F-036

構造化された概念ベースの特徴を利用した関連度計算方式
Calculating Degree of Association using Feature of Structured Concept-Base

岸本達也† 三品賢一† 土屋誠司‡ 渡部広一‡
Tatsuya Kishimoto Kennichi Mishina Seiji Tsuchiya Hirokazu Watabe

1. はじめに

人間はある語から関連性のある語を連想する能力があり，会話で役立てている．この能力をコンピュータに持たせることができれば，自然な会話ができるコンピュータの実現に近づくと考えられる．そこで我々は，人間の常識を判断するシステムとそれを支える語概念連想システムを構築している．語概念連想システムを実現するために，語の意味を理解するための概念ベース[1]や語と語の関連性の強さを計るための関連度計算方式[2]を用いている．

既存の概念ベースは属性の品詞情報や，同義や類義といった概念と属性の関係の情報は保持しておらず，概念と各属性との関係性を把握することができない．そこで属性の品詞ごとに，同義や類義などの関係をあらかじめ定義した構造化概念ベース[3]がある．しかし現在構造化概念ベースの特徴を利用した関連度計算方式が存在せず従来の関連度計算方式を使用している．問題点として概念ベースには名詞概念が最も多く登録されているため，形容詞や動詞の属性が使用されない可能性が考えられる．

以上より，本稿では構造化概念ベースの特徴を利用した関連度計算方式を提案する．

2. 概念ベース

概念ベースは電子化された国語辞書などから自動的に構築した知識ベースである．ある語を概念とし，概念の意味特徴を表す語（属性）とその重要さを表す数値（重み）の対の集合で定義される．また概念ベースでは全ての属性は概念としても登録されている． n 個の属性 a_i と重み $w_i(>0)$の対によって定義される概念 A を式(1)に示す．

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

既存の概念ベース（既存 CB）は，国語辞書や新聞記事，シソーラスなどを基にして構築する．既存 CB の概念数は約 9 万語，一概念当たりの平均属性数は 37 個である．

3. 関連度計算方式

関連度計算方式とは 2 つの概念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 までの実数値で表現され，関連が強いほど大きな値を示す．

既存の関連度計算方式では，概念 A と B の関連度($DoA(A,B)$)を求める際，概念 B の属性を概念 A の各属性との一致度の和が最大となるよう並び替える．そして以下によって関連度を求める．u_i と v_i は a_i と b_i の重みを表す．

$$DoA(A,B)=\Sigma_i DoM(a_i,b_i)\times(u_i+v_i)/2\times min(u_i,v_i)/max(u_i,v_i) \quad (2)$$

ここで，$DoM(a_i,b_i)$は属性 a_i と b_i の一致度である．一致度は概念 a_i と b_i それぞれの属性を比較し，一致した場合に小さいほうの重みを選択して足し合わせた値である．

4. 構造化概念ベース

4.1 構造化概念ベースの構造

概念として用いる語の品詞は単一で意味を持つことができる名詞，形容詞，動詞とする．構築する概念ベースの概念の定義を図 1 に示す．

N	$\{ n_{sa}, n_{si}, n_t, n_i, n_r, a_{sa}, a_{si}, a_i, a_r, v_{sa}, v_{si}, v_i, v_r \}$

- N ：名詞概念
- n ：名詞属性
- a ：形容詞属性
- v ：動詞属性
- sa：同義語
- si ：類義語
- t ：上位語
- i ：反意語
- r ：共起語

図 1 概念の定義

図 1 のように，属性の格納位置に同義や類義などの意味を付与し，その意味合いに沿った属性を追加する．例えば，概念 N の属性 n_{sa} には，概念 N の同義語である名詞属性を追加する．このように，概念の定義に従って属性を追加することで，属性の品詞や概念と属性の関係を把握することができる．図 1 では名詞概念の定義を示したが形容詞概念，動詞概念も同様の定義である．

属性の重みは既存 CB を用いた関連度と概念ベース idf の積で表す．概念 X の概念ベース idf は以下の式で求める．

$$idf(X)=\log(\text{全概念数} / \text{概念 } X \text{ を属性に持つ概念数}) \quad (3)$$

4.2 概念の登録

概念の登録には，基本語データベース− 語義別単語親密度− [4]を情報源として利用し，見出し語を概念とする．

4.3 属性の追加

概念構造情報[5]から属性候補となる語を取得し，図 1 の定義に沿って属性を追加する． 共起属性は，概念の語義文に茶筌[6]を用いて形態素解析を行い，属性候補を取得し追加する． また構造化概念ベースに存在する概念と一致する疑念を既存 CB 内から選び，それらの属性を共起属性として追加する．追加すべき属性が存在しない場合，空の意味を表す「×」を追加する．

4.4 構築結果

構築した概念ベースの概念数は 28815 個で，平均属性数は約 27 個である．

4.5 既存の関連度計算方式による問題点

既存 CB には名詞概念が最も多く登録されているため，名詞属性の重みが高い．そのため人が重要だと判断する同義や類義の属性でも名詞でなければ重みが小さくなる．既存手法では重み上位から属性を取得するため，名詞以外の属性が消滅する恐れがある．また，既存手法を用いる場合，適切な属性数は 30 個とされているが，構造化概

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

念ベースの平均属性数 27 個を上回るため，構造化概念ベースに対して関連度がうまく求められない可能性がある．

5. 提案手法

5.1 属性の品詞ごとに使用する割合を決める手法

各品詞から関連度計算方式に用いる属性数の割合をあらかじめ設定し関連度を計算する．関連度を計算する際の使用属性数を 6 個と仮定したときの例を表 1 に，また属性選別の例を図 2 に示す．

表 1　使用する属性数の品詞の内訳の例

品詞	名詞	形容詞	動詞
割合(%)	33	33	33
個数(個)	2	2	2

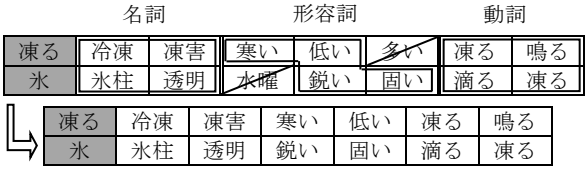

図 2　属性選別の例

あらかじめ各品詞の属性数の割合を決めることで属性の重みに依存せず名詞，形容詞，動詞の属性を取得できる．概念内に登録されたある品詞の属性数が足りない場合，設定した割合を満たすことができない．その場合，他の品詞の使用属性数を減らし設定した割合に近づける．上の例において形容詞属性の登録数が 1 個のみの場合，名詞と動詞を 1 個ずつ使用する．また形容詞属性が登録されていない場合，名詞と動詞の比を保ち使用属性数を計算する．上の例では，名詞と動詞は 1 : 1 の比率で使用するため個数は 2 個ずつとなる．この手法を提案手法 1 とする．

5.2 品詞ごとに一致度を計算する手法

既存手法では，属性の品詞に関わらず，一致度が最も高くなるように属性を並び替えていた．しかし本節の提案手法では，属性の並び替えを同じ品詞内で行う．そのため，品詞ごとに二つの概念間の属性数が揃う必要がある．二つの概念間の属性数が異なる場合，一致度が最大になるように並び替えた後，ペアができなかった属性を除去し関連度を計算する．この手法を提案手法 2 とする．

6. 精度評価

評価には *X-ABC* 評価を用いる．この評価では，任意の基準概念 *X* と，概念 *X* と関連が強い概念 *A*，関連がある概念 *B*，関連のない概念 *C* によって構成された 4 つの概念の組を 483 組用意する．例を表 2 に，その下に条件式を示す．

表 2　(*X-ABC*)セットの例

X	*A*	*B*	*C*
飲食店	食堂	客	得意

$$DoA(X,A)-DoA(X,B)>0 \quad (4)$$

$$DoA(X,B)-DoA(X,C)>0 \quad (5)$$

この評価を全ての組で行い，2 つの条件式を満たす割合を概念ベースの精度とする．既存手法，提案手法 1，提案手法 2 についてそれぞれ評価を行った．結果を表 3 に示す．

表 3　評価結果

手法	既存手法	提案手法 1	提案手法 2
精度(%)	56.5	60.7	56.1

表 3 より提案手法 1 の精度が最も高く，既存手法よりも 4.2%向上した．このときの品詞の割合は，名詞 82%，形容詞と動詞が 9%ずつであった．また使用属性数の上限は，可能な限り多く使用したほうが高精度であった．

7. 考察

既存手法と提案手法 1 で使用された属性を比較するため，以下の表 4 に既存手法と提案手法で「猛毒」と「毒」の関連度を算出する際の使用属性の一部を示す．

表 4　既存手法と提案手法における使用属性の比較

概念	属性			
	既存手法		提案手法 1	
猛毒	毒蛇	一酸化炭素	激しい	及ぼす
毒	害	人命	悪い	傷つける

表 4 より，既存手法で用いられた 4 つの名詞の属性の替わりに，提案手法 1 では形容詞や動詞の属性を使用できた．精度が向上したことから，重みのみに依存せず名詞，形容詞，動詞の属性を使用する本手法は有効であると考えられる．また改善点を考察する．本稿では関連度を求める際，概念の品詞は考慮せず使用属性の品詞の割合を決定した．しかし，例えば名詞概念中の動詞属性の重要度と，動詞概念中の動詞属性の重要度は異なる可能性がある．そこで品詞ごとに適切な割合を設定することで精度が向上すると考えられる．

続いて提案手法 2 について考察する．既存手法より精度が低い原因として使用属性数が挙げられる．そこで既存手法と提案手法 2 で使用された属性数の平均を調査するとそれぞれ約 20 個と約 19 個であった．精度と使用属性数の差がほぼ無いことから，この手法が有効でないかは判断できない．しかし構造化概念ベースにより多くの属性を登録することで既存手法の精度を上回る可能性はあると考えられる．

8. まとめ

本稿では構造化概念ベースの特徴を利用した関連度計算方式を提案した．その結果，既存手法よりも精度が向上し，形容詞や動詞属性を関連度計算方式に用いることが有効であることを示した．

謝辞

本研究の一部は，JSPS 科研費 16K00311 の助成を受けた．

参考文献

[1] 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41-64，2007.

[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，pp159－160，2002.

[3] 小川真路，芋野美紗子，土屋誠司，渡部広一，“概念の多義性を考慮した属性構造化による概念ベースの構築”，情報科学技術フォーラム FIT2013 pp. 223-224， 2013.

[4] NTT コミュニケーション科学基礎研究所，“基本語データベース– 語義別単語親密度– ”，株式会社学習研究社，2008.

[5] 小島一秀，渡部広一，河岡司，“常識判断のための概念ベース構成法—概念間論理関係を用いた概念属性の重み決定法”，信学技報，AI2010-58，pp.1-6，2011.

[6] ChaSen -- 形態素解析器，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)http://chasen-legacy.sourceforge.jp/，2013/1/10.