E-011

表記ゆれの補正と雑音除去による新聞から構築した概念ベースの精錬

Refinement of Concept-Base constructed from newspaper articles by correction of notation variability and noise reduction

山口周平† 岸本達也† 土屋誠司‡ 渡部広一‡
Shuhei Yamaguchi Tatsuya Kishimoto Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，人間と円滑なコミュニケーションをとる知的なロボットへの期待が高まっている．人間にはある単語から別の単語を連想する能力があり，人間同士のコミュニケーションは連想によって円滑に行われていると考えられる．そこで我々はコンピュータに連想機能を持たせることを目指している．そのためには語と語の関連性を知識として保持する知識ベースが必要であり，そのような知識ベースとして概念ベース[1]が存在する．既存の概念ベースは電子化された国語辞書などを情報源として構築されており，人間が日常で使用する基本的な語が網羅されている反面，時事用語・流行語などが欠けている．この問題点を解決する手法として新聞などの共起する情報群から概念ベースを自動構築する手法[2]が存在する．しかし，この手法で生成した概念ベースには表記ゆれが含まれている，雑音が存在するなどの問題点が存在する．以上より，本稿では表記ゆれの補正及び雑音の除去により共起する情報群である新聞記事を情報源とした概念ベースの精錬を行う．

2. 概念ベース

概念ベースは単語を概念とし，属性と重みの対の集合として定義している．属性は概念を特徴付ける単語であり，重みは概念に対する属性の重要度を表す．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．

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

3. 関連度計算方式

関連度計算方式[3]とは，2 つの概念間の関連の強さを定量的に表現する手である．関連度は 0.0 から 1.0 の実数値で，概念間の関連が強いほど大きな値となる．関連度は一致度を用いて算出される．一致度は 2 つの概念間で共通した属性がどれくらいあるかを示し，2 つの概念間の属性で表記が一致する属性の小さい方の重みの和をとる．小さい方の重みを計算に用いる理由は，2 つの概念で共通している重みの分が有効であると考えるためである．概念 A、B の一致度 は以下の式で表現される．

$$DoM(A,B) = \sum_{a_i=b_j} min(u_i, v_j) \quad (2)$$

概念 A、B の持つ 1 次属性をa_i，b_j，重みをu_i，v_jとする．このとき概念 A、B 及び関連度$DoA(A,B)$は次式で表される．

$$DoA(A,B) = \sum_i DoM(a_i, b_{x_i}) \times (u_i + v_j)/2 \times min(u_i, v_{x_i}) / max(u_i, v_{x_i}) \quad (3)$$

† 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

4. 共起する情報群からの概念ベース自動構築

国語辞書以外の情報源から自動的に概念ベースを構築するため，語群が共起して出現する情報源から概念と属性を自動的に獲得する手法[2]が存在する．本研究では電子化された 95 年度の毎日新聞の記事 1 年分を情報源とする．

4.1 概念と属性の獲得

記事中の句点で区切られた 1 文を共起範囲として茶筌[4]により形態素解析を行う．必要に応じて複合語処理（複数の形態素を連結する処理）を行い，概念及び属性を獲得する．概念とする語は単一で意味を持つ名詞・動詞・形容詞とする．図 1 の例では下線部が共起範囲である．

> …が十九日、明らかになった。検査は、国立大学病院で巨額の試薬代金を卸業者に滞納していたことが発覚したのを受け、全国規模で行われた。複数の関係者によると…

図 1　新聞記事における共起範囲の例

図 1 で示した共起範囲からは「検査，国立大学病院，巨額，…」といった語が獲得される．ここで獲得した語をそれぞれ概念とし，共起して出現した語を属性として定義する．先に述べた例では「検査」という概念に対して「国立大学病院」などの語を属性とする．獲得済みの概念が出現した場合，共起して出現した他の語を属性とする．

4.2 属性への重みづけ

獲得した概念の各属性に対して重みを付与する．重み付けは *tf-idf*[5]の考え方を概念ベースに適用した概念ベース *idf* を用いる．*tf-idf* は情報検索などで幅広く利用される特徴量の指標であり，*tf* はある語の出現頻度，*idf* は全文書において，ある語が出現した頻度の逆数の対数であり，これらの積を重みとする．概念ベースでは概念を文書，属性を文書中の語として重み付けする．重み$w(C, a_i)$，*idf* の値$idf(a_i)$は以下の式で定義される．

$$w(C, a_i) = tf(C, a_i) \times idf(a_i) \quad (4)$$

$$idf(a_i) = log_2(N/df(a_i)) + 1 \quad (5)$$

式中の C は概念，a_iは属性を表している．また$tf(C, a_i)$は概念 C に対する属性a_iの出現回数，N は全概念数，$df(a_i)$はa_iを属性として持つ概念の数を表している．

4.3 精度評価

構築した概念ベースの評価には *X-BC* 評価を用いる．基準概念を X と置き，概念 X とある程度関連がある概念 B，関連のない概念 C によって構成された 3 つの概念の組を人手で 340 セット作成した．評価セットの一部を表 1 に示す．ここで，概念 X と概念 B の関連度を *DoA*(*X*,*B*)，概念 X と概念 C の関連度を *DoA*(*X*,*C*)，とする．それぞれの関連度の値が以下の式を満たした場合を正解とする．

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

この評価を全ての組で行い，正解となった組の割合を概念ベースの精度とする．

表 1　評価セットの一部

X	B	C
馬券	単勝	山椒
…	…	…

構築した概念ベースの精度は 70.9%であった．また，概念数は 294187 個，平均属性数は 113 個となった．

4.4 既存手法の問題点

新聞記事から構築した概念ベースは概念の属性として「する」や「ある」のように多くの概念の属性となる概念が存在した．その結果，関連のない概念同士でも関連度が大きくなると考えられる．上述した属性は概念を表すのに不適と考え，除去することで関係のない概念同士の関連度が下がり，*X-BC* 評価の精度が向上すると考えた．

また，構築した概念ベースの概念を見ると，表記ゆれが含まれていた．表記ゆれとは同音・同意味の語に対して異なる文字表記が存在することである．具体的な例としては「引越」と「引っ越し」，「りんご」，「リンゴ」，「林檎」そして「バイオリン」と「ヴァイオリン」などが挙げられる．これらは同じ概念を表すべきであるが，別々の概念として登録されており，それぞれ属性やその数が異なるため関連度の計算に影響を与えていると考えられる．概念「りんご」，「リンゴ」，「林檎」を例としてそれぞれの属性数を表 2 に示す．

表 2　「リンゴ」・「りんご」・「林檎」の各属性数

概念	属性数[個]
リンゴ	1100
りんご	158
林檎	19

5. 提案手法

5.1 表記ゆれの補正

表記ゆれの補正手法として表記ゆれ辞書を構築し，参照する手法を用いた．表記ゆれ辞書は概念ベース構築の際に用いた新聞記事を情報源とし，茶筌の解析結果のうちの形態素の読みを用いて構築した表記ゆれ辞書①と英字表記とカタカナ表記からなる辞書を元に英字表記を利用して表記ゆれ辞書②を構築した．各辞書の一部を表 3 と 4 に示す．

表 3　読みと複数表記のリストの一部

読み	表記
リンゴ	リンゴ
	りんご
	林檎

表 4　英字表記と複数表記のリストの一部

英字表記	表記
violin	バイオリン
	ヴァイオリン

5.2 雑音除去

4.2 節において算出した *idf* の値に閾値を設け，雑音となる属性を削除する．概念数が約 30 万個のとき *idf* の値は定義式より最小は 1.0，最大は約 19 となる．*idf* の値が小さい場合，その概念は多くの概念の属性として頻繁に出現しており，概念の特徴を表す属性ではない．一方で，*idf* の値が大きい場合，属性としてほとんど出現しておらず，日常で使用する概念ではないと考えられる．そのため，これらの属性を雑音とする．閾値は 1.0 から 19 まで 1.0 ずつ設定し，概念ベース *idf* が上限以上，下限以下の属性を除去した．

6. 評価

表記ゆれを補正した概念ベースに対して *X-BC* 評価を行った結果 71.8%の精度だった．また，雑音除去を行った概念ベースでは，上限の閾値が 16 から 19 のときに精度が最も高くなり，70.9%となった．また，下限の閾値では 8.0 のときに精度が最も高くなり，71.8%となった．また，両手法を適応した概念ベースでの精度は上限の閾値 16,下限の閾値 8.0 のとき最大で 69.8%であった．

7. 考察

表記ゆれの補正を行った後概念ベースの精度は向上した．しかし，表記ゆれ補正を行った概念が 4361 個と概念ベース全体の 1.5%と少ないためより語数の多い表記ゆれ辞書を構築し補正を行うことで精度が向上すると考えられる．また，「ミュージック」や「音楽」などの外来語の表記ゆれを和英辞書などを用いて補正する必要があると考えられる．

雑音除去では，上限閾値では精度に変化はなく，下限閾値では 0.9%精度が向上した．精度に大きな変化がなかったことについて，概念ベース *idf* の分布を調べると 8.0 以下の概念は全体の 1.5%程度と非常に少なかったため効果がなかったと考えられる．一方，概念ベース *idf* が 16 以上の概念は全体の 13.4%程度であったため表記ゆれの補正によって少し精度が上がったと考えられる．

さらに，新聞記事から獲得できる特徴的な概念である固有名詞についても調査を行うと，固有名詞が全体の 12%ほど存在し，他の名詞概念より概念ベース *idf* が 1.0 程度小さくなる傾向が見られた．これより，固有名詞に対しては閾値を別で設ける必要があると考えられる．さらに概念ベース内の品詞の内訳を見ると名詞が 97%，形容詞が 0.38%，2.62%と偏りが大きいことが分かった．したがって各品詞で異なる閾値を用いて雑音除去を行う必要がある．

また，両手法を併用したときの精度低下については補正と除去によって概念同士の属性数に差が開いてしまったことが原因だと考えられる．

8. まとめ

本稿では共起する情報群を元に構築された概念ベースに対して表記ゆれの補正と雑音の除去を行った．その結果精度は向上し，手法は有効であることを示した．今後，辞書の語数増加や構築した概念ベースに含まれる概念の品詞の考慮することによってさらに精度が向上すると考えられる．

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] 芋野美紗子，吉村枝里子，土屋誠司，渡部広一，“共起する情報群からの概念ベース自動生成手法”，信学技報，HCS2014-114，pp25-30，2015.

[3] 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.

[4] 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.

[5] 徳永健伸（編），“情報検索と言語処理”，東京大学出版会，1999.