HOKKAIDO UNIVERSITY

Title	Webデータを用いた単語に対する比喩的説明文収集手法
Author(s)	近村, 亮一; ジェプカ, ラファウ; 荒木, 健治
Description	第27回ファジィシステムシンポジウム(27th Fuzzy System Symposium). 2011年9月12日～14日. 福井大学文京キャンパス, 福井市.
Citation	ファジィシステムシンポジウム講演論文集, 27, 112-115 https://doi.org/10.14864/fss.27.0.27.0
Issue Date	2011-09-12
Doc URL	https://hdl.handle.net/2115/63615
Type	conference paper
File Information	FSS27-2011_112-115.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Web データを用いた単語に対する比喩的説明文収集手法

A Method of Collecting Figurative Explanations from the Web

○近村　亮一　　ジェプカ　ラファウ　　荒木　健治
Ryoichi Chikamura　　Rafal Rzepka　　Kenji Araki

北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

Abstract: In this paper, we propose the system extracting figurative explanations from the web. Our system collects candidates for figurative descriptions for the target words using the Web, and extracts a word that shows to what it should be compared from those candidates. Each extracted word are set as "word of comparison" in each extraction source candidate. Then, our system selects appropriate candidates by a similarity calculation proposed in this paper, and outputs selected candidate. We evaluate the validity of our system output.

1. はじめに

私達が文章を読むときに，意味の知らない単語に出くわすことがある．このとき，辞書を引くといったことで単語の意味を調べ，理解することが必要となる．しかし，辞書を引いた際，そこに記載されている説明を理解するための知識がない等により，単語の意味を調べるためにまた辞書を引き直す，または調べたが全くわからない結果に終わるという場合がある．

そのような問題を解決する方法として，比喩による説明に着目した．これは比喩を用いた説明は，日常的に広く用いられているからである．本研究の目的は，単語に対し，比喩による説明の付与を通し，単語の意味の直観的な理解を促すことである．

2. 関連研究

比喩を用いた説明生成の研究として，事前に整備した知識ベースと比喩テンプレートを用いて比喩表現の生成を行うものがある[1]．

これに対して我々は，知識ベース構築等のコストを削減するため， Web データをもとにした，比喩説明の自動収集・及び生成を目指す．そこで本稿では，比喩説明の自動収集について述べる．

3. 提案手法

まず,Web 検索エンジンを用いて入力単語(名詞)に関連があると考えられる比喩表現を収集する．具体的には以下の計４種のクエリで検索エンジン[2]から検索結果のスニペットを 20 件ずつ取得する．

- 「“入力単語とは”　“に例え”」
- 「“入力単語”　“に例え”」
- 「“入力単語とは”　“のような”」
- 「“入力単語”　“のような”」

得られたスニペットを句読点,省略記号（“...”等）を文の区切りとし複数の文に分割する．各文の中から「に例え」「のような」を含むものを候補文として抽出する．候補文各々に CaboCha[3]で係り受け解析を行い，「例え」を含む文節に係る直近の文節の中の名詞,及び「のような」を含む文節から「のような」の直前に現れる名詞を抽出する．つまり，「に例え」「のような」の直前に現れる名詞を抽出する処理を行う．これを文に対する“例えの語”として設定する．以下に処理による設定例を表 1 に示す．

表 1.“例えの語”の設定例

候補文	例えの語
水に/例えますと、/池に/貯まって…	水
…電流は、/水に例えて/考えると…	水
電圧は/よく水圧/に例えられる/…	水圧
…川の流れのような電気…	流れ

次に,比喩文それぞれに対して設定された“例えの語”と入力単語の間での関連性を計算する．これは候補文が入力単語についての比喩であるかどうかを判断するためである．入力単語と“例えの語”が比喩関係にあると判断できれば，“例えの語”に結びつく文が入力単語に対する比喩文であると判断できる．

また，加えて“比喩による説明文”であることを考慮しなければならない．そこで，関連度を測るために接続動詞とその格関係及び動詞の目的語や補語の情報を用いる方法を提案する．これは，ものを説明する際には，対象となるものの動作や役割，働きといった面に着目した例えがなされる場合が多いと考えたためである．具体例としては，よく CPU は脳に例えた説明がなされるが，これは CPU が情報を処理する等，脳と同様の役割を持つ

ためであると考えられる．以上の観点に基づき，あるものの動作や役割，働きを表現するためには，あるものに接続する動詞が適当と考えた．動詞だけでなく格関係等の文脈情報を用いるのは，動詞の用法が文脈により異なるものが多く，動作等を表現するものとして用いる場合は用法による区別が必要と考えたためである．

名詞単語に接続する動詞，及び文脈情報の収集には，Web 検索を用いる等の方法も考えられるが，ある程度網羅的に収集するのは，検索や処理コストの面から容易ではない．そこで京大格フレーム検索[4]を利用することにした．京大格フレーム検索では，Web コーパスから収集した名詞に対する接続動詞とその時の格関係，及び頻度が登録されており，また，その用法についてクラスタリングによる分類情報が登録されている[5]ためである．

4. 関連度計算

入力単語と先の方法で収集した，“例えの語”候補一つをそれぞれ京大格フレーム検索により接続動詞の情報（動詞，格関係，用法の分類番号）を収集する．これにより，語に対し格ごとに頻度で並べ替えられた接続動詞群が得られる．次いで，入力単語の動詞群の中に，“例えの語”の動詞群にも存在するものの割合を格ごとに計算する．この総和を関連度とした．対象とする格の種類は，登録されている全十種から“ノ格”を除く計 9 種とした．これは，“ノ格”が動詞との関係を表すものというよりも，“A の B”といった名詞同士の関係を示す側面が強く，動詞との関係を示すには不適当と考えたためである．

5. 出力

候補収集，関連度計算の後，関連度の高かった“例えの語”が持つ文を提示する．現段階では一つの“例えの語”に対して複数の文が設定されている場合がある，したがってその中から最も説明として適当なものを選び出す処理が必要である．この点についてはより深い文脈処理が必要と考えており，今後の課題である．そのため，今回は設定される文すべてを出力することとした．

6. 評価及び考察

6-1. システム出力について

入力単語に対して“比喩による説明”を収集できたか，という点について評価する．入力単語は，

A)eWords[6]月間注目ランキング上位 20 語
B)中学理科用語[7]からランダムに 20 語
C)Weblio[8]ヘルスケア・キーワードランキングから上位 10 語

の 3 種のカテゴリから計 50 個を使用した．出力できたものは 34 件で，その内訳を表 2 に示す．

表 2. システムの出力数

A	B	C	計
16(80%)	13(65%)	5 (50%)	34(68%)

出力できなかったものは全て入力単語を格フレーム検索した際に結果が得られないものであった．

一つの入力単語に対して，先の関連度計算により順位付けられた”例えの語”上位 3 つを選定し，3 つの”例えの語”に関連付けられた文全てを入力単語に対する“例えの説明”として出力する．一つの入力単語に対し，システムが抽出した候補文数と設定した”例えの語”の異なり数，及び出力する文数を表 3 に示す．

表 3. 1 入力あたりの候補文，例えの語，出力文

候補文数	“例えの語”数	出力文数
44.89	27.9	3.74

候補文数に対して，“例えの語”の数が 6 割程度となった，このことから，候補文から設定される“例えの語”は平均して 1.6 語程度の重複が存在する．

6-2. システム出力の妥当性評価

理系男子大学院生４名に，入力単語とそれに対する出力文をセットで提示し，入力単語に対しての“例えの説明”として，それぞれの文が説明として妥当かどうかを判定してもらった．客観的な判定基準の設定が困難なため，基本的には主観に任せることとした．以下に，全ての被験者が比喩の説明として妥当とした文を含む出力が得られた割合，同様に 2 名以上が妥当と判断した文を含む出力が得られた数とその割合を表 4 に示す．

表 4. 関連度による出力で妥当なもの含む割合

全員	2 名以上
8(23.5%)	13(38.2%)

候補文からの“例えの語”抽出時には，1 文に対して 1 語が抽出される．この時，前述したように“例えの語”が重複する場合も多い．この時の重複した回数を“例えの語”の頻度としている．以下に頻度の設定例を表 5 に示す．

表 5. 頻度の設定例

設定された候補文	例えの語	頻度
水に例えますと、池に貯ま…	水	2
…電流は、水に例えて考え…		
電圧はよく水圧に例えら…	水圧	1
…川の流れのような電気…	流れ	1

そこで比較のため，“例えの語”をこの頻度により順位付けを行い，上位 3 個の“例えの語”に設定される候補文を全て出力し，同様に評価を行った．頻度による順位付け時の平均出力文数は，一つの“例えの語”に対する候補文の数がここでいう頻度に一致するという理由から，関連度計算による方法よりも高い 6.54 文であった．なお，抽出される候補文の数，及び“例えの語”の異なり数は表 2 に示したものと同じである．表 6 にこの結果を示す．

表 6．頻度による出力に妥当なものを含む割合

全員	2 名以上
7(20.6%)	16(47%)

関連度計算により提示した出力，頻度により提示した出力について考察する．関連度計算を用いた場合，頻度を用いた場合での成功数という点では大きな違いは確認できなかった．

頻度による出力と，関連度計算による出力での差異を確認するため，評価揺れの少ないと考えられる全ての被験者が妥当とした文を含む出力が得られた場合について考察する．

4 人の被験者が妥当と判断した文を正解とし，同じ入力単語に対する，頻度による方法と関連度計算による方法の出力の間で正解が重複したものは 5 件存在した．重複しなかったものについて，その理由を考察する．

頻度による方法で出力した場合のみに正解が含まれていたケースは 2 件，関連度計算による方法で出力した場合でのみ正解が含まれたケースは 3 件あった．これらについて，システムが行った“例えの語”の順位付けを確認する．

頻度による方法で出力した場合のみに正解が含まれていたケース 2 件について，そのうち 1 件は正解とされた文に対する“例えの語”が格フレーム検索に失敗したことにより，関連度が計算できなかったため頻度による方法でのみ出力されたものであった．残りの 1 件については関連度計算による方法では 4 位に正解があったため，出力に現れなかったケースであった．この結果から，関連度計算よりも，頻度による順位付けが有効であるケースが存在することが確認された．

関連度計算による方法のみで正解が出力された 3 件は，“例えの語”の頻度にばらつきが多く，頻度では適切に順位付けできないケースであった．

関連度計算による方法で出力した場合でのみ正解が含まれたケース 3 件のうち 1 件についてシステムが判断に用いた“例えの語”上位 3 件の具体例を表 7 に示す．なお頻度による順位付けでは 1 位から 3 位まで同頻度（2 回）であった．

表 7．頻度に偏りのない時のシステム内での順位

入力単語：SNS			
	1 位	2 位	3 位
頻度	mixi	ある形	恐竜
関連度	ブログ	mixi	**サークル**

表 7 に示したそれぞれの“例えの語”から出力される文を表 8 に示す．

表 8．“例えの語”（表 7）に設定された候補文

例えの語	結びつく候補文
mixi	・上に挙げた mixi は国内最大手の SNS ですが、現在は SNS に会員として参加するのではなく、SNS 自体を作成するサービス、つまり mixi のようなサイトを自分で作成し、
	・昨年 12 月 1 日から、「改正出会い系サイト規制法」が順次施行されていて、今年 2 月 1 日から、出会い系サイトでは、登録時の年齢確認が義務化されたそうですが、mixi のようないわゆる一般的な SNS は、この出会い系サイトにははいらないという
ある形	・問八幡市域の形は、ある形に例えられるが、それはどんな形に似ているか
	・八幡市域の形は、ある形に例えられるが、それはどんな形に似て
恐竜	・5 巨大化する SNS は、やがて絶滅した恐竜のような運命に
	・恐竜のような巨大な SNS の時代ではない
ブログ	・ブログのような印象を受ける
サークル	**・同じ趣味仲間の他にも、年代別や地域別に仲間を集めることができる、ちょっとしたサークルのような機能だと思えば分かりやすいと思います**

4 名の被験者全員が妥当と判断したものは，“サークル”から得られた説明文である．表 5 に記載した他の“例えの語”から得られた文の中には，2 名以上が妥当だと判断したものはなかった．

以上から，抽出される“例えの語”の頻度に偏りがない場合は，関連度計算による順位の方が有効である可能性が高く，偏りがある場合には頻度情報も有効であることが確認された．

4 人の被験者が妥当と判断した文について，頻度による方法のみに含まれたもの，関連度計算による方法のみに含まれたもの，共通して含まれたものとその時の入力単語については巻末付録，表 9-11 に示す．

7. まとめ

ユーザの直感的理解を促すため，単語の比喩による説明をWebから収集する方法について検討した．接続動詞とその格関係，及び動詞の近似的な文脈による用法という情報を用いて，説明文脈という点を考慮した単語同士の比喩としての関連度計算方法を提案し，評価を行った．その結果，頻度を用いることで改善できるケースも存在することが確認できた．

今後の課題としては，5章でも述べたが，“例えの語”に対して複数の候補文が設定されている場合も多いので，そのような場合に対して，適切なものを選択できるようにする必要がある．現段階でこの方法については明言が出来ないが，候補文収集にはWeb検索エンジンを用いているので，候補文が収集されたスニペットの検索順位情報を用いるといった方法が考えられる．

また，本稿での出力の評価は“妥当さ”に基づいたものであり，“わかりやすさ”については考慮できていない．今後はその部分について調査を行う必要があると考えている．

8. 参考文献

[1]石田 修一 ， 蔵方 隆宏 ， 森 辰則：知的マニュアルシステムにおける比喩説明文の生成，情報処理学会研究報告.自然言語処理研究会 6(7),1-6

[2]YahooAPI ：http://developer.yahoo.co.jp/

[3]CaboCha：http://chasen.org/~taku/software/cabocha

[4]京大格フレーム検索 http://reed.kuee.kyoto-u.ac.jp/cf-search

[5] 河原 大輔，黒橋 禎夫：高性能計算環境を用いたWebからの大規模格フレーム構築情報処理学会研究報告. 自然言語処理研究会報告 2006(1), 67-73, 2006-01-12

[6] IT用語辞典 e-words：http://e-words.jp/

[7]仙台市立六郷中学校HP 理科のページ：http://www.sendai-c.ed.jp/~rokugojh/rika/rika.html

[8]weblio 辞書：http://www.weblio.jp/

連絡先

近村亮一
北海道大学情報科学研究科
E-mail: chika@media.eng.hokudai.ac.jp

付録

表9. 頻度による方法のみで出力に含まれた正解

入力単語	被験者全てが妥当とした候補文
ドメイン	・インターネット上の住所のようなもので、重複しないように発行・管理されている
アフィリエイト	・アフィリエイトとは、広告代理店のようなものです

表. 10 関連度計算による方法のみで出力に含まれた正解

入力単語	被験者全てが妥当とした候補文
SNS	・同じ趣味仲間の他にも、年代別や地域別に仲間を集めることができる、ちょっとしたサークルのような機能だと思えば分かりやすいと思います
スクリプト	・スクリプトとは、部品を操作する命令を箇条書きにしたテキストのようなものです
細胞膜	・しかし、実際は、限られた物質だけを通過させることができる「魔法のフィルム」のようなものです

表11.共通して出力に含まれた正解

入力単語	被験者全てが妥当とした候補文
VPN	・VPN(VirtualPrivateNetwork)とは不特定多数が使用する共有ネットワーク上に、あたかも専用線のようなネットワークを作り出す
API	・ちなみに、APIとは、ApplicationProgramInterfaceの略で、プログラムを集めたライブラリのようなものです
プロトコル	・ネットワークでコンピューターが使う言語のようなもので、双方が理解できる同じプロトコルを使わないと通信は成立しない
Cookie	・Cookieとはホームページを運営しているサーバが、あなたのパソコンに勝手に植え付けていく、ちょっとしたメモのようなものです
黒点	・黒点とは太陽表面に見えるシミのような部分で、強い磁場を持ち、周辺では「フレア」と呼ばれる大爆発が多く発生します