E-002

否定表現から肯定表現への変換による文間関連度計算方式
Measuring the Degree of Association between Sentences with Change Expression of Negation to Expression of Affirmation.

谷 裕一朗† 芋野 美紗子‡ 土屋 誠司‡ 渡部 広一‡
Yuichiro Tani Misako Imono Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年インターネットの普及により，個人が入手できる情報量は膨大なものとなり，必要とする情報の収集が困難になってきている．それに伴い，利用者の要求にあった適切な情報の検索が必要とされている．特に単語表記の意味だけではなく，文としての意味を据えた検索を行うことによって正確な情報検索を行うことができる．そこで文の正確な類似性や関連性を求める定量化技術が必要であると考える．

本稿では文中の否定表現を肯定表現へ変換する処理を，文間の関連性を定量化する手法である文間関連度計算方式[1]に適用する．否定表現を考慮することにより，より正確な関連性を導き出すことを目的とする．具体的には，反対語辞書を用いて単語にかかる否定表現を肯定表現に変換する変換アルゴリズムを構築する．

2. 関連技術

2.1 概念ベース

概念ベース[2]は，国語辞書等から自動構築された知識ベースである．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．

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

また，概念は約 9 万語格納されており，概念に定義されている意味特徴を一次属性，一次属性の意味特徴を概念に対する二次属性，というように概念は任意の次元で表現できる．

2.2 関連度計算方式

関連度計算方式[2]とは，概念ベースに定義されている 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 の間で値が変動し，概念間の関連が強いほど大きな値を示す．尚，関連度計算方式にはお互いの概念が持つ属性の一致度と重みを利用している．

2.3 一致度

任意の概念 A, B について，それぞれ一次属性を a_i, b_j とし，対応する重みをそれぞれ u_i, v_j とする．それぞれが持つ属性数が M 個，N 個とすると，概念 A, B はそれぞれ以下で表される．

$$A = \{(a_1,w_1),(a_2,w_2), \cdots (a_M,w_M)\} \quad (2)$$

$$B = \{(b_1,v_1),(b_2,v_2), \cdots (b_N,v_N)\} \quad (3)$$

このとき概念 A, B の一致度を以下の式で定義される．

$$DoM\ (A, B) = \sum_{a_i = b_j} \min(u_i, v_j) \quad (4)$$

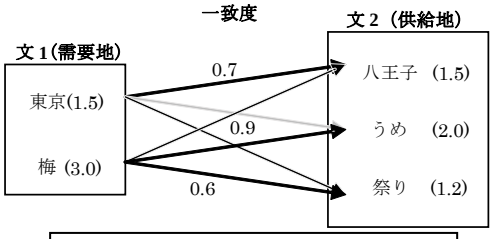
$a_i = b_j$ は属性同士が表記的に一致した場合を示している．つまり，一致度とは概念 A と概念 B 双方が共通して持つ属性のうち，小さいほうの重みを足し合わせたものとなる．両概念の属性と重みが完全に一致する場合には一致度は 1.0 となる．

† 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 Engineering, Doshisha University

2.3 EMD を用いた文間関連度計算方式

文間関連度計算方式とは，文と文との関連の強さを定量的に表すことができる計算方式である．本稿ではEMD[3]を用いて文間関連度を算出する．EMDとは線形計画問題の一つであるヒッチコック型輸送問題において計算される距離尺度である．二つの文に出現する自立語群を離散集合と見なすことで文間関連度を算出できる[1]．

例えば，文1「東京の梅」と文2「八王子のうめの祭り」という2文の関連度を求める場合，2文の自立語をそれぞれ需要地，供給地と見立てることができる．そして需要地と供給地の距離は自立語間の関連性と見立てることができ，概念ベースを用いた一致度計算により求めることができる．需要地の自立語である「東京」の場合，供給地の自立語と最も関連が高い「千代田区」の値が用いられる．一致度は関連度が高いと値も大きくなるため，1から一致度の値を引いた値に変換する．尚，各自立語には*tf・idf*[4]重み付けを行い，重みを輸送量として見立てる．以上の値を用いてすべての単語を考慮して関連性の定量化を行う．図1に計算例を示す．各自立語の数値は重みを示している．



輸送コスト(距離×輸送量)
東京⇒八王子: (1-0.7)×1.5 = 0.45
梅⇒うめ : (1-0.9)×2.0 = 0.2
梅⇒祭り : (1-0.6)×1.0 = 0.4
EMD＝(0.45+0.2+0.4)÷4.5 =0.233
結果 文間関連度= 0.233

図1 計算例

3. 文間関連度計算方式の問題点

文間関連度計算方式では，文に出現する自立語に重みづけを行うことで関連度を求める．そのため，否定が含まれた文の場合，否定表現である「ない」等は自立語でないため無視されてしまい，考慮されない．例えば，「私は試合で負けない」と「私は試合で負ける」の 2 文の場合「ない」を考慮しないので，自立語がすべて同じになってしまい，同じ意味と判断される．従って，関連度は最大の 1.0 が算出される．しかし，本来 2 文は違う意味である．

このように，否定表現が含まれている場合，文同士の正確な関連度を求めることができない．従って，否定表現を考慮するアルゴリズムが必要である．

4. 否定表現を考慮するアルゴリズム

本稿では文中の否定辞「ない」に着目し，肯定表現に変換するアルゴリズムの構築を行った．例えば，「私は試合で負けな

い」という文であれば，「私 試合 勝つ」のように変換する．否定表現が含まれた文を考慮した形に変換した後，文間関連度計算方式に適用する．

4.1 否定辞「ない」が含まれた文節の抽出

否定辞「ない」が含まれた文に対して，係り受け関係を解析するソフトである CABOCHA[5]を用いて文節に区切る．ここで文節とは，言語として不自然でない程度に区切ったときに得られる最小の単位である．「私は試合で負けない」という文の場合は，「私は」「試合で」「負けない」の 3 つの文節から成り立つ．

上記の例において，否定辞「ない」が含まれている「負けない」という文節を抽出する．

4.2 肯定表現への変換

否定辞「ない」が含まれた文節に対して，肯定表現に変換するために反対語を用いる．そこで，述語が対象となる反対語辞書の構築を行った．反対語大辞典[6]を参考に，改定常用漢字表[7]の例文に記載されている 1079 語の述語に対応する反対語辞書を構築し，これを基に肯定表現へ変換を行う．尚，反対語辞書は 1 対 1 対応の形で作成を行った．反対語辞書の例を図 2 に示す．

外れる ⇔ 付ける
寂れる ⇔ 賑わう
簡素だ ⇔ 複雑だ

図 2 反対語辞書(一例)

変換対象となる述語に対応する反対語が反対語辞書にあった場合，対応する語に変換を行う．例えば「負けない」という文節の場合，「負ける」の反対語「勝つ」が反対語辞書に存在するので，「勝つ」に変換する．

変換対象となる述語に対する語が反対語辞書に記載されていない場合は関係語辞書や概念ベースを用いて変換する語を派生させることで反対語を取得する．ここで，関係語辞書とは，同義語，類義語，上位語の関係にある語を登録したデータベースである．

提案手法を関連度計算方式に適用した場合，「私は試合で負けない」と「私は試合で負ける」の 2 文であれば，前文は「私 試合 勝つ」と変換される．変換された 2 文の関連度は 0.502 が算出される．以上のように，否定表現を考慮した文間関連度を求めることができる．

5. 精度評価

アンケートを用いて，提案手法の評価精度を行った．反対語辞書に記載されている語をランダムに 100 語選択し，選択した述語を含めた肯定文をそれぞれ 100 文作成した．これら 100 文を肯定文 S とする．

被験者には，1 つの肯定文 S に対して同義となる同義文 T と「ない」等の否定表現を用いない反対の意味をもつ反対語 Y を 1 つずつ作成してもらった．被験者 1 名につき 100 文についてそれぞれ考えてもらい，4 名から 1 文につきそれぞれ 4 パターンを得た．

肯定文 S に「ない」を付け加えた文を作成し，これを否定文 A とする．これら 4 つの文のセットを 400 セット用意する．表 1 に評価セット例を示す．

表 1 評価セット

	S	T	Y	A
1	私は試合で負ける	私は試合で敗れる	私は試合で勝つ	私は試合で負けない
2	:	:	:	:

否定文 *A* と同義文 *T* との関連度を *DoA(A,T)*，否定文 *A* と反対文 *Y* の関連度を *DoA(A,Y)*とした．本来の文としての意味を考慮した場合，「ない」等の否定表現が文に含まれていれば文の意味が変わり，関連度も変化する．しかし，既存の文間関連度計算方式では表記一致してしまい，否定文 *A* と肯定文 *S* の場合必ず関連度は 1.0 となってしまう．そこで表記一致を避けるために肯定文と意味が似ている同義文を用いることで精度評価を行った．以上より，次の条件を満たすものを正解とした．

$$DoA(A,Y) > DoA(A,T) \quad (5)$$

次に否定文 *A* に対して，提案手法である否定表現「ない」を肯定表現に変換したもので関連度を求めた．尚，同様に式(5)とを正解とした．

5.2 評価結果

評価結果を表 2 に示す．

表 2 評価結果

	既存手法	提案手法
精度(%)	50.5	55.1

評価結果より，既存手法より提案手法の方が 4. 6%精度が上がったことが分かる．

6. 考察

評価結果より，既存手法に比べると提案手法の方が精度が高いため，否定表現が考慮出来ていると考えられる．したがって，否定辞「ない」を取り除き，肯定表現に変換した文として計算を行うことで適切な関連度が求められる．

今回精度が低かった原因の1つとして，変換するべき述語に問題があると考えられる．実際，ある述語に対する反対語は 1 つとは限らず複数存在する．したがって,1 対多の反対語辞書が必要と考えられ，時と場合によって相応しい反対語を条件で適用出来れば精度は向上すると考えられる．以下反対語辞書における1対多の例を図 3 に示す．

押す ⇔引く，戻す，抜く，離す
進める ⇔止める，戻る，休む，維持する

図 3 反対語の 1 対多の例

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] 藤江悠五，渡部広一，河岡司，“概念ベースと Earth Mover's Distance を用いた文書検索”，信学技報，Vol.108, No.456, pp.111-116, 2009.
[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] Y.Rubner, C.Tomasi, L.Guibas：The earth mover’s distance as a metric for image retrieval, Int.J.Comput.Vision, Vol.40, pp.99-121, 2000.
[4] G.Salton and C.Buckley. Term-weighting approaches in automatic textretrieval. Information Processing and Management, Vol.41, No.4, pp.513-523, 1998.
[5] 工藤拓，松本裕治，”チャンキングの段階適用による係り受け解析”，情報処理学会論文詩，Vol.43, No.6, 2002.
[6] 中村一男，“反対語大辞典”，東京堂出版，1965.
[7] 文化審議会，“改定常用漢字表”，文化庁，2010.