言い換えのための機能表現シソーラスの作成

桝田 達也† 佐藤 理史† 藤田 篤†
†名古屋大学大学院工学研究科
masuda@sslab.nuee.nagoya-u.ac.jp, {ssato, fujita}@nuee.nagoya-u.ac.jp

1 はじめに

言い換えとは、ある表現をその意味内容を変えずに別の表現に置き換えることを言う。近年、電子文書が爆発的に増加しており、利用者や利用形態に適した形式にテキストを自動的に言い換えることの必要性が高まってきている。これが可能となれば、子供や外国人など、利用者の言語能力にあわせて文章を言い換えたり、機械処理が容易な表現に文章を言い換えたりすることが可能となる。

ここでは、文節を言い換えることを考える。文節を構成する要素は、「犬」、「動く」、「楽しい」のような内容語と、「が」、「に対して」、「なければならない」のような機能表現に分けられる。これらを言い換える場合、同義関係を定義したシソーラス (同義語辞書) を利用する方法が一般的である。

内容語のシソーラスは、大規模で実用的なものが多数存在する。これに対して、機能表現のシソーラスで電子的に利用できるものは、我々が知る限り、機能表現辞書『つつじ』[1] しかない。『つつじ』には、16,801 の機能表現 (表層形) が収録されており、これらの全てに対して接続情報、文体、難易度などが付与されている。さらに、『つつじ』では、意味的に等価な (言い換え可能な) 機能表現のグループに対して、意味コードと呼ばれる ID が付与されている。この意味コードは、199 種類存在する。

一般に、文節は複数の機能表現を含むことができる。例えば、文節「動か/ ない/わけがない」は、「ない」と「わけがない」という 2 つの機能表現を含む。この 2 つの機能表現が接続した「ない/わけがない」という機能表現列は、1 つの機能表現「に違いない」に言い換えることができる (「動く/に違いない」)。文節の言い換えを実現するためには、このような長さが異なる機能表現列間の言い換えが必要となる場合がある。

機能表現辞書『つつじ』の意味コードによって定義される同義関係は、機能表現間 (1 対 1) の同義関係に限定されており、そのままでは、上記のような長さの異なる機能表現列間の言い換えを実現することはできない。そこで、我々は、『つつじ』で定義されている意味コードを利用し、このようなの同義関係を提供するシソーラスの作成に取り組んだ。本稿では、作成したシソーラスの概要と、長さが異なる機能表現列間の言い換えを実現するための規則の作成方法、および、作成した規則の評価について述べる。

2 機能表現シソーラスの概要

今回作成したシソーラスは、長さ 3 以下の機能表現列 X に対し、X と言い換え可能な機能表現列 Y を列挙する機能を持つ。本シソーラスは、(1) シソーラス本体と、(2) 意味ラベル書き換え規則集合、の 2 つの要素から構成されている。

2.1 シソーラス本体

シソーラス本体は、機能表現列エントリーの集合である。それぞれのエントリーは、見出し語 (機能表現列)、意味ラベル、接続情報の 3 つの要素から構成される。エントリーの例を表 1 に示す。

エントリーの見出し語は、長さ 3 以下の機能表現列である。機能表現の区切りは “/” で表す。見出し語の長さに基づき、全エントリーは、1-gram エントリー、2-gram エントリー、3-gram エントリーの 3 種類に分けられる。

意味ラベルは、見出し語 (機能表現列) を構成する機能表現の『つつじ』意味コードの列である。意味コードの区切りは、見出し語と同様に “/” で表す。同義関係にあるエントリーは、同一の意味ラベルを持つ。

接続情報は、左接続情報と右接続情報の 2 つの情報から構成される。左接続情報は、どのような形態素の直後にそのエントリー (機能表現列) が接続するかということを表す情報である。例えば、「からして」の左接続情報は “名詞” である。これは、名詞の直後に「からして」が接続することを示している。

右接続情報は、そのエントリー (機能表現列) の直後にどのような形態素が接続しうるかを間接的に表す情報で、特定の品詞もしくは活用形が記述される。例えば、「に決まっている」の右接続情報は、“動詞基本形” と記述されている。これは、動詞の基本形の直後に接続できる形態素が、「に決まっている」の直後に接続できるということを表している。

表 1: エントリーの例

	見出し語	意味ラベル	左接続情報	右接続情報
1-gram	からして	a31	名詞	接続助詞 “て”
	に決まっている	I21	名詞, 用言基本形, 助動詞 “た, だ”	動詞基本形
2-gram	ない/わけがない	y41/y11	用言未然形	形容詞基本形
	しかない/から	D21/s22	名詞, 動詞基本形	接続助詞 “から”
3-gram	ことはない/という/が	M11/i11/t25	用言基本形	接続助詞 “が”
	まで/だ/と	f11/D41/r21	名詞, 動詞基本形	接続助詞 “と”

2.2 シソーラス本体の作成

シソーラス本体であるエントリー集合は、機能表現辞書『つつじ』[1] を用いて作成した。

『つつじ』は、形態素解析にも利用できるように設計されているので、機能表現の異形 (表層形) のほとんどを収録している。その中には、あまり使用されない表層形も数多く含まれる。これらの表層形を、毎日新聞 1991-2005 年版 (計 15 年分) のコーパスを利用して除外した。具体的には、このコーパスに 1 回以上出現するものを抽出し、これらを 1-gram エントリーの見出し語として定義した。エントリーの意味ラベル・接続情報には、『つつじ』の意味コード・接続情報をそのまま使用した。なお、『つつじ』には、同一の表層形をとる複数のエントリー (異なる意味コードを持つ) が定義されている。このような表層形がコーパスに出現した場合は、異なる意味ラベルを持つ複数のエントリーをシソーラスに作成した。最終的に、4,623 エントリーを作成した。

次に、こうして作成した 1-gram エントリーを用いて、2-gram エントリーを作成した。具体的には、1-gram エントリーのあらゆる並びに対し、それが次の 2 つの条件を満たすかどうか調べ、条件を満たすもののみを 2-gram エントリーとして採用した。

1. 左側のエントリーの右接続情報と、右側のエントリーの左接続情報から、これらが接続できると判定できる。
2. 2 つのエントリーの見出し語を接続した機能表現列が、上記の毎日新聞コーパスに 1 回以上出現する。

2-gram エントリーの意味ラベルは、2 つの 1-gram エントリーの意味ラベルを接続して作成する。左接続情報は左側の 1-gram エントリーの左接続情報を、右接続情報は右側の 1-gram エントリーの右接続情報をコピーして作成する。

3-gram エントリーも、2-gram エントリーと同様の方法で作成した。

作成した機能表現シソーラスのエントリー数と意味ラベルの異なり数を表 2 に示す。この表に示すように、

表 2: エントリー数と意味ラベルの異なり数

	1-gram	2-gram	3-gram	合計
エントリー数	4,623	76,734	299,254	308,611
意味ラベル数	199	8,164	62,245	70,608

シソーラス本体のエントリー数は 308,611 である。

2.3 意味ラベル書き換え規則

このように作成したシソーラスにおいて、機能表現 (1-gram エントリー) の同義関係の合成によって定義される 2-gram エントリー間の同義関係、および、3-gram エントリー間の同義関係は、意味ラベルによってすでに定義されている。例えば、2-gram のエントリー「てもよい/ので」と「てもかまわない/から」は、どちらも同じ “F11/s22” という意味ラベルを持ち、これらのエントリー間に同義関係が定義されていることになる。

残された問題は、長さが異なる機能表現列エントリー間の同義関係をどのように定義するかという点である。これを実現するために、我々は、**意味ラベル書き換え規則**を定義する。

意味ラベル書き換え規則とは、以下のような規則である。

y41/y11 → I21

この規則は、“y41/y11” という意味ラベルが、“I21” という意味ラベルに書き換え可能であるということを示す。この書き換え規則により、例えば、「ない/わけがない (y41/y11)」と「に違いない (I21)」の間に、同義の関係が設定されることになる。

我々は、エントリーの意味ラベルの一致によって定義される同義関係と意味ラベル書き換え規則によって結びつけられる同義関係を区別する。このような区別を導入するのは、少数ながら、方向性がある意味ラベル書き換え規則が存在するからである。

2.4 言い換え生成機能

本シソーラスでは、長さ 3 以下の機能表現列 X に対し、次の方法で、X と言い換え可能な機能表現列 Y を生成する。

1. X のエントリーから、意味ラベル s_X を取得する。

2. 意味ラベル s_X を持つエントリーをすべて列挙する (出力 1)。
3. 意味ラベル s_X に適用できる意味ラベル書き換え規則を 1 回だけ適用し、書き換えられた意味ラベル s_Z を得る。この後、意味ラベル s_Z を持つエントリーをすべて列挙する。これを適用できるすべての規則に対して実行する (出力 2)。

3 意味ラベル書き換え規則の作成

本研究では、以下の 2 つのデータを参照し、意味ラベル書き換え規則を作成した。

1. 人間が自由に機能表現を言い換えたデータ
2. 非命題的意味構造間の類似性規則

3.1 言い換えデータを利用した規則の作成

松吉らは、機能表現の言い換えを定式化するための予備調査として、人間が機能表現をどのように言い換えるかを調査した [2]。この調査データには、機能表現列間の言い換えが含まれている。例を以下に示す。

例)「説明/によれば」→「説明/で/は」
「会社員/にすぎない」→「会社員/で/しか/ない」

我々はまず、このデータから、意味ラベル書き換え規則の作成に利用できると考えられる 94 個の言い換え例を抽出した。次に、抽出した例に含まれるすべての機能表現を、適切な『つつじ』の意味コードに置き換えた。

例)「説明/a31」→「説明/O21/k13」
「会社員/D43」→「会社員/O21/m31/y41」

最後に、内容語を削除し、以下の 2 点を考慮して、意味ラベル書き換え規則を作成した。

1. 書き換え規則は十分な一般性を有しているか。(一般性を有していないと判断した場合は、規則を作成しない。)
2. 方向性はないか。(方向性がある場合は、片方向の規則のみを作成する。方向性がない場合は、両方向の規則を作成する。)

以上の方法で、総計 110 個の意味ラベル書き換え規則を作成した。

3.2 類似性規則を利用した規則の作成

首藤らは、次のような非命題的意味構造間の類似性規則を定義している [3, 4]。

(1) (否定 $_1$(否定 $_1$(S))) = (S)
(2) (必要性 (S)) = (否定 $_{12}$(可能 (否定 $_{12}$(S))))

ここで、S は骨格文を表す。(1) は、骨格文に否定 $_1$ の意味を付加し、さらに否定 $_1$ の意味を付加すると、骨格文そのものと等価な意味になるということを示している。

我々はまず、このような類似性規則のうち、意味ラベル書き換え規則の作成に利用できそうな 18 規則を抽出した。次に、これらの類似性規則を意味ラベルを用いて再現した。

(1) y41/y41(否定/否定) = (機能表現なし)
(2) D11(当為) = y41/E11/y41(否定/可能/否定)

最後に、先の手法と同様の 2 点を考慮し、45 個の意味ラベル書き換え規則を作成した。

2 つの方法で作成した意味ラベル書き換え規則の概要を表 3 に示す。ここで、「2 → 2」は同一長の機能表現列エントリー間を結びつけるものである。これは、次の例に示すような、1-gram エントリー間の同義関係からは結びつけられない同義のエントリー対のために作成した意味ラベル書き換え規則である。

例)「ない/かもしれない (y41/I11)」
→「に決まっている/わけではない (I21/y41)」

表 3: 意味ラベル書き換え規則の概要

M → N	手法 1	手法 2	合計
1 → 2	40	13	53
1 → 3	10	5	15
1 → 4	2	0	2
2 → 0	0	1	1
2 → 1	40	13	53
2 → 2	4	8	12
2 → 3	1	0	1
3 → 1	10	5	15
3 → 2	1	0	1
4 → 1	2	0	2
合計	110	45	155

4 意味ラベル書き換え規則の評価

作成した意味ラベル書き換え規則の適切さを、以下のような方法で評価した。まず、『日本語能力試験出題基準 [改訂版]』[5] の例文から『つつじ』に収録されている機能表現を含む文節、200 文節を抽出した。次に、作成したシソーラスを用いて、それらの文節の機能語部に対する言い換えを生成した。ただし、生成する言い換えは、意味ラベル書き換え規則を使用して生成されるものに限定した。その後、生成された言い換えを毎日新聞コーパスでの出現頻度で順位付けし、最後に、文節の内容語部と接続して、順位付けられた形で、言い換え文節を出力した。この結果、200 文節中の 78 文節に対して、言い換え文節が出力された。文節「動か/ない/わけがない」に対する出力を表 4 に示す。

生成された言い換えのうち、上位 5 位までに出力された 360 件をの適否を判定した。360 件の言い換えを生成するために使用された意味ラベル書き換え規則の異なり数は、計 44 であった。

表 4:「動か/ない/わけがない」を入力したときの出力例

順位	出現頻度	代替表現
1 位	7,288	動く/に違いない
2 位	1,608	動く/に決まって
3 位	1,209	動く/に決まってい
4 位	1,162	動く/にちがいない
5 位	850	動く/に決まっている
6 位	476	動く/には違いない
7 位	175	動く/に相違ない
:	:	:

表 5: n 位で生成された言い換えの分類結果の内訳

順位	○	×	合計	○の割合
1 位	62	16	78	79%
2 位	58	16	74	78%
3 位	51	19	70	73%
4 位	60	10	70	86%
5 位	50	18	68	74%
合計	281	79	360	78%

生成された代替表現を、1 人の作業者が主観に基づき以下の 2 つに分類した。

- 言い換えとして適切である (○)
- 言い換えとして適切ではない (×)

分類結果を表 5 に示す。評価対象の 360 件の出力のうち、281 個 (78%) が適切な言い換えであった。このことから、今回の言い換え生成で用いられた 44 件の意味ラベル書き換え規則には、大きな問題はないと考える。しかし、今回の実験では使用されなかった 111 個の意味ラベル書き換え規則の適切さは不明である。これらの規則を評価するためには、より大規模なテストデータに対する評価が必要である。

次に、生成された不適切な言い換え 79 個を分析した。その結果、不適切な言い換えは次の 3 種類に大別できた (表 6)。

第 1 のグループは、次の例に示すような「た (完了)」の言い換えに起因する誤りである。

「鳴っ/たかと思うと (o21)」
→ ○「鳴っ/た/と同時に (B21/o11)」
→ ×「鳴っ/ちゃう/や (B21/o11)」

意味ラベル書き換え規則 (o21 → B21/o11) により、「たかと思うと」に対して言い換え表現「た/と同時に」が生成される。これは適切な言い換えである。しかし、意味ラベル “B21” をとるエントリーには「てしまう」があり、この異形 (口語体) として「ちゃう」が定義されている。また、意味ラベル “o11” をとるエントリーには、「と同時に」の他に「や (部屋に入ってくる や 窓を開けた」がある。「ちゃう」と「や」は接続条件を満たし、かつ、「ちゃうや」という文字列が毎日新聞コーパスに存在するため、2-gram エントリーとして存在し、

表 6: 生成された不適切な言い換えの分析結果

グループ 1	「た (完了)」の言い換え	36 個
グループ 2	意味ラベル書き換え規則の無条件適用	14 個
グループ 3	入力文節の機能表現の意味の不特定	29 個

言い換えとして出力される。『つつじ』では、「ちゃう」は口語体、「や」は堅い文体と定義されているので、これらの文体情報を用いれば、このような不適切なエントリーを排除できる可能性がある。

第 2 のグループは、意味ラベル書き換え規則を無条件に適用していることに起因する誤りである。

(1)「覚え/た/にすぎない」→ ○「覚え/た/だけ/だ」
(2)「会社員/にすぎない」→ ×「会社員/だけ/だ」

「にすぎない」から「だけ/だ」への言い換えは、(1) のように「動詞＋た」の直後に接続する場合は適切であるが、(2) のように「名詞」の直後に接続する場合は不適切である。この問題を解決するためには、意味ラベル書き換え規則の適用にある種の条件を導入する必要がある。

第 3 のグループは、入力文節中の機能表現の意味を特定せずに言い換えを生成することに起因する誤りである。以下に例を示す。

(1)「走っ/てならない (不許可)」
→ ○「走っ/てよく/ない」
(2)「寂しく/てならない (自然発生)」
→ ×「走っ/てよく/ない」

機能表現「てならない」には、不許可と自然発生の 2 つの意味があり、不許可の場合 (1) は「てよく/ない」と言い換えられるが、自然発生の場合 (2) は言い換えることはできない。これを避けるためには、入力文節中の機能表現の意味的曖昧性の解消が必要である。本シソーラスで定義されている意味ラベル書き換え規則は、C11(不許可) → F11/y41 であり、“不許可” であることが特定できれば、このような不適切な言い換えは生成しない。

謝辞 本研究では、毎日新聞 1991-2005 年版を使用した。

参 考 文 献

[1] 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂. 自然言語処理, Vol.14, No.5, pp. 123–146, 2007.

[2] 松吉俊, 佐藤理史. 文体と難易度を制御可能な日本語機能表現の言い換え. 自然言語処理, Vol.15, No.2, pp. 75–99, 2008.

[3] 本田聖晃, 田辺利文, 高橋雅仁, 吉村賢治, 首藤公昭. 日本語文末表現の言い換え. 福岡大学工学集報, 79, 2007.

[4] K. Shudo, T. Tanabe, M. Takahashi, and K. Yoshimura. MWEs as Non-propositional Content Indicators. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing (MWE-2004)*, pp. 32–39, 2004.

[5] 国際交流基金, 日本国際教育支援協会. 日本語能力試験 出題基準 –改訂版–. 凡人社, 1994.