語の相関関係を考慮した概念ベースの連想語取得手法の検討

奥村 紀之　　豊嶋 章宏
香川高等専門学校 情報工学科
okumura@di.kagawa-nct.ac.jp, i09146@sr.kagawa-nct.ac.jp

1　はじめに

近年，スマートフォンなど IT デバイスの普及により，オンラインのテキスト資源に容易にアクセス可能となってきている．しかし一度に表示できる情報量の限界など，ストレスなくテキストデータを扱うには多くの課題が残されている．

本稿では，テキスト情報を選別するための基盤システムとして動作する連想システムに関する報告を行う．特に，連想システムの中核となる概念ベースは，未だ完全に自動構築することが困難である．本研究では，概念ベース構築の新たな素材として相関関係による連想語群収集手法について検討する．

2　研究目的

本研究では，語の連想機能をコンピュータ上で実現し，語と語の関連の強さを定量的に扱うための概念ベースを自動的に構築するための手法について検討している．従来，概念ベースは国語辞書や新聞記事などから半自動的に構築されている．一方で，そのサブシステムとして動作する常識判断システム等で使用する知識ベースを概念ベースに追加するなど，手作業による精錬作業が施されており，完全に自動で構築されていないため，再現性に乏しいという問題を抱えている．

そこで，概念連鎖による連想語群取得手法に加えて，国語辞書等での語と語の相関関係を利用して連想語群収集を行い，概念ベースを完全に自動構築することを目指す．本稿では特に，概念連鎖による連想語と相関関係によって取得した連想語の違いを比較検討している．

3　概念ベース

概念ベース [1] は，電子化辞書や電子化新聞から自動構築された大規模知識ベースである．概念 A は，その概念から常識的に連想可能な語や，その概念を意味的に特徴付ける語 (a_n) と，その語群を特徴付ける重み (w_n) の対の集合で定義されている (式 1).

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

本稿では，概念ベースの自動構築を行うため，その基盤となる初期概念ベースとして，EDR 電子化辞書 [2] の概念辞書から見出し語を概念とし総数約 22 万，見出し語に対する説明文を形態素解析することによって平均連想語数 3 を各概念に付与したものを構築している．なお，連想語群としては EDR 電子化辞書に見出し語として登録されているもののみを抽出している．なお，本稿では重み付けに関しては検討していない．

4　概念連鎖による連想語群拡張

概念ベースに対する連想語群は，[3] のように概念に対する連想語群を連鎖的に展開することで取得し，取捨選択されてきた．しかし，概念総数が有限である以上，たかだか n 次の連鎖で取得される連想語が収束すると推定される．そこで，構築した概念ベースに対し，93 のサンプル概念に対して概念連鎖の収束性を確認したところ，最大で 33 次の連鎖で取得される連想語群に変化がなくなることが確認された [4].

33 次の概念連鎖によって取得される連想語群は平均して 1000 あるが，93 のサンプルについてすべて目視によって評価したところ，基準となるサンプル概念に対して確からしいと考えられる連想語群は平均して 30 程度しか取得できていない．表 1 に「うどん」を 33 次展開した集合に含まれる適切な連想語群の例を示す．

表 1 に示した「うどん」の例では，33 次まで連鎖展開した場合およそ 1100 の語群が取得されているが，その中で正解であると考えられる連想語は 44 であり，その割合はきわめて小さい．従って，概念連鎖が収束するまで連鎖展開を行い連想語群を取得する場合，雑

表 1: うどんの概念連鎖集合

食用	練り	麺	水
食べ物	品物	代金	かける
つくる	健康	粉	安い

音となる語群を除去する手法が非常に重要な役割を果たす．連想語群の信頼性評価尺度として属性信頼度 [5] が提案されているが，適切な論理関係が抽出できなければ信頼度を設定することそのものが困難である．そこで，新しい重み付け手法の検討，および獲得した語群の選別方法が必要となる．

5　相関関係を用いた連想語群抽出

前節で述べたとおり，概念連鎖による連想語群取得手法では，たかだか 33 次の連鎖で収束し，それ以上の新規連想語群獲得は望めない．また，取得される語群もきわめて質が悪く，その選別に問題が残る．

そこで，EDR 電子化辞書をテキストマイニングツールを用いて解析することによって，概念連鎖ではない新たな連想語群獲得を検討する．本稿では IBM 社の Content Analytics に EDR 電子化辞書の概念辞書を解析させ，相関語群を取得している．表 2 に Content Analytics によって取得された「うどん」の相関語群の例を示す．

表 2: うどんの概念連鎖集合

料理	そば	かける	煮る
肉	月見	汁物	てんぷら
ゆでる	ネギ	粉	小麦

表 2 に示したように，表 1 の概念連鎖によって取得された語群と重複する語も取得されるが,「ゆでる，てんぷら，ネギ」といった「うどん」に関して適切と思われる連想語群が多く取得される傾向にあった．なお，Content Analytics では「うどん」に関しておよそ 100 の語が取得された．100 の語群の中には辞書特有の情報である読み仮名などが含まれており，概念ベースを構築するに当たっては，これらの情報は雑音となるため適切に除去してから解析させる必要がある．また，取得された語群は相関値によって選別されており，およそ半数の語が「うどん」に対して適切であると考えられ，他のサンプルについても同様の傾向であった．

6　考察

表 1，表 2 から，概念ベースを構築する上で重要となる語群を適切に取得するには，概念連鎖を利用する場合，その選別手法が重要な課題であり，相関値を利用する場合は，テキストの事前処理が重要であることがわかった．

また，相関値は連想語群に対する重み付けに有効活用できると考えられるので，従来の $tf \quad idf$ 法などと組み合わせることで，適切な重みを付与できるのではないかと期待している．

7　おわりに

本稿では，従来の概念ベース構築法に加えて，テキストマイニングツールによる相関分析によって連想語群を拡張する手法について検討した．今後は，サンプル数を増やし，概念ベースの構築を完全に自動化するための手法を開発していく必要がある．

謝辞

本研究の一部は科研費 (23720222) の助成を受けたものである．

参考文献

[1] 「概念間の関連度計算のための大規模概念ベースの構築」奥村 紀之, 土屋 誠司, 渡部 広一, 河岡 司 : 自然言語処理 Volume14 Number5 p.41-64,2007.

[2] 「EDR 電子化辞書」独立行政法人情報通信研究機構, 2007.

[3] 「連想システムのための概念ベース構成法－語間の論理的関係を用いた属性拡張」小島 一秀, 渡部 広一, 河岡 司 : 自然言語処理，Volume11，Number3，pp.21-38,2004.

[4] 「概念ベースにおける属性連鎖の傾向と属性集合の評価」豊嶋 章宏, 奥村 紀之, 情報処理学会第 75 回全国大会,1Q-3,2013

[5] 「連想システムのための概念ベース構成法 -属性信頼度の考え方に基づく属性重みの決定-」小島 一秀, 渡部 広一, 河岡 司 : 自然言語処理，Volume9，Number5，pp.93-110,2002.