で扱う言い換え語の選択には，ある単語の指示的意味および語彙的制約を抽出する必要がある．その

ため，まず (1) 語釈文から表層的意味と意味差分を取得し，さらに (2) 意味差分を指示的意味と語彙的制約に分類するという処理によって指示的意味，語彙的制約の自動抽出を行う．

3.2 語釈文からの意味差分の自動抽出

本稿では，ある単語の語釈文は以下の 4 種類の語句から構成されると仮定する．

- **表層的意味**: その表現の属する意味分類を示す語句．
- **意味差分**: 同一の意味分類に属する他の表現との意味の差異を示す語句．指示的意味と語彙的制約に分類できる．
- **不要語**(stop word): 「～こと」「～の意」など，どの意味分類においても共通に頻出する語句．
- **その他**: 助詞などの付属語および記号類．

この仮定をもとに，角川類語新辞典 [6] に記述された意味分類および語釈文の情報を用いて，まず語釈文から表層的意味と意味差分を以下の方法により抽出した．

1. 形態素解析辞書に各形態素の属する意味分類の情報を付加し，それを用いてある単語の語釈文を形態素解析する．
2. 語釈文内にその単語と同じ意味分類に属する語句があれば，それを語釈文の暫定的な表層的意味とする．
3. 語釈文中において暫定的な表層的意味と共起する語句のうち，意味分類情報を与えられているものを意味差分として抽出する．
4. 暫定的な表層的意味を示す語句について再帰的に語釈文を辿っていき，表層的意味が得られなくなるまで意味差分を取得する．

また，この方法では不要語も意味差分として取得されてしまうため，各意味差分に以下のように重要度の重みを与えることで，意味差分と不要語とを区別する．

- 単語の語釈文から直接得られる意味差分の重要度を 1 とし，表層的意味の語釈文を辿る過程で得られる意味差分は，その過程に沿って重要度を半分にしていく．
- 辞書全体の語釈文における単語の出現確率で重要度を割る．

以上の方法により，例えば「悲憤慷慨」の意味差分および重要度が以下のように抽出できる．

悲憤慷慨 時世・運命などに対して悲しみかつ憤ること．

- 表層的意味: 憤る → 怒る
- 意味差分: 時世 (477)　こと (164)
  運命 (2253)　恨む (1127)
  悲しむ (770)

() 内が各意味差分の重要度である．

3.3 意味差分の分類

次に，得られた意味差分から類義語間の順位付けに必要な指示的意味，語彙的制約を，以下の仮定に基づいて分類する．

- ある単語が指示的意味を持つとき，その指示的意味は単語に含まれる意味であるため，その単語とは共起しない．
- ある単語が語彙的制約を持つとき，その単語と語彙的制約は特によく共起する．

抽出を行うためには，各意味差分がその意味差分を持つ単語と共起するかどうかを調べる必要がある．本稿では，毎日新聞 [7] および青空文庫 [8] 合計およそ 193 万文から共起情報を取得することにより，類義語とその意味差分との共起関係を取得した．その結果から，指示的意味および語彙的制約を以下の手順により抽出した．

1. コーパスに出現しないエントリを削除する．
2. 意味分類 C に属するエントリ e の各意味差分情報 d に対し，C 全体における d の共起確率 $P(d|C)$，および e に対する d の共起確率 $P(d|e)$ を以下の式により算出する．

$$P(d|C) = \frac{C\text{ 全体における }d\text{ の共起頻度}}{C\text{ の共起カテゴリの総数}} \quad (1)$$

$$P(d|e) = \frac{e\text{ に対する }d\text{ の共起頻度}}{e\text{ の共起カテゴリの総数}} \quad (2)$$

3. $P(d|C) = 0$ となる意味差分情報 d を削除する．
4. $P(d|e) = 0$ ならば，d 自体の重みと $P(d|C)$ との積を重みとして指示的意味に分類する．
5. $P(d|e) \neq 0$ ならば，d 自体の重みと $\frac{P(d|e)}{P(d|C)}$ との積を重みとして語彙的制約に分類する．

3.4 実験および考察

角川類語新辞典の全 2,924 カテゴリ 57,130 エントリについて以上の処理を行ったところ，語釈文から表層的意味および意味差分を取得できるエントリが 36,434 得られた．また，その他にこれらのエントリから表層的意味として参照されている 1,857 エントリについても取得した．意味差分を取得したエントリについては，3.3 節で述べた方法により意味差分の分類を行った結果，1 エントリあたり平均で 4.7 個の指示的意味，5.1 個の語彙的制約が抽出された．

本稿における自動抽出方法の評価のため，意味差分を持つとされたエントリから無作為に選択した 50 語について，(1) 表層的意味および意味差分の取得，(2) 意味差分の分類，の 2 点に対し，語釈文を人手で解析

し表層的意味，指示的意味および語彙的制約の抽出を行ったものと自動抽出結果との比較を行った．(1) および (2) の結果をそれぞれ表 1，表 2 に示す．ただし表 2 は抽出された意味差分のうち不要語でないものについての分類結果である．

表 1: 表層的意味の抽出結果

結果	エントリ数
正解	40
不正解	10
(表層的意味が異なる)	(4)
(表層的意味の経路が異なる)	(6)
精度	80.0 %

表 2: 意味差分の分類結果

		出力		再現率 (%)
		指示的意味	語彙的制約	
正解	指示的意味	56	13	81.2
	語彙的制約	22	20	47.6
精度 (%)		71.8	60.6	

意味差分を示す 2 つの情報のうち，指示的意味については精度，再現率ともに良好な値を示したが，語彙的制約については特に再現率が低い．これは，ある単語の語彙的制約となるべき語がコーパス上でその単語と共起しないために，指示的意味として扱われてしまう，もしくは意味差分から削除されてしまうという問題による．今後，この 2 つの意味差分情報の取得については，コーパスの量と分野をさらに増やすこと，語釈文の解析を進めて構文的に指示的意味，語彙的制約，不要語の判定を行うことで改善できる可能性がある．また，本稿では表層的意味を示す語を同一意味分類からのみ抽出したが，こちらも語釈文の解析により他の意味分類の語との関係を示すことも可能であると考えられる．

4 類義語群の順位付け

前章の方法により得られた指示的意味の情報を用いて，入力文中のある表現の類義語群を，文脈との意味的適格性によって順位付けする手法について述べる．

4.1 選択対象

前章の実験でカテゴリ数より表層的意味を示すエントリの数が多かったことからも分かるように，同一の意味分類に属する類義語でも必ずしも同じ表層的意味を持つとは限らない．したがって本稿では，指示的意味を持つ類義語群と言い換えの対象語とは，まず表層的意味が同一であることを前提とする．

4.2 指示的意味と文脈との比較

指示的意味は，エントリに含まれる意味を示す情報である．入力文中の対象語句 w とその類義表現 s_i との比較を行なう際，それぞれが持つ指示的意味の情報によって以下のように判定する．

(1) w と s_i に共通の指示的意味が含まれる
⇒ そのまま言い換え可能

(2) w および s_i が指示的意味を持たない
⇒ そのまま言い換え可能

(3) w に含まれる指示的意味が s_i に含まれない
⇒ その指示的意味を入力文に付加することで言い換え可能

(4) s_i に含まれる指示的意味が w に含まれない
(a) s_i の指示的意味が入力文中にある
⇒ 入力文からその語句を削除することで言い換え可能
(b) s_i の指示的意味が入力文中にない
⇒ 言い換え不可

それぞれの指示的意味は重みを持つため，言い換え可能であればその重みの分だけスコアを増加させ，言い換え不可ならば重みの分だけスコアを減少させる．ただし，(3) の場合は類義表現に指示的意味がないため，スコアの変更を行わない．また，指示的意味を示す語句の入力文への付加や削除についての情報を保持しておく必要がある．

以上から，ある類義表現の指示的意味 d_i がそれぞれ重み e_i を持つとき，その類義表現の指示的意味による意味的適格性のスコア S_d は以下に示す式で表される．

$$S_d = \sum_i p e_i \tag{3}$$

$$\left(p = \begin{cases} 1 & (d_i \text{に関して言い換え可}) \\ -1 & (d_i \text{に関して言い換え不可}) \end{cases} \right)$$

4.3 語彙的制約と文脈との比較

語彙的制約における言い換えの可否は，語彙的制約を示す語句が入力文中に存在するかどうかによって決定される．本稿ではこの点について，入力文中の語句と語彙的制約との類似度をもとに評価する．類似度はシソーラスの類似度計算を元に行うが，ある程度以上離れた意味分類間ではスコアを減少させることを考え，以下の式により類似度を計算する．

$$\log\left(\frac{d_c \times 4}{d_i + d_j}\right) \tag{4}$$

ここで，1つのエントリに複数の語彙的制約が含まれる場合，入力文中のどの語句にどの語彙的制約を割り当てるかを判断する必要がある．この問題に対しては，最も類似度の高い語句の組合せから順に抽出していき，それぞれの語彙的制約の重みと類似度との積から，スコアを決定する．つまり，ある類義表現の語彙的制約 v_i がそれぞれ重み e_i を持ち，入力文中の語句 q_i に割り当てられているとき，その類義表現の語彙的制約による意味的適格性のスコア S_v は以下の式で表される．

$$S_v = \sum_i (e_i \cdot sim(v_i, q_i)) \quad (5)$$

$(sim(v_i, q_i) \cdots v_i$と q_iの類似度)

4.4 類義語群の順位付け

以上により得られた各類義語の2種類のスコアをそれぞれ正規化し合計をとったもの，およびそれぞれについてスコアを元に類義語群の順位付けを行う．また，4.2節で述べた指示的意味と文脈との関係から，入力文へ付加/削除すべき語句も同時に与える．

4.5 実験および考察

コーパスから無作為に選んだ日本語の単文40文に対して言い換えの対象語を適当に選択し，順位付けの実験を行った．順位付けの例を図1に示す．

「乱れた世の中を立て直す.」

1.	実社会	6.	世
2.	世俗	7.	世界
3.	乱世 (削除: 乱れる)	8.	世上
4.	社会	9.	天地
5.	民間		

図 1: 順位付け結果の例

評価は，正解となる言い換え候補語を各文ごとに人手により複数用意し，文書検索の順位付け評価に利用される非補間平均精度を用いて行った．順位付けを行わない場合との比較結果を表3に示す．

表 3: 順位付けの精度

対象語の多義性 (文数)	非補間平均精度 (%)			
	本手法			無作為
	S_d	S_v	$S_d + S_v$	
なし (21)	74.2	63.8	71.2	60.0
あり (19)	48.8	48.3	51.0	42.1
両方 (40)	52.9	48.1	52.4	43.9

多義性のない語に対しては，本稿で提案した手法が特に類義語間の意味的適格性の判定に有効であることが分かる．しかし，多義性を含む語に対しては曖昧性解消が十分に行われておらず，精度の悪化を招いている．

また，指示的意味のみを用いた場合より語彙的制約のみを用いた場合の方が全体に精度が低い．これは，表2で示された語彙的制約の精度の低さの他に，語彙的制約における類似度の決定がシソーラスの意味分類のみを用いていることが原因の一つと考えられ，意味分類とは別に意味的類似度を計算するための何らかの尺度を利用することで改善できる可能性がある．

順位付け結果に含まれる付加/削除の情報に対する定量的評価は行っていないが，削除は低頻度で高精度，付加は高頻度だが低精度であるという傾向にあり，特に付加の情報については提示方法や選別方法を検証する必要がある．

5 おわりに

本稿では，推敲支援を目的とした言い換えにおける類義語群の順位付けについて，その意味的なアプローチに対する語釈文の情報の有効性を示すことができた．今後取り組むべき課題として，語釈文の解析による意味差分抽出精度の向上，類義語群における多義性の解消，語彙的制約と文脈との関係の定式化などが挙げられる．

参考文献

[1] 菅沼明, 牛島和夫: 日本語文章推敲支援ツール『推敲』におけるとりたて詞「は」の抽出法とその評価. 情報処理学会論文誌, 32(11), 1991.

[2] 乾裕子, 岡田直之: 「わかりにくい」表現の検出規則作成—推敲支援システムの実装を目指して—. 言語処理学会第6回年次大会発表論文集, pp. 179–182, 2000.

[3] 藤田篤, 乾健太郎: 語釈分を利用した普通名詞の同概念語への言い換え. 言語処理学会第7回年次大会発表論文集, pp. 331–334, 2001.

[4] 鍛冶伸裕, 黒橋禎夫, 佐藤理史: 国語辞典に基づく平易文へのパラフレーズ. 情報処理学会自然言語処理研究会, NL-144-23, pp. 167–174, 2001.

[5] P. Edmonds: Semantic Representations of Near-Synonyms for Automatic Lexical Choice. *Ph.D. thesis, Department of Computer Science, University of Toronto*, 1999.

[6] 大野晋, 浜西正人: 角川類語新辞典. 角川書店, 1981.

[7] CD-ROM 毎日新聞'92, '95: 日外アソシエーツ株式会社.

[8] 青空文庫, http://www.aozora.gr.jp/main.html