4ZB-2

あらゆる概念表記への対応・精度向上を目指した意味的類似度算出ツール

野口　洋平†　清水　諒†　杉本　邦弘†　石川　勉†
拓殖大学工学部情報工学科†

1　まえがき

我々は, 単語 (概念) 間の意味的な類似性判別を主な目的とした概念ベース (以下 GB) の構築を進めてきた [1]. これは国語辞書の語義文を基に 1 概念 1 表記で構築しており, その単純な概念間での類似性判別能力に関しては, シソーラスを上回ることを確認している [2]. しかし, この構築は機械的に行っているため, 概念によっては人間の感覚と大きくずれる類似度を与えることがある. また, 表記揺れや複合語等にも対応していない. ここでは, これらに対処した類似度算出ツールについて報告する.

2　構築の考え方と構成概要

2.1　あらゆる概念表記への対応

GB は本体 (約 26 万語) と清書ファイルで構成されている. 本体は各概念ごとにその概念に関連する属性とその属性値の対を記録したものである. また, 清書ファイルは各概念の概念番号と標準的な表記を記録したものである. GB を使用する場合, まず清書ファイルで概念の表記から概念番号を調べ, それを基に本体からその属性及び属性値を得る. 現在, この清書ファイルは前述したように 1 概念 1 表記となっているため, ここでは, 表記揺れを含むようこれを拡充する (複数の表記が 1 概念番号に対応). さらに, いかなる概念表記に対しても類似度が計算できるように, 複合語や造語にも対応可能とする.

2.2　類似度の精度向上

本 GB では, 概念の属性を日本語語彙体系 [3] の 2715 のカテゴリとしており, 類似度 (0〜1 の値) は, 各概念をこれを次元とするベクトル (概念ベクトル) とみなし, その内積で算出している. 概念の属性は, 基本的には語義文中の自立語が属するカテゴリであることから, 語義文が短かったり, 共通する自立語によりそれが構成されている場合には, 非類似の概念間でも高い類似度を算出するという問題点がある. 例えば,「視野」と「磁場」では, あまり類似しないにも関わらず高い類似度 (0.87) となる. 従って, これを解決するため, ここでは人手で作成された日本語語彙体系のシソーラスをもとに算出した類似度も考慮し, その精度向上を図る. 具体的には, 両者の類似度を統合して, 最終的な類似度とする. さらに, 同義語間の類似度も必ずしも 1 に近い値とならない場合もあるので, 同義語辞書を作成し, これを利用することとする.

以上を考慮して改良した類似度導出ツールの全体構成を図 1 に示す.

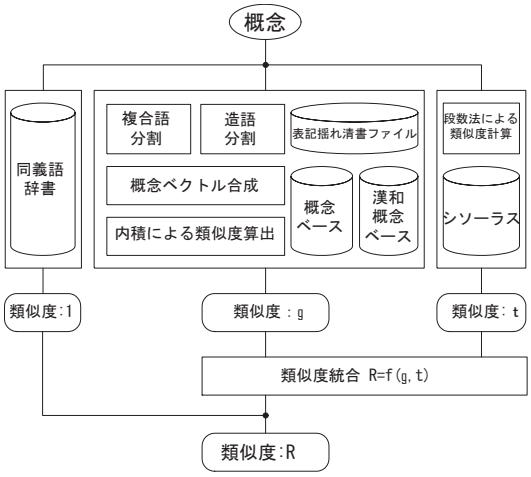

図 1: 類似度算出ツールの全体構成

3　対応可能語数の増加

3.1　表記揺れ

表記揺れには様々な形が存在するが, 大きく 5 種類に分け清書ファイルに登録した.

(1) 片仮名表記揺れ

片仮名の表記揺れは, 長音や濁点の有無, 部分的な表記の違いが見られる (例:「メモリ」→「メモリー」). これについては文献 [4] で示されている片仮名表記揺れパターンに属する全ての表記揺れを登録した. その他, 片仮名表記される傾向が見られる果物, 魚, 動物などの特定カテゴリに属する単語についても登録した.

(2) 漢字表記揺れ

日常使用しない難しい漢字は平仮名で表すことが多い (例:「宝籤」→「宝くじ」). これについては, 常用漢字 1945 以外の漢字を平仮名に変更し, 表記揺れとした.

(3) 略語表記揺れ

片仮名表記の単語について, 表記を短縮した略語がある (例:「アマチュア」→「アマ」). これは, 語義文がその見出し語の一部のみからなるとき, それを表記揺れとした.

(4) 送り仮名表記揺れ

単語に関する送り仮名の有無の違いにより多くの表記揺れが生まれる (例:「引き受ける」→「引受ける」,「ひき受ける」). これは, 国語辞書の見出し語中の括弧書きされた平仮名について, 括弧内の平仮名の有無全ての組み合わせを表記揺れとした.

(5) 平仮名・英字表記揺れ

全ての概念の平仮名読みは表記揺れとした. また,「メモリー」等の片仮名は「memory」と英字で表記されることがあるため, これも表記揺れとした.

以上の表記揺れ対応により, 清書ファイルは 26 万概念, 約 51 万語に対応可能となった.

An Improved Tool for Measuring Semantic Similarity between Words.

†Yohei Noguchi,Ryo Shimizu,Kunihiro Sugimoto, Tsutomu Ishikawa
†Department of computer Science,Takushoku University

3.2 複合語

入力された概念が清書ファイルに存在しない場合, 複合語とみなし, 清書ファイルに登録された概念に分割する. この処理は, 以下の例のように, 入力語の前方からみて, 清書ファイル中に存在する最長単語を取り出し, 次に, これを取り除いた語に対して, この処理を繰り返すことにより行われる.

例:1. 瞬間最大風速
2. 瞬間/最大風速
3. 瞬間/最大/風速

このように分割された単語の概念ベクトルを合成し複合語全体の概念ベクトルとする. この合成の方法としては各種考えられるが, ここでは全ての重みを均一としたベクトル合成とした.

3.3 造語

入力語が清書ファイルに存在せず, 複合語でもなく, かつ漢字のみからなる場合, 造語とみなす (例:「激安」). この場合, 漢和辞書を基に漢字 1 文字ごとのについて構築した GB(漢和 GB) を用いて漢字の概念ベクトルを得て, そのベクトルを合成し, その語の概念ベクトルとする [5].

4 類似性判別能力の向上

4.1 GB とシソーラスの類似度の統合法

GB とシソーラスの類似度を統合する場合, 以下について考慮する必要がある.

① 平均的な類似性判別能力は GB が優れている.
② 全く類似しない概念は, シソーラス上離れたカテゴリに属すると期待できる.
③ 日本語語彙体系のシソーラスでは, 同一カテゴリに属していても, あまり類似しない概念が存在する.

①は文献 [2] から言える. ②については, シソーラスは人手で作られているため, その仮定はほぼ成り立つと言えよう. すなわち, 全く類似しない概念間の類似度は低くなると期待できる. また,③については, 同シソーラスでは例えば「設備」と「摩天楼」は同カテゴリに属している. 従って, シソーラスを利用したこれらの類似度は距離法でも段数法でも 1 となってしまう. すなわち, シソーラスの類似度が高い場合, その値は必ずしも信頼できるとは言えないことになる. これらを考慮して,GB とシソーラスの類似度の統合式を以下のように設定する.

$$R = g \qquad (|g-t| \leq \alpha) \quad (1)$$

$$R = g\left(\frac{1-z}{1-\alpha}\right) + t\left(1-\frac{1-z}{1-\alpha}\right) \qquad (g-t > \alpha) \quad (2)$$

$$R = g\left(\frac{1-z}{1-\alpha}\right) + \frac{g+t}{2}\left(1-\frac{1-z}{1-\alpha}\right) \qquad (t-g > \alpha) \quad (3)$$

ここでは,R は統合後の類似度, g はGB の類似度,t はシソーラスの類似度 (段数法),α は閾値,$z = |g-t|$ としている.

(1) 式は,GB とシソーラスの類似度の差が閾値以下の場合には,①から GB の類似度をそのまま採用することを意味する. (2) 式は,GB の類似度が高くシソーラスの類似度が低くその差が閾値以上の場合, その差を考慮し線形結合したものである. これは, 両者の類似度の差が閾値に近い場合は,①から GB での類似度を優先し, 逆にその差が極めて大きい (1 に近い) 場合,②からシソーラスの類似度を優先するものである. また,(3) 式は,(2) 式と逆の場合に,③を重視し結合したものである. すなわち, シソーラスの類似度が高く GB の類似度が低い場合には, シソーラスの類似度を重視しすぎないようにしたものである.

また, この統合式では, 最適な閾値の値を設定する必要がある. これは平均的な類似性判別能力が高く人間の感覚とずれた類似度が得られづらいように設定することが望ましい. しかし, ここでは詳細は述べないが, 類似性判別能力は機械的評価法 [2] では $\alpha = 1$ のとき最大となり 0.6 以下となるとかなり低下する傾向があった. また, 類語を使った類似度評価では,α が小さいほど人間の感覚とあった高い類似度が算出される傾向があった. すなわち, これらは相反する傾向があるため, ここではそのバランスを考慮し 0.6 と設定した.

4.2 同義語の利用

広辞苑・大辞林・学研国語大辞典の 3 つの辞書を利用し, 同義語を機械的に抽出した [5]. この結果, 合計 11898 語の同義語候補を抽出し, 同義語辞書として構築した. さらに, 正確な同義語辞書を作成するために, シソーラスの同一カテゴリに属していない単語同士 (多義の場合は全てのカテゴリが一致しない場合) は同義語とみなさず除去する方法をとった. その結果,3684 語の同義語を獲得した.

5 評価

対応語数について, 新聞記事 1 年分のデータから何語の名詞を抽出できるかにより評価した. この結果, 全 59666 種類の語から, 従来の GB では 45076 種類 (約 76 %) の語しか抽出できなかったのに対し,58633 種類 (約 98 %) の語を抽出できた. また, 本ツールの速度は, 普通名詞 3000 組 (新聞データから抽出) を用いて測定した結果,1 回の類似度計算あたり約 1.1ms であった (CPU クロック 2.8GHz, メモリ 8GB). なお, 本ツールの動作に必要な記憶容量は約 1.4GB である.

6 まとめ

表記揺れ, 複合語, 造語に対応でき, かつ類似性判別能力を向上した類似度算出ツールについて述べた. 本ツールは,web 上で公開を行っているので試用されたい (http://sund.cs.takushoku-u.ac.jp/ruiji.html).

参考文献

[1] Nguyen Viet Ha, 帆苅譲, 石川勉, 笠原要:単語の意味の類似性判別のための大規模概念ベース, 情報処理学会論文誌.Vol.43,No.10,pp.3127-3136(2002)

[2] 川島貴広, 石川勉:言葉の意味の類似性判別に関するシソーラスと概念ベースの性能評価, 人工知能学会論文誌 20 巻 5 号 B(2005 年)

[3] 池原悟, 他:“日本語語彙体系 1 意味体系”, 岩波書店

[4] 久保村千明:「片仮名異表記処理能力を備え持つ情報検索システム」電子情報通信学会論文誌 Vol,j86-D-II No.3 pp.418-428 2003 年 3 月

[5] 帆苅譲, 石川勉, 笠原要:言葉の意味に関する階層型大規模概念ベースの構築, 電子情報通信学会信学技報 AI98-65(1999-01)P.25