4R-6

概念辞書を用いた比喩性判定

吉田　光一† 　岸　義樹‡

†茨城大学大学院理工学研究科情報工学専攻　　‡茨城大学工学部情報工学科

1　はじめに

計算機による理解が難しい言語現象の一つに比喩表現が存在する．比喩とは聞き手に新しい概念を理解させ情報を円滑に伝達する手段であり，人間の概念体系に深く関わっている．我々の使用する言語には比喩的な表現が多く含まれており，機械翻訳や検索など自然言語処理技術で言語の意味や文脈を扱っていくためには，比喩理解の問題は避けては通れない部分と言える．そのため，自然言語処理において自動的に比喩文を認識する必要があると考えられる．

そこで本研究では，EDR 電子化辞書の概念情報を用いて，比喩表現の中でも比喩指標を含んだ基本的な直喩表現の文を対象に，その文が比喩表現であるかリテラル（字義通り）な文であるかを判定する手法を提案する．本手法では構造として「A のような B」という文のみ扱うとし，A,B についてその単語間の比喩性を計る．

2　EDR 電子化辞書

EDR 電子化辞書は単語辞書，概念辞書，対訳辞書，共起辞書，専門用語辞書，EDR コーパスから構成されており，本研究では主に概念辞書を用いる．

2.1　概念辞書

概念辞書は単語辞書などの各辞書から参照される概念を規定するための辞書であり，約 41 万の概念についての知識が記述されている．

2.2　概念体系辞書

概念体系辞書は概念間の関係のうち，特に上位-下位関係を用いて概念全体を体系化したものである．この上下関係の構造は概念間の包含関係にあり，その主要部は５つの基本語概念体系の項目が存在する．

Calculation of similarity using a concept dictionary
Kouichi Yoshida†, Yoshiki Kishi‡
†‡Ibaraki University
4-12-1 Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan

3　提案手法

本手法の流れは以下の通りである．

1. 文章データを用意し,「のような」が含まれる文を収集する
2. 比喩性を判定する二単語を抽出する
3. リテラル語における判定を行う
4. 概念体系上位層によるカテゴリー分類を行う
5. 概念情報による類似度計算を行う
6. 4,5 における情報から比喩性を判定する

3.1　処理 1．扱う文章データ

文章データは国立国語研究所の現代日本語書き言葉均衡コーパス [3] の中から新聞，雑誌および書籍のデータを利用する．また,「このような，あのような，…」といった指示代名詞を含む文は扱わないものとする．

3.2　処理 2．扱う単語

判定する単語は修飾語句のついていない名詞に限定し,「〜のようなこと、もの」といった非自立語の名詞は曖昧性が生じるためあらかじめ判定対象としないものとする．

3.3　処理 3．リテラル語による判定

判定には抽出した二単語を扱うが，比較せずにリテラルと判定できるような例示に用いられる語がいくつか存在する．表 1 に示すような語を含む文はあらかじめリテラルであると判定する．

種類	例
代名詞	僕，あなた，…
文の場所を指す語	図，次，上記，…
鉤括弧	「」,『』，…

表 1: リテラル語

3.4　処理 4．概念の上位層によるカテゴリー分類

A のような B における喩辞 (A)，被喩辞 (B) それぞれについて基本語概念体系の上位 2 レベルまで上位概念を求め，2 レベル目の項目によるカテゴリーで分類を行う．ただし，項目「もの」に関しては多くの単語

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.

が分類される傾向にあったため，さらに下位概念の項目で分類を行うとする．全23項目の上位層概念で分類し，喩辞 (A) において「昨日」「従来」といった時間点に分類されるものは，リテラルな文であると判定する．

3.5 処理5．類似度計算

抽出された二単語について，次の手順で単語間の類似度を計算する．

1）類似度 α の計算

A のような B において，各単語の概念集合をそれぞれ C_A，C_B としたとき，C_A と C_B 間に存在する共通の概念数によって α を計算する．集合 X に存在する要素の数を $|X|$ としたときのαの式を以下に示す．

$$\alpha = |C_A \cap C_B|$$

2）類似度 β の計算

A のような B において，各単語の k 段目の上位概念からなる集合をそれぞれ S_k^A，S_k^B としたとき，階層 k での類似度 β_k を計算する．各単語の k 段目の上位概念集合内の異なる概念数をそれぞれ N_{kA}，N_{kB}，S_k^A と S_k^B の間の共通な上位概念の数を CS_k としたときの β_k の式を以下に示す．

$$\beta_k = (1 + K_{\beta 1} * CS_k)(1 + K_{\beta 2} * (\frac{CS_k}{N_{kA}} + \frac{CS_k}{N_{kB}})) - 1$$

$K_k(k = 1, 2, \cdots N)$ を用いて各段階での類似度の重みを調整し，β を計算する式を以下に示す．

$$\beta = K_1 * \beta_1 + K_2 * \beta_2 + \cdots + K_N * \beta_N$$

3）類似度 δ の計算

最終的な類似度となる δ を計算する式を以下に示す．

$$\delta = 1 - e^{-(K_\alpha * \alpha + K_\beta * \beta)}$$

3.6 処理6．比喩性の判定

カテゴリー分類においては，同一カテゴリーならば包含関係にあり例示を表している可能性が高く，異なったカテゴリー関係ならば比喩が使われている可能性が高いと判断する．

類似度は0〜1の値をとり，1に近いほど単語間の類似度は高く，リテラルである可能性が高いと判断する．

4 手法の評価

4.1 内容

現代日本語書き言葉均衡コーパスから「A のような B」で表現された文を100抽出し，人間による判断と比較し評価を行う．各文において，比喩性が高い，低い，およびリテラルであると判定する．

また，類似度 β の計算は上位二段目まで，各式の重みはそれぞれ $K_{\beta 1} = 2.5, K_{\beta 2} = 8.0, K_1 = 1.0, K_2 = 0.5, K_\alpha = 0.4, K_\beta = 0.03$ として計算を行った．

4.1.1 評価手法

各文において，異なったカテゴリーに分類され，かつ類似度の値が0.3以下のものを比喩性が高いと判定し，同一カテゴリーかつ類似度の値が0.7以上となったものをリテラルな文と判定する．以上に含まないものは比喩性の低い文と判定する．

4.2 結果および考察

抽出した100の文の表現のうち人間が高い比喩性と判断したのは31，低いと判断したのは12，リテラルな文と判断したのは57となった．評価手法によってシステムが高い比喩性と判定したのは39，低いと判定したのは15，リテラルな文と判定したのは46であった．各表現の再現率，適合率，F 値を表2に示す．

	比喩性高い	比喩性低い	リテラル
再現率	0.641	0.40	1.0
適合率	0.806	0.50	0.807
F 値	0.714	0.444	0.893

表 2: 各表現の再現率，適合率，F 値

結果から，判定したリテラルの多くはリテラル語による分類が多かったが，典型的な例示を示す文や高い比喩性を持つ文についてはある程度人間に近い判定ができていると考えられる．比喩性の低い文については人間的な見方でも曖昧性があったことや，評価手法による信頼性が低かったことが伺える．

5 まとめ

本研究では概念辞書を用いて比喩性を判定する手法を提案し，人間による判断と比較してある程度比喩表現が使われているか，リテラルな文であるかが判定できることを確認した．

参考文献

[1] 崔進，小松英二，安原宏：EDR 電子化辞書を用いた単語類似度計算法，情報処理学会研究報告自然言語処理（NL）Vol. 93 , No. 1 pp.1-6, 1993

[2] 田添丈博，桝井文人，椎野努：’名詞’ のような’名詞’ の分類と比喩性の判定モデル，情報処理学会研究報告，2001(9), pp.27-32

[3] コーパス整備計画 KOTONOHA
http://www.ninjal.ac.jp/kotonoha/

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.