5RE-02

シソーラスを用いた単語間類似度の判定法に関する研究

小林 雅暢†, 杜 小勇‡, 石井 直宏†
†名古屋工業大学,‡中国人民大学

1 はじめに

テキストデータに対する情報検索や自動分類ではテキスト内の単語に着目し処理を行なう。このようなシステムでは優れた単語間類似度の判定法が随所で必要となる。

本稿では既存の単語間類似度の判定法を紹介し、その問題点について述べる。そして、これらの問題を解決する方法として正規化を導入した類似度判定法を提案する。

2 単語間類似度の評価方法

一般的に単語には様々な意味があり、それを概念と呼ぶ。単語を比較し類似度を判定するには、その概念を比較する必要がある。よってこの判定には以下の2つの情報が必要となる。

1. **単語と概念の関係**
   単語と概念の関係は表1で表現できるような単語から概念集合への対応付けである。

表 1: 単語と概念の関係

単語	概念 1	概念 2	...
mouse	nouse(動物)	mouse(入力装置)	...
head	head(頭部)	head(指導者)	...
:	:	:	:

2. **概念と概念の関係**
   概念間の関係は図1のような木構造のシソーラスで表現される。図中で各ノードは概念ノードに相当し、エッジはIS-A関係を表す。

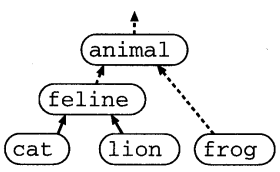

図 1: 概念と概念の関係

Using thesaurus to measure word similarity
Masanobu Kobayashi†,Xiaoyong Du‡, Naohiro Ishii†,
†Nagoya Institute of Technology,‡Renmin University of China

この2つの関係に基づき、単語 w_1 と w_2 の類似度を以下の過程で判定する。またその過程を図2に示す。

1. 単語 w_1,w_2 を概念の集合 C_1,C_2 にそれぞれ変換する。($|C_1| = m, |C_2| = n$,)
2. 概念集合 C_1,C_2 のそれぞれの概念ペアを比較し概念間類似度 $Sim(c_i^1, c_j^2)$ を計算する。
3. $m \times n$ 個の概念間類似度から単語間類似度 $Sim(w_1, w_2)$ を求める。

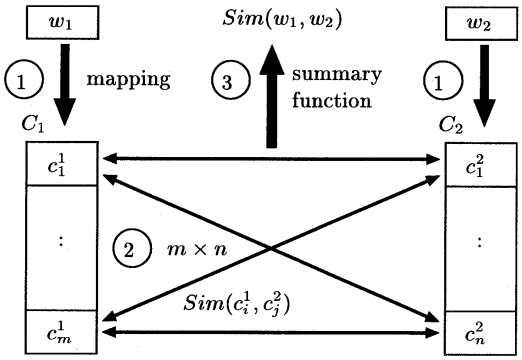

図 2: 単語間類似度の判定プロセス

一般に単語間類似度は概念間類似度の最大値とする。これは比較に値しない低い概念間類似度を示す概念ペアの影響を防ぐためである。

この様に単語間類似度は概念間類似度から求めるので、概念間類似度の判定法が重要である。次に実際の概念間類似度の判定法をいくつか示す。

3 距離を用いた測度と情報量を用いた測度

シソーラスを用いた概念間類似度の判定法は既にいくつか存在する。中でも最も単純なものが概念ノード間の距離を用いる方法である [1]。木の深さを H, 概念ペア c_1, c_2 間の距離を $dis(c_1, c_2)$ とした場合、概念間類似度 $Sim_{edge}(c_1, c_2)$ は以下の式で計算できる。

$$Sim_{edge}(c_1, c_2) = 2H - dis(c_1, c_2) \quad (1)$$

この場合、概念ノード間の距離が小さいほど両者は似ていることになる。一般的にノード間距離はエッジの数で定義しているが、シソーラス毎に概念ノードやエッジの数や構成が異なるため、全てのエッジを同じ長さ(重み)で扱うのは適切ではない。この問題点を解決するた

めに情報量に基づく方法が提案された [2]。この方法では距離を利用せず、概念ペアが共有する上位概念の情報量を利用し類似度を判定する。

概念ペア $c_1(\in C_1)$、$c_2(\in C_2)$ が共有する上位概念の集合を $CS(c_1,c_2)$ とすると、概念間類似度 Sim_{info} は以下の式で計算できる。

$$Sim_{info}(c_1,c_2) = \max_{c\in CS(c_1,c_2)}[info(c)] \quad (2)$$

ただし、情報量はコーパス (ドキュメントの集合) を用いて以下の式で計算する。

$$info(c) = -\log_2(p(c)) \quad (3)$$

$$p(c) = \frac{c\text{と下位概念の出現回数}}{\text{コーパスの総単語数}} \quad (4)$$

P.Resnik 氏の実験 [2] では情報量を用いた方が距離を用いる類似度よりも精度が良いことが示されている。しかし、情報量は用意したコーパスに依存しやすく、情報量の計算も複雑である。また、抽象的な概念ペアの類似度が必ず小さくなるといった問題も存在する。

4 情報量を用いた類似度の正規化

式 (2) より概念ペアの類似度はそれ自身の情報量を超えることはない。なぜなら概念間類似度が最大となるペアとは同じ概念同士の場合で、

$$\max_{c\in CS(c_1,c_2)}[info(c)] = info(c_1) = info(c2) \quad (5)$$

だからである。そのため抽象的な概念ペアを比較した場合、類似度の取り得る最大値は必ず小さくなってしまい判定の精度が落ちてしまう。

上記の問題を解決するために、本稿では概念ペアの情報量に基づいた概念間類似度の正規化を提案する。

$$Sim_{nrml}(c_1,c_2) = \max_{c\in CS(c_1,c_2)}[info(c)\times adj(c_1,c_2)] \quad (6)$$

$$adj(c_1,c_2) = \frac{\text{概念間類似度の最大値}}{Max(info(c_1),info(c_2))} \quad (7)$$

5 比較実験

英単語のペア 40 組を作成しアンケートにより人間の判定した類似度 (平均値) Sim_{human} を求めた。また同じペアについてシソーラスによる 3 つの類似度を計算し、人間のものと比較した。

なお実験ではシソーラスとして WordNet を利用した [3]。また情報量を用いた手法の実装においてコーパスが必要となるが、単純化のために実験ではコーパスを用いず式 (4) を以下の様に定義しシソーラスを用いて計算した。

$$\hat{p}(c) = \frac{c\text{の下位概念の数}}{\text{シソーラスの総概念数}} \quad (8)$$

表 2 に示す実験の結果から、我々の類似度 Sim_{nrml} と人間の判定との相関は 0.8160 と非常に高く, 既存の手法 $Sim_{edge}(0.7366), Sim_{info}(0.6268)$ よりも人間の判定に近いことがわかる。また、Sim_{human} とアンケート対象者それぞれとの相関の結果から、我々の手法が人間の判断する個人差の中にあることがわかる。

表 2: Sim_{human} との相関関係

順位	対象	相関
1	人 1	0.9254
:	:	:
21	人 21	0.8249
22	$\mathbf{Sim_{nrml}}$	**0.8160**
23	人 22	0.7938
:	:	:
27	人 26	0.7481
28	Sim_{info}	0.7366
29	人 27	0.6941
30	Sim_{edge}	0.6268

6 まとめ

本論文ではシソーラスを用いた単語間類似度の判定法を提案した。正規化された情報量を用いた我々の手法は、従来の類似度判定法よりも人間の判断に近いことが実験により示された。

今後の課題として、提案した類似度判定法の改良と実際の情報検索/ 文章分類システムへの実装・評価などが挙げられる。

参考文献

[1] Lee,Joo Hoo et al(1993)Information Retrieval Based on Conceptual Distance in IS-A Hierarchies;Journal of Documentation,49(2),p.188-207.

[2] Resnik, Philip(1995)Using Information Content to Evaluate Semantic Similarity in a Taxonomy;Proceedings of IJCAI-95,p.448-453.

[3] Miller,A.,George et al(1990)Introduction to WordNet:An On-Line Lexical Database;In Princeton University (Ed.),Cognitive Science Laboratory: Five Papers on WordNet ,p.1-10,Princeton University. http://www.cogsci.princeton.edu/~wn/