E-017

大規模格フレームを用いた概念ベースへの動詞属性の追加
Supplement of verb attributes to Concept Base using Case Frame

小泉 政弥 † 芋野 美紗子 † 土屋 誠司 ‡ 渡部 広一 ‡
Masaya Koizumi Misako Imono Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，情報処理システムの高性能化，高機能化に伴いその操作方法は複雑化を辿る一方であり，ユーザが特別な知識や技能を必要としないシステムの構築が求められる．そこで，人間が日常会話で使用する自然言語を用いて，人間同士が会話を行うように情報処理システムを扱うことが出来れば，ユーザの負担が軽減されると考えられる．本稿では概念ベース [1]や関連度計算方式 [2]を用いてこのシステムを実現している．

概念ベースとは複数の電子国語辞書から機械的に構築された大規模な知識ベースである．概念ベースを用いることで，ある語から他の語を連想することができる．しかし既存の概念ベースは名詞が大多数を占めており，概念に付与されている動詞属性が少ないため，動詞に対して適切な連想が行いづらいという問題がある．そこで本稿では概念ベースに対して動詞属性の追加を目的とする．

2. 関連技術

2.1 概念ベース

概念ベースでは様々な語を概念として定義しており，概念はその意味特徴を表す属性と，属性の重要性を表す重みの対の集合によって構成されている．ある概念 A は m 個の属性 a_i と重み w_i（>0）の対により次のように定義される．

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_m, w_m)\} \quad (1)$$

ここで属性 a_i を概念 A の一次属性と呼ぶ．これら一次属性は概念ベースの中で概念として定義されている語で構成されている．つまり概念 A の持つ属性 a_i を概念とみなし，更に属性を導くことができる．概念 a_i の持つ属性を元の概念 A の二次属性と呼ぶ．このように概念ベースは任意の次元までの属性連鎖集合により定義されている．

2.2 関連度計算方式

関連度計算方式とはある 2 つの概念間の関連の強さを定量的に表現する手法である．関連度は 0.0 から 1.0 の実数値で算出され，概念間の関連が強いほど高い数値となる．

2.3 大規模格フレーム

大規模格フレーム [3]とは，用言とそれに関係する名詞を用言の用法ごとに整理したものである．*Web* 上の約 5 億文の日本語テキストから自動的に構築されており，約 4 万用言が格納されている．なお，ここでいう用言とはサ変名詞，動詞，形容詞を表す．名詞から検索すると，その名詞に対する用言，格，頻度が出力される．頻度とは *Web* 上での出現頻度である．大規模格フレームには，約 2 万語の動詞が用言として登録されている．

† 同志社大学大学院理工学研究科
Graduate School of Engineering, Doshisha University
‡ 同志社大学理工学部
Faculty of Science and Technology, Doshisha University

3. 名詞概念の動詞属性評価

現在の概念ベースからランダムに 300 語の名詞概念を選び，その概念に付与されている動詞属性が妥当であるかを目視により評価した．評価は 11 名で行い，6 名以上が正しいと判断した属性を正解とした．品詞の判断には茶筌 [4]を用いた．評価の基準として，概念中で重みが最大の動詞属性が妥当だと判断した場合は○，妥当な動詞属性は存在するが，重みが最大の動詞属性は妥当でないと判断した場合は△，動詞属性が全て妥当でないと判断した場合は×とした．その結果，○が 31%，△が 41%，×が 28%となった．これにより，既存の概念ベース内は名詞概念の多くに適切な動詞属性が付与されておらず，また付与されている場合でも適切な重み付けがされていないと考えられる．

4. 大規模格フレームを用いた動詞属性の追加

概念ベースの名詞概念に対し属性追加を行う．本稿の目的は動詞属性の追加だが，サ変名詞も動詞と同じ働きをするため，サ変名詞属性を追加する場合も考慮する．そこで動詞のみを追加した場合，サ変名詞のみを追加した場合，両方を追加した場合の 3 種類の概念ベースを作成する．

4.1 大規模格フレームによる属性追加候補語の選出

大規模格フレームにより名詞概念を検索して出力された用言に対して形態素解析を行い，動詞及びサ変名詞を抽出した．その内，既存の概念ベースに概念として登録されている語のみを選別し，属性追加候補語とした．名詞概念「椅子」に対する属性追加候補語の一部を表 1 に示す．

表 1　概念「椅子」の属性追加候補語(一部)

概念	属性追加候補語	候補数
椅子	座る，腰掛ける，立ち上がる，…，買う，並ぶ，…，奪う，投げる，…	978 語

4.2 閾値の設定

属性追加候補語は膨大な量があり雑音も含まれる．そこで大規模格フレームにおける頻度に以下の 3 種類の閾値を設け，閾値以上の頻度を持つ語を属性として追加する．

① 属性追加候補語に付与されている頻度の平均値
② ①で選別された語に付与されている頻度の平均値
③ ②で選別された語に付与されている頻度の平均値

閾値を用いて選別した追加属性の比較例を表 2 に示す．

表 2　閾値による追加属性の比較

概念	閾値	追加属性	追加数
椅子	①	座る，…，買う，…，奪う	102 語
	②	座る，…立ち上がる，…買う	12 語
	③	座る，腰掛ける	2 語

大規模格フレームでは「痛みを伴う」と「痛みが伴う」のように同じ用言でも格の用法によって別物と定義されている場合がある．本稿では格の用法による意味の違いには着目せず，概念に対して同じ用言が複数現れた場合，それ

らの頻度を合計した値をその概念に対する用言の最終的な頻度とする．4 章で述べた 3 種類の概念ベースに対して上記の閾値調整を行い，全 9 通りの属性追加を行った．

5. 属性への重み付け

属性への重み付けには概念ベース *idf* [5]を用いる(式 2)．$idf_3(B)$とは重みを付与される属性 *B* の概念ベース *idf* であり，本稿では三次属性まで展開することで求める [6]．V_{all} は概念ベースに定義されている全概念数で，$df_3(B)$は三次属性集合内で概念 *B* を属性として持つ概念数である．

$$idf_3(B) = log(V_{all}/df_3(B)) \quad (2)$$

大規模格フレームによる頻度が高いほど概念に対して重要な属性であると考え，概念ベース *idf* に以下の 2 種類の補正値の内いずれかを掛けた値を追加属性の重みとする．基本的には補正値①を用い，追加属性が概念に対しすでに属性として存在する場合のみ補正値②を用いる．補正値①は大規模格フレームでの頻度により式 3 で算出する．

$$\text{【補正値①】} = 1 + \frac{\text{属性}B\text{の頻度}}{\text{概念}A\text{に対する追加属性の合計頻度}} \quad (3)$$

補正値②は元の概念ベースの属性重み上位 30 語 [7]における追加属性の重みの割合により式 4 で算出する．上位 30 語に追加属性がない場合は重み変更を行わない．

$$\text{【補正値②】} = 1 + \frac{\text{属性}B\text{の元の重み}}{\text{概念}A\text{の属性重み上位30語の重み合計値}} \quad (4)$$

6. 精度評価

6.1 *X-ABC* 評価

基準概念 *X* に対し，*X* と高関連の概念 *A*，中関連の概念 *B*，関連のない概念 *C* の 4 つの概念のセットを 500 セット用意する．本稿では属性追加が行われた名詞概念を *X* とした．*DoA*(*X*,*A*)，*DoA*(*X*,*B*)，*DoA*(*X*,*C*)を *X* と *A*，*B*，*C* の関連度，*AveDoA*(*X*,*C*)を評価セット全体における *DoA*(*X*,*C*)の平均とする．式(5)及び(6)を満たすものを正解とする．

$$DoA(X,A) - DoA(X,B) > AveDoA(X,C) \quad (5)$$

$$DoA(X,B) - DoA(X,C) > AveDoA(X,C) \quad (6)$$

500 セット中，正解のセットの比率を概念ベースの精度とする．また 500 セットの中から概念 *A* 又は *B* が動詞又はサ変名詞のセットを 203 セット抜き出し動詞評価セットとした．正解のセットの比率を名詞概念と動詞属性及びサ変名詞属性の関連性の評価とする．500 セットによる評価結果を表 3 に，動詞評価セットによる評価結果を表 4 に示す．

表 3　概念ベース評価結果（500 セット）

追加前	閾値	動詞のみ	サ変名詞のみ	両方追加
79.4%	①	79.6%	79.0%	80.0%
	②	79.6%	79.6%	80.0%
	③	79.0%	80.2%	79.2%

表 4　概念ベース評価結果（動詞評価セット: 203 セット）

追加前	閾値	動詞のみ	サ変名詞のみ	両方追加
78.8%	①	80.7%	78.3%	81.2%
	②	80.2%	80.2%	81.2%
	③	79.8%	80.7%	79.8%

6.2 目視評価

目視評価は 7 名で行い，概念に追加される属性が妥当であるかを判断する．4 名以上が正しいとした属性の比率を追加属性の精度とする．目視評価の結果を表 5 に示す．

表 5　追加属性の目視評価結果

閾値	動詞のみ	サ変名詞のみ	両方追加
①	49.4%	65.4%	52.7%
②	68.4%	78.4%	70.2%
③	80.8%	74.4%	79.9%

7. 考察

表 5 より追加属性は高い閾値で選別した語ほど人間が適切であると感じた割合が高くなった．大規模格フレームの頻度を用いることで，名詞概念に対して適切な属性を選別することが出来たと言える．また表 4 より動詞評価セットを用いた場合，作成した概念ベースはほぼ全てで精度が向上した．属性を追加することで動詞及びサ変名詞に対し，より適切な連想を行えるようになったと言える．同じく表 4 より動詞とサ変名詞の両方を追加した場合，閾値②から③にかけて精度が低下しているが，これは閾値③によって除去された属性の中に重要な属性が含まれていたためであると言える．閾値①と②では精度は同じだが，これは閾値②で除去された属性が重要でなかったため *X-ABC* 評価でほとんど使用されなかったためであると考えられる．つまり閾値②で両方を追加した場合の概念ベースが，高い精度でより多くの属性を追加できたと言える．なおこの概念ベースは名詞概念 1 語あたり平均約 8.1 語の属性が追加され，表 3 より概念ベース全体の精度も向上している．実際に閾値②により追加された動詞及びサ変名詞の例を表 6 に示す．

表 6　閾値②で両方追加した場合の追加属性例

概念	追加属性例	属性数の変化
英語	話す，勉強，喋る，理解，…	57 語→86 語
林檎	かじる，剥く，食べる，…	72 語→83 語

8. おわりに

本稿では大規模格フレームを用いて，概念ベースの名詞概念に対する動詞属性及びサ変接続属性の追加と重み付けを行った．結果として，概念ベースの精度を保ちつつ，名詞概念に対して平均約 8.1 語の属性を増やすことができた．

謝辞

本研究の一部は，科学研究費補助金(若手研究(B)24700215)の補助を受けて行った．

参考文献

[1] 笠原要, 松澤和光, 石川勉, “国語辞書を利用した日常語の類似性判別”, 情報処理学会論文誌, vol38-No7, pp.1272-1283, 1997.

[2] 井筒大志, 渡部広一, 河岡司, “概念ベースを用いた連想機能実現のための関連度計算方式”, 情報科学技術フォーラム FIT2002, pp.159-160, 2002.

[3] 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, 自然言語処理研究会 171-12, pp.67-73, 2006.

[4] 形態素解析器, http://chasen-legacy.sourceforge.jp/, 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室), 2012.

[5] 天野真家, 石崎俊, 宇津呂武仁, 成田真澄, 福本淳一, “IT Text 自然言語処理”, オーム社, 2007.

[6] 小島一秀, 渡部広一, 河岡司, “概念ベースにおける概念属性の確からしさによる概念属性の重み決定法”, 信学技報, AI2001-39, pp.39-46, 2001.

[7] 荒木孝允, 渡部広一, 河岡司, “共通・類似属性を考慮した概念間関連度計算方式”, 情報処理学会第 68 回全国大会講演論文集, 4N-2, 2006