意味情報を考慮した単文言い換えシステム

峯脇さやか† 新見道治† 河口英二†
†九州工業大学工学部

1 はじめに

「言い換え」とは，同じ意味内容を表す複数の表現を結びつける変換であり [1]，人間の言語処理能力の一部である．我々人間は，話し言葉・書き言葉にかかわらず，ある概念を同じ意味（同義）のまま別の表現に変えるだけでなく，抽象化したり，具体化したりすることもできる．

言い換えは，機械翻訳，自動要約，情報検索などその処理の一部として使われてきた [2, 5]．また，テキストベースの情報秘匿技術にも応用できる [3, 4]．このように，言い換えの応用分野は広いので，近年独立したテーマとして研究されるようになってきたが，研究事例はまだあまり多くない．言い換えの研究のほとんどは，言語学的知識に基づく言い換えである [1, 2, 5]．これらは，既存の文法パターンをルールとして蓄え，自然言語文をそのルールに直接当てはめて変換するというものである．多くの場合は実現できているが，意味的に不自然な文章が生成されてしまうことがある．これからは，言語知識だけでなく人間の持つ意味情報に基づいた言い換えが必要である．

本稿では，著者らの研究グループが自然言語の意味表現形式の一つとして提案している SD 式（Semantic-structure Description Form）[6] を利用した言い換え手法を提案する．SD 式は，自然言語における個々の概念，陳述表現，感情表現，あるいはシステムに与える知識データなどを記述するための一種の中間言語である．自然言語概念を SD 式として捉え，その記述データを基にして意味処理を行なおうとするモデルが SD 式意味モデルである．SD 式意味モデルの特徴は，概念間の意味的な近さを定量的に扱うことができることである．

本稿における言い換えとは，意味表現レベルでの言い換えのことであり，必ずしも自然言語文を文法的に変換するものではない．すなわち，ある SD 式を別の SD 式に変換することである．その前提としては，“自然言語文と SD 式の相互変換”が機械的に可能でなければならない．このことに関しては，文献 [7, 8] で議論している．

1 つの SD 式から別の SD 式への言い換えでは，システムに実装された概念体系を利用する．SD 式意味モデルにおける概念体系とは，SD 式で記述した概念を節，詳述量を枝とする意味ネットワークである．ここで詳述量とは，2 つの概念の詳述の程度を表したものである．概念体系において，ある 1 つの枝で結ばれた 2 つの節には，詳述関係による上位－下位関係があり，これを言い換え対とみなすことができる．つまり，ある概念の言い換えをその概念の上位概念または下位概念とする．

言い換えた概念が言い換え対象概念の上位概念ならば，言い換え対象を抽象化したことになり，下位概念ならば具体化したことになる．言い換えを利用するアプリケーションが言い換えシステムに抽象化した言い換えを出力するよう要求したとき，言い換えシステムが抽象化した言い換え結果のみを出力できれば，言い換えを利用するアプリケーションにとって有効である．そこで，アプリケーションの要求と一致する言い換え結果を出力できるように，抽象化／同義／具体化の分類を行う．この分類には，SD 式意味モデルで定義された枠組を用いた定量的な判定を利用する．この処理を，「詳述レベルのカテゴリ化」と呼ぶ．

表 1: SD 式の記述例

SD 式	自然言語文
$[s$(自分)$, v$(テニス/時/毎日)$]$	私は毎日テニスをする．
$[s$(自分)$, v$(質問)$, o$(相手)$,$ $c([s$(当該)$, v$(である)$, o$(何)$])]$	それは何ですか．
$[a$(ジョン)$]$	ジョン．(呼びかけ)

表 2: 知識データの記述例

SD 式	自然言語文
(九州)$incl$(沖縄)	九州は沖縄を含む．
(りんご)$ptof$(果物)	りんごは果物の一種である．

以下，**2** では，SD 式意味モデルの概要について述べる．**3** では，SD 式意味モデル利用した単文の言い換え手法ついて述べる．**4** で実験例を示し，本手法について検討する．最後に**5** でまとめる．

2 SD 式意味モデルの概要 [6]

SD 式意味モデル（Semantic-structure Description Form Semantics Model）は，自然言語の意味を定量的に分析するための枠組である．このモデルに従った意味記述を SD 式と呼ぶ．SD 式では，自然言語における個々の概念，陳述表現，感情表現，システムに与える知識データなどを記述することができる．SD 式意味モデルの特徴は，与えられた 2 つの概念の意味的な差異を定量的に扱えることである．

2.1 SD 式の記述例

SD 式の構文は，SDG（SD-form Grammar）と呼ばれる曖昧さのない文脈自由文法で規定されている．SD 式は，概念ラベル，修飾子，規定子，結合子，機能項目記号，区切り記号の 6 種類の「SD 式記号」から構成される記号列である．概念ラベルとしては，既成の単語（日本語や英語）を借用しており，変数ラベル，単純ラベルなどに分類している．

陳述文，会話文の自然言語概念とそれに対応する SD 式の例を表 1 示す．表 1 において，s(D)，v(D)，$\cdots$ の形式におけるD を機能項目の内容と呼ぶ．システムに与える知識の例を表 2 に示す．表 2 における「$incl$」，「$ptof$」は結合子記号である．表 3 にいくつかの結合子とそれぞれの用法を示す．

2.2 SD 式の意味的情報量

意味を定量的に処理するために，各 SD 式記号に意味素量と呼ばれる値が与えられている．ある SD 式に対して，その SD 式を構成する SD 式記号の意味素量を全て足し合わせた値（意味素量の総和）をその SD 式の「意味量」として定義している．任意の SD 式をD とするとき，その意味量を次のように表し，単位を $semit$ としている．

$$si(\mathrm{D}) = n$$

表 3: 一部の結合子記号の用法

結合子記号	用法	結合子記号	用法
equa	等価	*incl*	包含
para	並列関係	*kdof*	種類
plus	結合	*ptof*	部分

以下に意味素量の例を示す.

- 変数ラベル　“X, Y, ···”　: 1 [*semit*]
- 単純ラベル　“車, 買う, ···”　: 10 [*semit*]
- 修飾子　“/”　: 1 [*semit*]
- 規定子　“*nega, only,* ···”　: 2 [*semit*]
- 結合子　“*equa, para,* ···”　: 1 [*semit*]
- 機能項目記号　“*s, v, c,* ···”　: 1 [*semit*]
- 区切り記号　“[ ]”　: 1 [*semit*]
- 区切り記号　“( )”, “,”　: 0 [*semit*]

意味量の例を以下に示す.

例 1 $si([s(\text{自分}), v(\text{テニス/時/毎日})]) = 45$

SD 式意味モデルでは，このような“絶対的な”意味素量そのものを定めているのではなく，“相対的な”意味素量の設定例を示している.

2.3 2 概念間の詳述関係

2 つの概念 D_1, D_2 に関して，D_1 の意味をより具体化したものが D_2 であり，かつ，D_2 の意味をより抽象化したものが D_1 であるとき,「D_1 と D_2 には詳述関係がある」という．このとき，D_1 を D_2 の先祖，D_2 を D_1 の子孫と呼ぶ．D_1 から D_2 への詳述関係を次のように表す.

$$elab(D_1, D_2) = n$$

ここで，n $(0 \leq n < \infty)$ は詳述量といい，詳述の程度を表すものである.

2.4 意味差の尺度

SD 式意味モデルでは，概念間の「意味差の尺度」を次のように定義している．2 つの概念 D_1, D_2 に共通する全ての先祖 D_{01}, D_{02}, ··· の中で D_1, D_2 に最も近い先祖を「D_1, D_2 の最近共通先祖」と呼び，D_0 で表す.「D_1 と D_2 の意味差 n_0」を次のように表す.

$$diff(D_1, D_2) = n_0$$

意味差は D_0 から D_1 の詳述量と D_0 から D_2 の詳述量との和である.

2.5 概念体系

概念体系は，概念集合と詳述量の集合からなる意味ネットワークである．概念集合のそれぞれの概念を節，詳述量を枝としている．概念体系において，最上位概念を変数ラベル X とする．また，全ての念の集合を Ω で表す．与えられた概念の集合を体系化する操作（概念の体系化）は，一定のアルゴリズムに従ってシステム内で実行可能である [9]．概念体系化の過程において，新たな概念が導出される．この概念を導出概念と呼ぶ.

概念体系の例を以下に示す.

例 2 表 4 に示す知識データがシステムに与えられていたとする．このときの概念体系図を図 1 に示す．また，導出概念 IC_1 と概念集合 Ω を次に示す.

$$IC_1 : (\mathsf{Y})kdof(\text{和菓子})$$
$$\Omega = \{\mathsf{X}, F_1, F_2, \cdots, F_7, IC_1\}$$

表 4: 知識データ

	知識データ	自然言語文
F_1	和菓子	和菓子
F_2	ういろう	ういろう
F_3	ようかん	ようかん
F_4	芋ようかん	芋ようかん
F_5	(ういろう) *kdof*(和菓子)	ういろうは和菓子の一種
F_6	(ようかん) *kdof*(和菓子)	ようかんは和菓子の一種
F_7	(芋ようかん) *kdof*(ようかん)	芋ようかんはようかんの一種

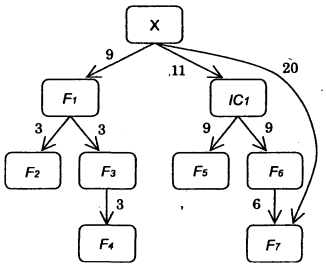

図 1: 概念体系図

2.6 認識

SD 式意味モデルでは，2.5 で述べた概念体系利用することにより，認識動作を形式化している.

外部入力として，ある概念 D_{in} が与えられたとする.「D_{in} を認識する」とは，システム内の概念体系において，D_{in} に最も近い先祖 D_{rec} を見出すことである．つまり，

$$D_{rec} = \arg\min_{D}\{elab(D, D_{in}) \mid D \in \Omega\}$$

である．ただし，Ω は概念体系における概念集合である.

3 単文の言い換え手法

3.1 言い換えシステム

試作した言い換えシステム（以下，本システム）の構成図を図 2 に示す．本システムは,「言い換え生成部（Paraphrasing engine）」,「意味フィルタ（Semantic filter）」,「SD 式意味モデル実験システム ver.4（SDENV-4）[10]」,「知識データ（Knowledge data）」,「概念体系データ（Concept hierarchy）」で構成している．言い換え生成部では，言い換え対象の SD 式（SD-Form）を入力とし，適宜 SDENV-4 や概念体系データにアクセスしながら言い換え生成処理を行なう．意味フィルタでは，得られた全ての言い換え結果について，詳述レベルのカテゴリ化を行なう．SDENV-4 とは，2 で示した枠組を計算機上に実装したものであり，意味差の計算や概念の体系化を行う Perl プログラムである．本システムで用いる知識データは，人手で作成し，あらかじめ登録しているものである．概念体系データは，与えられた知識データを SDENV-4 が体系化したものであり，あらかじめシステムに登録しておいたものである.

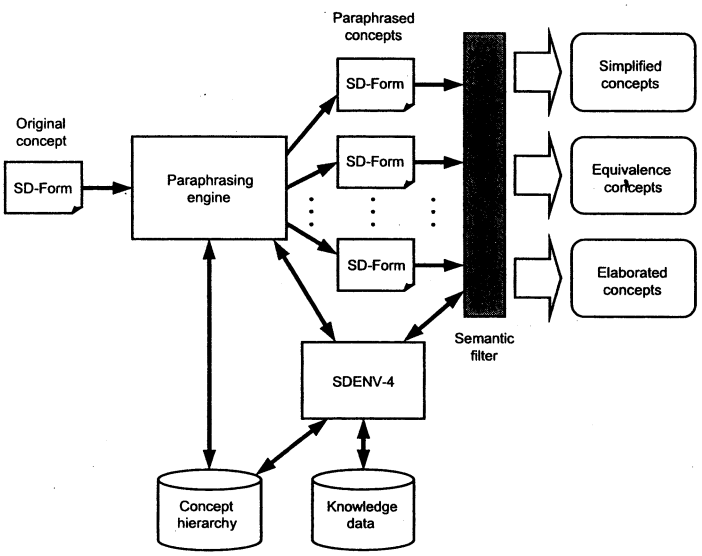

図 2: 言い換えシステムの構成

3.2 言い換え手法

上述のように，本研究における言い換えとは，概念体系を用いてある SD 式を別の SD 式に変換することである．まず，SD 式で記述した言い換え対象を認識する．次に，認識された SD 式について，変数ラベルを含んでいるならば，その変数ラベルを具体的な SD 式に置き換える．そして，認識された SD 式および，その先祖や子孫を基本概念として，言い換え候補を生成する．

以上の処理手順を以下に示す．

Step 1 言い換え対象の SD 式を D_{in} とする．D_{in} を認識し，認識結果を D_{rec} とする．ただし，$\mathsf{D}_{rec} = \mathsf{X}$ の場合，処理を中止する．

Step 2 D_{rec} に変数ラベルが含まれている場合に，その具現化の処理を行なう．変数ラベルの具現化とは，D_{rec} に含まれている変数ラベルと詳述関係をもつ SD 式を D_{in} 中から見出すことである．見出した SD 式のことを，具現 SD 式と呼ぶ．

Step 3 $elab(\mathsf{HC}_s, \mathsf{D}_{rec}) \geq 0$，$elab(\mathsf{D}_{rec}, \mathsf{LC}_t) > 0$ を満たす HC_s，$\mathsf{LC}_t(s, t = 1, 2, \cdots)$ を求める．ただし，$\mathsf{HC}_s \neq \mathsf{X}$ とする．

Step 4 $\mathsf{D}_{rec}, \mathsf{HC}_1, \mathsf{HC}_2, \cdots, \mathsf{LC}_1, \mathsf{LC}_2, \cdots$ を一般に $\mathsf{Q}^m(m = 1, 2, \cdots)$ と表す．全ての Q^m について **Step 5**〜**9** の処理を行う．

Step 5 Q^m の機能項目の内容を一般に $\mathsf{D}_i^m(i = 1, 2, \cdots)$ と表す．

Step 6 全ての $\mathsf{D}_i^m(i = 1, 2, \cdots)$ について，その先祖や子孫を $\mathsf{D}_{ij}^m(j = 1, 2, \cdots)$ とする．D_i^m が変数ラベルである場合，その具現 SD 式と概念体系中にある具現 SD 式の先祖や子孫を D_{ij}^m とする．ただし，$\mathsf{D}_{ij}^m \neq \mathsf{X}$ とする．

Step 7 D_i^m を D_{ij}^m に置換し，置換後の SD 式を Q_j^m とする．

Step 8 $\mathsf{Q}_0^m = \mathsf{Q}^m$ とする．

Step 9 $\mathsf{Q}_j^m(j = 0, 1, \cdots)$ が D_{in} と同一の SD 式でなく，変数ラベルが含まれていないならば，Q_j^m を言い換え候補 $\mathsf{P}_k(k = 1, 2, \cdots)$ とする．

3.3 詳述レベルのカテゴリ化

詳述レベルのカテゴリ化とは，言い換え生成部で得られた言い換え候補 P_k を抽象化／同義／具体化の 3 つに分類することである．それぞれの言い換え結果が，抽象化／同義／具体化のいずれに属するかの判定は，意味差，最近共通先祖との詳述量，意味量を順次用いる．

詳述レベルのカテゴリ化の手順を以下に示す．

Step 1 意味差

(a) $diff(\mathsf{P}_k, \mathsf{D}_{in}) = 0$ → 「同義」.
(b) $diff(\mathsf{P}_k, \mathsf{D}_{in}) \neq 0$ → **Step 2** へ.

Step 2 最近共通先祖からの詳述量の大小

(a) $elab(\mathsf{D}_0, \mathsf{P}_k) < elab(\mathsf{D}_0, \mathsf{D}_{in})$ → 「抽象化」.
(b) $elab(\mathsf{D}_0, \mathsf{P}_k) > elab(\mathsf{D}_0, \mathsf{D}_{in})$ → 「具体化」.
(c) $elab(\mathsf{D}_0, \mathsf{P}_k) = elab(\mathsf{D}_0, \mathsf{D}_{in})$ → **Step 3** へ.

Step 3 意味量の大小

(a) $si(\mathsf{P}_k) < si(\mathsf{D}_{in})$ → 「抽象化」.
(b) $si(\mathsf{P}_k) > si(\mathsf{D}_{in})$ → 「具体化」.
(c) $si(\mathsf{P}_k) = si(\mathsf{D}_{in})$ → 「カテゴリ外」.

ここで，**Step 3(c)** は，意味差が 0 でないので「同義」ではなく，最近共通先祖からの詳述量と意味量がそれぞれ等しいので，「抽象化」または「具体化」のどちらにも属さない場合である．このような場合は，抽象化／同義／具体化のいずれにも属さないという意味で「カテゴリ外」と表す．

4 実験と考察

4.1 実験

本システムを試作し，新聞記事 [11, 12, 13] の背景にある知識データを人手で作成したものを用いて動作実験を行なっ

表 5: 実験結果の一部

SD 式	自然言語文	カテゴリ
$[s(\text{日本相撲協会}), v(\text{承認/過去}), o(\text{開催/韓国公演})]$	日本相撲協会は，韓国公演の開催を承認した．	抽象化
$([s(\text{日本相撲協会 (\$)}), v(\text{開く/過去}), o(\text{臨時理事会})])$ $pseq([s(\$), v(\text{承認/過去}),$ $o(\text{事柄}/[s(*\text{省略}), v(\text{行う/場所/ソウル}), o(\text{公演})])])$	日本相撲協会は，臨時理事会を開き，ソウルで公演を行うことを承認した．	同義
$([s(\text{日本相撲協会 (\$)}), v(\text{開く/過去}), o(\text{臨時理事会})])$ $pseq([s(\$), v(\text{合意/過去}),$ $o(\text{事柄}/[s(*\text{省略}), v(\text{行う/場所/ソウル}), o(\text{公演})])])$	日本相撲協会は，臨時理事会を開き，ソウルで公演を行うことを合意した．	同義
$([s(\text{日本相撲協会 (\$)}), v(\text{開く/過去}), o(\text{臨時理事会})])$ $pseq([s(\$), v(\text{合意/過去}),$ $o(\text{事柄}/[s(*\text{省略}), v(\text{行う/場所/ソウル/韓国}), o(\text{公演/大相撲})])])$	日本相撲協会は，臨時理事会を開き，韓国のソウルで大相撲の公演を行うことを合意した．	具体化

表 6: 知識の量と言い換え結果の個数

知識の量	言い換え結果の個数
12	8
19	15
25	63

た．知識データとしての SD 式の個数（知識の量）は 25 個である．この実験での言い換え対象（SD 式）と，それに対応する自然言語文を以下に示す．

言い換え対象（SD 式）
$[s(\text{日本相撲協会}), v(\text{承認/過去}),$
$o(\text{事柄}/[s(*\text{省略}), v(\text{行う/場所/ソウル}), o(\text{公演})])]$

言い換え対象（自然言語文）
“日本相撲協会は，ソウルで公演を行うことを臨時理事会で承認した．”

言い換えられた SD 式とその自然言語文[1]，および，意味フィルタにおけるカテゴリを表 5 に示す．また，システムに与える知識の量を変化させて実験を行なった．使用した知識の量と生成された言い換え結果の個数を表 6 に示す．

4.2 考察

表 6 より，知識の量が増せば，より多くの言い換え結果が得られることが予想される．本システムでは，人手で作成した知識データを与えている．知識データが多ければ，より多くの概念体系データが得られ，概念体系データが多ければ，多くの言い換えが生成できる．よって本システムに，より多くの知識データを与えれば，より多くの言い換え結果が得られることが，本システムの特徴といえる．

意味フィルタにおける詳述レベルのカテゴリ化は，いずれの場合も適切でると考えられる．しかし，一般の人間にはどれも「同義の言い換え」のように見える．本研究において，「同義」と判定するのは意味差が 0 のときだけである．意味差が 0 でない場合，つまり「抽象化」あるいは「具体化」と判定したものについて「同義」と感じてしまうのは，人間が「同義」と感じる範囲が機械よりも広いからではないかと考えられる．

5 おわりに

SD 式意味モデルを利用した単文の言い換え手法を提案した．また，本手法を計算機上に実装し，動作実験を行なった．知識データが少ない場合，得られた言い換え数も少なく，知識データが多い場合，多くの言い換えが得られた．本手法の特徴は，与える知識データの量が多ければ多いほど，言い換え結果も多く得られることである．意味フィルタにおける詳述レベルのカテゴリ化について，どの場合も適切な判定だった．

今後の課題は，文脈を考慮して複数の文章を言い換える手法へと発展させることである．

[1] SD 式から自然言語文への変換は，文献 [8] における生成アルゴリズムに従って人手で行なった．

参考文献

[1] 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).

[2] 近藤恵子，佐藤理史，奥村学: 格変換による単文の言い換え，情報処理学会論文誌，Vol. 42, No. 3, pp. 465–477 (2001).

[3] 中川裕志，三瓶光司，松本勉，柏木健志，川口修司，牧野京子，村瀬一郎: 意味保存型の情報ハイディング—日本語文書への適用，情報処理学会論文誌，Vol. 42, No. 9, pp. 2339–2350 (2001).

[4] 峯脇さやか，伊藤友和，新見道治，野田秀樹，河口英二: SD 式を利用した言語ステガノグラフィ，情報処理学会研究報告，Vol. 2002, No. 68, 2002–CSEC–18, pp. 137–144 (2002).

[5] 乾健太郎: 言語表現を言い換える技術，言語処理学会第 8 回年次大会チュートリアル (2002).

[6] Masahiro Wakiyama, Hideki Noda, Koichi Nozaki, Eiji Kawaguchi: Computation Algorithm of Semantic Difference Measure in the SD-Form Semantics Model, 情報処理学会論文誌，Vol. 40, No. 3, pp. 1065–1079 (1999).

[7] 榊原正典: 自然言語文からの SD 式生成，修士論文，九州工業大学大学院工学研究科 (2001).

[8] 林勝仁: SD 式からの自然言語文生成システムに関する研究，修士論文，九州工業大学大学院工学研究科 (2002).

[9] Guifeng Shao and Eiji Kawaguchi: A Self-Organization Process of Concept Hierarchy in the SD-Form Semanticx Model, *International Symposium on Artificial Intelligence*, pp. 393–400 (1992).

[10] 吉原将太，峯脇さやか，脇山正博，河口英二: CGI を用いた SD 式意味処理実験システムの試作，電子情報通信学会技術研究報告，Vol. 101, No. 484, TL2001–22, pp. 29–36 (2001).

[11] 朝日新聞社: 朝日新聞 アサヒ・コム (2002). (`http://www.asahi.com/`).

[12] 毎日新聞社: 毎日新聞 Mainichi INTERACTIVE (2002). (`http://www12.mainichi.co.jp/`).

[13] 読売新聞社: 読売新聞 YOMIURI ON-LINE (2002). (`http://www.yomiuri.co.jp/`).