直喩と比較による間接表現を利用した用語説明の自動生成

伊藤 玄暉† 　　椛沢 直樹‡ 　　藤井 敦‡
† 東京工業大学工学部情報工学科
‡ 東京工業大学大学院情報理工学研究科計算工学専攻

1 はじめに

近年、インターネットの普及により、多くの情報がWeb 上に存在するようになった。その中には、用語説明などの記述も多く含まれ、同じ用語に対して様々な説明が混在している。例えば、「ハクビシン」という動物の体長について、「体長は約 61 - 66cm」のように直接的に表現する場合と、「タヌキと同じぐらいの体長」のように他の用語を媒介として間接的に表現する場合がある。

本研究は、上述した 2 つの表現、すなわち「直接表現」と「間接表現」の違いに着目する。間接表現は具体的には直喩もしくは比較による表現であり、媒介として使用される用語を「媒介語」と呼ぶことにする。なお、「媒介語」という用語は一般的な用法では、言語学習において意思疎通のために使用される学習対象以外の言語を指す場合がある。しかし、本論文内ではそれとは異なる意味で使用する点に注意を要する。

直接表現を用いた場合は、内容の正確性が高く誤解の余地が少ない。間接表現を用いた場合は、媒介語に関する知識があれば、対象の用語について直感的に理解することが容易になる。

このように、直接表現と間接表現にはそれぞれ別々の利点が存在するため、それぞれの表現を状況に応じて使い分けることには意義がある。そうした目的のための第一歩として、本研究では直接表現を利用した用語説明から、間接表現を利用した用語説明を生成する手法を提案する。

2 関連研究

本研究では直喩または比較を利用した間接表現の生成を行うため、比喩表現に関する研究と比較表現に関する研究に関連がある。よって、それらの研究について以下で解説する。

比喩に関する研究を大別すると、比喩表現とそうでない表現を区別する比喩認識の研究 [1]、比喩表現の意味的な分析を行う比喩解釈の研究 [1]、比喩表現の生成を目指す比喩生成の研究 [2, 3, 4, 5, 6] の三つに分けられ、本研究は比喩生成の研究に関連する。

比喩生成に関する研究 [2, 3, 4, 5] は、対象語に対する適切な説明を生成することを目的としている。具体的には、多くの語の意味、特徴、カテゴリなどの情報を記述したデータベースを用意し、説明対象の語と記述された情報が似ている別の一語を選択して比喩表現を生成する。しかし、前述のデータベース内の語は全てユーザーがよく知っている語であるという前提に基づいて比喩生成を行うため、比喩に利用する語の一般性を評価しない。語の一般性を考慮した比喩生成モデルを提唱した研究 [6] も存在するものの、比喩生成モデルの概要を述べただけにとどまり、具体的な計算方法の設定や比喩表現の生成には至っていない。

比較表現の生成に関する研究 [7] では、コンピュータによる学習支援システムにおいて分かりやすい説明を行うことを目標とし、速さや高さといった物理量について比較表現を生成する。媒介語は、ユーザーによって登録された語のリストから選択されるため、語の一般性を自動的に評価する手法は提案されていない。さらに、説明対象に近い物理量をもつ語を媒介語として優先的に利用して対象語との物理量の比を出力するため、2 以上の自然数 n に対し、対象語の n 倍もしくは $1/n$ というきりのいい物理量を持つ媒介語を優先的に利用できない。

3 提案する手法

本研究は、媒介語として利用する語の適切さを測る基準として、属性値類似度、語の類似度、把握容易度という尺度を提案する。この属性値類似度によって、対象語の n 倍もしくは $1/n$ に近い属性値を持つ媒介語を優先して利用可能となる。さらに、把握容易度によって語の一般性を反映できる。この三つについて、高い値を持つ媒介語を利用した間接表現ほど適切であるとして、以下の流れで間接表現を生成する。

1. **入力として対象語と属性-属性値の組を受け取る**

   例えば、「ハクビシンの体長は約 61 - 66cm である」という直接表現から間接表現を生成する場合、入力する対象語は「ハクビシン」、属性-属性値の組は、属性「体長」- 属性値「約 61 - 66cm」となる。また、「ハクビシンは体の色が灰褐色で、足の色は黒色である」などの、属性-属性値の組を複数含む直接表現からの間接表現生成も考慮し、複数の属性-属性値の組の入力も許す。

2. **媒介語候補を収集する**

   あらかじめ多くの語の属性-属性値ついて、< 語, 属性, 属性値 > という形で記述したデータベースを、人手もしくは既存の自動手法 [8] により構築し

ておく。以下、本論文ではこのデータベースを属性データベースと呼ぶ。この属性データベースから、入力として与えられた属性を持つ語を媒介語候補として収集する。例えば属性データベース内に、< タヌキ, 体長, 約 50-60cm>、< コイ, 体長, 60cm 程度 >、< キツネ, 体重, 5.2-5.9kg> という記述が存在したとする。この場合、入力された属性が「体長」であれば、この属性とそれに対応する属性値を持つ「タヌキ」と「キツネ」が媒介語候補として収集される。

3. **各媒介語候補ごとに、属性値類似度、語の類似度、把握容易度を計算する**

   対象語の複数の属性について、一語の媒介語を用いて間接表現を生成する場合、対象語と類似した属性値をより多くもつ媒介語を使用するのが望ましい。また、対象語の属性 i について、媒介語 m と 2 以上の自然数 n を利用して「m の n 倍 $(1/n)$ の i」と表現する場合には、n が小さく、かつ対象語と媒介語の属性値の比が n または $1/n$ に近いほど適切な表現になる。そのため、こういった媒介語の適切性を反映する値として、属性値類似度を計算する。

   媒介語が対象語と同義でない場合においては、対象語と概念的により類似した媒介語候補ほど、間接表現への利用に適する。例えば、「ハクビシンの体長は約 61 - 66cm」という直接表現から、< コイ, 体長, 60cm 程度 > と < タヌキ, 体長, 約 50-60cm> のどちらかの語を媒介語として間接表現を生成する場合を考える。この場合、「コイと同じぐらいの大きさ」と表現するよりも、「ハクビシン」により概念的に近い語である「タヌキ」を利用して「タヌキと同じぐらいの大きさ」と表現するほうが適切である。そのため、こういった媒介語の適切性を反映する値として、語の類似度を計算する。

   一般性の高い媒介語候補ほど、間接表現への利用に適する。例えば媒介語候補として「キツネ」と「フクロギツネ」が存在した場合、より多くの人がよく知っているであろう「キツネ」を使用するほうが適切である。そのため、こういった媒介語の適切性を反映する値として、把握容易度を計算する。

4. 3. **の値から媒介語候補ごとのスコアを計算する**

   その媒介語候補を使用して間接表現を生成した際の適切さを表すスコアを、式 (1) で計算する。

   $$\text{属性値類似度} \times \text{語の類似度} \times \text{把握容易度} \qquad (1)$$

5. **間接表現を生成して出力する**

   最大のスコアを持つ媒介語候補 M と入力された属性 i を用いて、「M のような i」、「M と同じくらいの i」、「M の n 倍の i」、「M の $1/n$ の i」のいずれかの間接表現を生成する。(n は 2 以上の自然数)

以下で、3. のそれぞれの尺度の計算方法について、それぞれ 3.1 節、3.2 節、3.3 節で解説していく。

3.1 属性値類似度の計算方法

前述の属性データベースを用いて計算する。媒介語候補 m の属性値類似度 $S(m)$ は、入力と m に共通する一つの属性 i に対する個別の属性値類似度を $s(i)$ とし、入力された属性の総数を I として、式 (2) で計算される。

$$S(m) = \frac{\sum_i s(i)}{I} \qquad (2)$$

$s(i)$ の計算方法は、対象語と媒介語候補の属性値が比較可能であるかどうかで、計算方法が変わる。

二つの属性値がどちらも数値を含み、なおかつ属性値の単位 (kg、 cm など) が一致する場合は、それらの属性値は比較可能であるとする。この場合、二つの属性値の数値部分のみを抜き出し、二つのうち大きいほうの数値を V とし、小さいほうの数値を v とする。なお、示された属性値が「約 61 - 66cm」などのように幅を持つ場合、その平均を属性値として扱う。$V/v = r$ とし、r を小数点第一位で四捨五入した自然数を n とする。この媒介語候補を利用して生成する間接表現は、n が 1 の場合「M と同じぐらいの i」という形となり、n が 2 以上の場合「M の n 倍 $(1/n)$ の i」という形となり。$s(i)$ を、[0,1] の値を取る定数 a と、自然数 N を用いて、式 (3) で計算する。

$$s(i) = \begin{cases} a^{n\ \ 1} \times (1 - 2 \times |r - n|) & (n \quad N) \\ 0 & (n > N) \end{cases} \qquad (3)$$

二つの属性値が比較可能でない場合は、二つの属性値が一致すれば $s(i) = 1$ とし、一致しなければ $s(i) = 0$ とする。$s(i) = 1$ となる i について生成する表現は、「M のような i」という形をとる。

3.2 語の類似度の計算方法

シソーラスなどの語の階層的な分類情報から、対象語と媒介語を共通して含む最少範囲の分類を探し、その分類に対応する定数を与えることによって決定する。その分類の範囲が狭いほど対応する定数は大きくなるように、あらかじめ決めておく。例えば、二つの動物名について語の類似度を計算する場合には、動物学的な分類体系を利用する。まず、あらかじめ「界」「門」「綱」「目」「科」「属」「種」の七種の分類階級に対応する定数を決めておく。そして、この七種の分類階級から、二つの動物を共通して含む最少範囲の分類を探し、それに対応する定数を語の類似度とする。「ハクビシン」と「タヌキ」の語の類似度の計算であれば、この二つの動物を含む最少範囲の分類は「食肉目」であるので、分類階級「目」に対応する定数を与える。

3.3 把握容易度の計算方法

Wikipedia の記事において他の語を説明する際により多く参照される語ほど一般性が高い、という考えに基づき、Wikipedia 内での語に対応する記事の被リンク数を利用して計算する。属性値データベース内の全ての語に

ついて Wikipedia での被リンク数を調べる。そこで出現した被リンク数の最大値で、各語の被リンク数を割って正規化したものを把握容易度とする。

4 本手法の実行例

本手法で提案した三つの尺度が適切な間接表現の生成に貢献することを確認するために、特定の入力に対して間接表現の出力を行った。属性データベース、各語の階層的な分類情報、各分類に対応する定数については、簡易的なものを用意して使用した。なおこの実行例においては、対象語と媒介語を全て動物名とし、属性値類似性の計算式 (3) におけるパラメータは、$a = 0.7$、$N = 4$ とした。

入力は「ハクビシン」という動物の、色に関する複数の属性、体長、体重、及びそれらに対応する属性値とした。詳細は表 1 の通りである。入力 1 では 5 組の属性-属性値の組を一度に入力し、入力 2 と入力 3 では単一の属性-属性値の組を入力する。なお、ここでの属性-属性値の情報は、WIkipedia の「ハクビシン」の記事から人手で抽出した。

属性データベースは、いくつかの生物名について Wikipedia の該当記事に書かれた情報から、色に関する情報と、体長、体重の情報を抽出することで構築した。詳細は表 2 の通りである。

各語の階層的な分類情報として、対象語「ハクビシン」と属性データベースに存在する生物名の全てについて、「界」「門」「綱」「目」「科」「属」「種」の分類情報を Wikipedia のインフォボックスから抽出した。また、各分類階級に対応する定数は、今回の実行においては表 3 のように定めた。

また、Wikipedia での各語の被リンク数と、それを正規化した把握容易度は表 4 の通りである。なお、表 4 における数値は、小数点第四位を四捨五入して記載した。

5 実行例に対する考察

入力 1〜3 について、生成された間接表現は表 5 のようになった。さらに、各入力に対して収集された媒介語候補のスコアと三つの尺度の詳細は、表 6、表 7、表 8 の通りである。なお、この三つの表における数値は、小数点第四位で四捨五入したものを記載した。 以下、各入力に対する出力結果についての考察を行う。

入力 1 では、属性値類似度の効果により、媒介語として「オオカミ」ではなく「タヌキ」が選択されている。入力 1 に対する各媒介語候補のスコアの詳細は、表 6 の通りである。ここで媒介語候補「タヌキ」と「オオカミ」の各尺度の値とスコアに注目する。語の類似度と把握容易度においては、「オオカミ」が「タヌキ」以上の値を示している。しかし、属性値類似度において「タヌキ」が「オオカミ」より大きな値をもっているため、「タヌキ」が全体のスコアにおいて「オオカミ」を上回り、媒介語として選択されている。このことから、属性値類似度がより適切な間接表現の生成に貢献するケースの存在が確認できた。

表 1: 入力の詳細

	対象語	属性	属性値
入力 1	ハクビシン	体の色	灰褐色
		頭の色	黒色
		足の色	黒色
		尾の色	黒色
		頬の色	白色
入力 2	ハクビシン	体長	約 61 - 66cm
入力 3	ハクビシン	体重	2 - 3kg

表 2: 属性データベース

生物名	属性	属性値
タヌキ	体の色	灰褐色
	目の周りの色	黒色
	足の色	黒色
	体長	約 50-60cm
	体重	3-10kg
キツネ	体重	5.2-5.9kg
フクロギツネ	体の色	灰色
	腹の色	乳白色
	尾の色	黒色
	鼻の色	ピンク色
	体長	35-55cm
	体重	1.2-4.5kg
オオカミ	体の色	灰褐色
	体重	25 - 50kg
コイ	体長	60cm 程度

表 3: 各分類階級に対応する定数

	界	門	綱	目	科	属	種
対応する定数	0.4	0.5	0.6	0.7	0.8	0.9	1.0

表 4: 各語の被リンク数と把握容易度

語	被リンク数	把握容易度
タヌキ	350	0.680
キツネ	515	1
フクロギツネ	25	0.049
オオカミ	436	0.847
コイ	304	0.590

入力 2 では、語の類似度の効果により、媒介語として「コイ」ではなく「タヌキ」が選択されている。入力 2 に対する各媒介語候補のスコアの詳細は、表 7 の通りである。ここで媒介語候補「タヌキ」と「コイ」の各尺度の値とスコアに注目すると、属性値類似度の値は「コイ」のほうが高く、把握容易度は「タヌキ」のほうがわずかに高い。従って、この二つのみを用いてスコアを計算した場合、媒介語として「コイ」が選択される。しかし、語の類似性において「タヌキ」が「コイ」を上回っているため、「タヌキ」が全体のスコアにおいて「コイ」を上回り、媒介語として選択されている。このことから、語の類似度が、より適切な間接表現の生成に貢献するケースの存在が確認できた。

入力 3 では、把握容易度の効果により、媒介語として「フクロギツネ」ではなく「キツネ」が選択されている。入力 3 に対する各媒介語候補のスコアの詳細は、表 8 の

表 5: 各入力に対する出力結果

入力	出力された間接表現
入力 1	「タヌキのような体の色, 足の色」
入力 2	「タヌキと同じぐらいの体長」
入力 3	「キツネの 1/2 の体重」

表 6: 入力 1 に対する媒介語候補のスコア詳細

媒介語候補	属性値類似度	語の類似度	把握容易度	スコア
タヌキ	0.4	0.7	0.680	0.190
フクロギツネ	0.2	0.6	0.049	0.006
オオカミ	0.2	0.7	0.847	0.119

表 7: 入力 2 に対する媒介語候補のスコア詳細

媒介語候補	属性値類似度	語の類似度	把握容易度	スコア
タヌキ	0.691	0.7	0.680	0.329
フクロギツネ	0.178	0.6	0.049	0.005
コイ	0.883	0.5	0.590	0.261

表 8: 入力 3 に対する媒介語候補のスコア詳細

媒介語候補	属性値類似度	語の類似度	把握容易度	スコア
タヌキ	0.098	0.7	0.680	0.047
キツネ	0.392	0.7	1	0.274
フクロギツネ	0.72	0.6	0.049	0.021
オオカミ	0	0.7	0.847	0

通りである。ここで媒介語候補「キツネ」と「フクロギツネ」の各尺度の値とスコアに注目すると、属性値類似度の値は「フクロギツネ」のほうが高く、語の類似度は「キツネ」のほうがわずかに高い。従って、この二つのみを用いてスコアを計算した場合、媒介語として「フクロギツネ」が選択される。しかし、把握容易度において「キツネ」が「フクロギツネ」を大きく上回っているため、「キツネ」が全体のスコアにおいて「フクロギツネ」を上回り、媒介語として選択されている。このことから、属性値類似度がより適切な間接表現の生成に貢献するケースの存在が確認できた。

6 おわりに

本論文では、直接表現から間接表現を生成するために、媒介語を選択するための三つの基準を提案した。そして、それらの尺度を利用した間接表現生成手法の概要を示し、簡易的なデータ用いて実行を試みた。その結果、提案した三つの尺度について最低限の有効性を確認できた。今後の課題は、より大規模なデータを用意しての実験を行い、その結果をもと三つの尺度とスコアの計算方法を経験的に改良していくことである。

謝辞

本研究の一部は、科学研究費補助金基盤研究（B）（課題番号 22300050）によって実施された。

参考文献

[1] Ekaterina Shutova, Simone Teufel, and Anna Korhonen. Statistical metaphor processing. *Computational Linguistics Accepted for publication*, pp. 1–92, 8 2012.

[2] 蔵方隆宏, 石田修一, 森辰則. 用語説明システムにおける比喩説明文の生成. 言語処理学会 第 2 回年次大会 発表論文集, pp. 217–220, 3 1996.

[3] 小澤宏也, 岡本紘幸, 斎藤博昭. 色・形状情報を用いた比喩生成. 言語処理学会 第 13 回年次大会 発表論文集, pp. 222–225, 3 2007.

[4] 酒匂孝之, 中村順一, 吉田將. 概念間の外見的な類似性と心理的な評価を利用した比喩表現の生成. *IEICE technical report. Natural language understanding and models of communicaton 93(131)*, pp. 17–24, 7 1993.

[5] 阿部慶賀. 言語統計解析に基づく比喩表現生成システムの構築. 青山インフォメージョン・サイエンス, Vol. 37, No. 1, pp. 55–90, 2009.

[6] 山崎澄, 桝井文人, 河合敦夫, 椎野努. 比喩生成モデルにおける単語親密度の有効性について. 信学技報 NLC2001-10, pp. 85–90, 10 2001.

[7] 土門真城, 中原浩昭, 下大澤孝晴, 藤井敏史. CAI における比較表現の生成. Proceedings of the IEICE General Conference 1995 年. 情報・システム (1), p. 341, 3 1995.

[8] Weiyi Meng Hai He, Clement Yu, and Zonghuan Wu. Automatic integration of web search interfaces with WISE-Integrater. *VLDB Journal*, Vol. 13, No. 3, pp. 256–273, 2004.