語の影響度と編集距離を用いた文の類似度計算手法

三浦直子† 高木友博†
†明治大学大学院理工学研究科基礎理工学専攻
Email :{n_miura,takagi}@cs.meiji.ac.jp

1. はじめに

ウェブ上には多様な種類の文書が存在しており,中でも SNS(Social Networking Service)が台頭している.我々が簡単に利用することができ,量も豊富であることから,企業戦略を立てる際や,マーケティングを行う際にも SNS は活発に利活用されている.

従来より,テキスト処理において, ベクトル空間モデルが広く利用されているが,単語の出現する順序は考慮されず,出現頻度や前後で共起する単語等を使用し,文書をベクトルとして表現する.

しかし,SNS は新聞の記事などに比べて短い文で構成されているため,ベクトル空間モデルは,SNS の解析には難点がある.長い文や記事等と比較して,短い文は構成する語数や共起単語が少ないからである.

そこで本研究では文間の類似度を単語の出現順序を考慮するため編集距離を用いて求め,さらに文内の単語の意味的類似度や影響度を考慮した手法を提案する.

第 2 節で関連技術と関連研究について述べ,第 3 節で提案手法,第 4 節で実験と考察,第 5 節でまとめとする.

2. 関連研究

近年,文を解析する際に辞書を導入する研究の例が多く見受けられる.[2] Aziz[2]らは,ブログに投稿された記事同士の類似度算出に文間類似度を用いている.各文を構成する名詞,動詞の単語間の類似度を階層構造を持つ辞書を利用して求め,それらを合計したものをブログ記事同士の類似度としている.

Li ら[1]は単語間の類似度を階層構造をもつ辞書を利用して求め,さらに語順を組み合わせた手法を提案している.階層構造を持つ辞書として WordNet がある.2005 年現在,WordNet のデータベースは約 11 万 5000 の synset に分類された約 15 万語を収録している. 名詞,動詞,形容詞,副詞を文法上の扱いが異なることから,区別して収蔵している.この関係の種類は品詞によって異なっている.

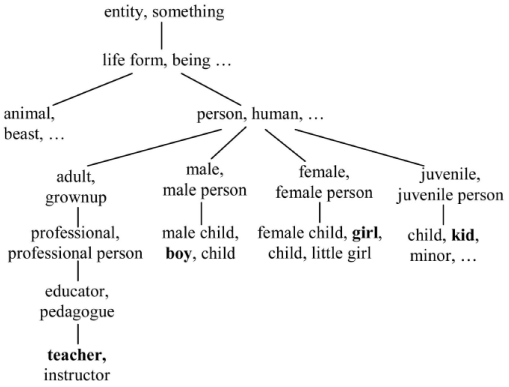

図 1 階層構造辞書

上位・下位関係は名詞と動詞には存在するが,形容詞,副詞には存在していない.例として名詞の場合を図 1 に示す.

Li ら[1]は単語間のパスとルートからの深さを用いて語同士の類似度を求める手法を提案している.本研究も単語同士の類似度はこの方法を利用し,単語 w_1 と w_2 間の類似度 $s(w_1,w_2)$ は

$$s(w_1,w_2)=e^{-\alpha l}\cdot\frac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}} \quad (1)$$

によって算出する.l は単語間のパスの総数である.図 1 を用いて算出例を示す.”boy” と ”girl” を結ぶパスは,*boy-male-person-female-girl* となり,パスは 4 本であるため l=4 となる.h は w_1 と w_2 の共通の親の synset のルートからの深さである. ”boy”と”girl”の場合共通の親の synset は”person,human…”の synset であり,この synset のルートからの深さが h となる.

さらに,階層構造を持つ辞書において,単語は上位層に存在するにつれて,より汎用的な意味を持つため,下位層の単語同士で比較するときよりも意味的な類似度は小さくなる.Li らは定数 α , β を, α =0.2 と β =0.45 に設定している.

語同士の類似性のみを比較するたけでなく,語順を考慮することが重要である.例えば,“dog bites Mike” と “Mike bites dog”の 2 文において構成する単語は同じであるが,意味は異なる.この 2 文にベクトルを用いた方法を適用すると,各文に対するベクトルは同一のものとなり,2 文間の意味の違いは考慮されず,類似度は高くなる. そこで,本研究では編集距離の概念と語の類似性の両方に基づく文間類似度を求める手法を提案する.

3. 提案手法

提案手法は,語の出現順序を考慮するために編集距離をベースとして,辞書を用いて求めた単語間の意味的類似度を導入し,さらに文における語の影響度を考慮に加える.

文 S_1(以下 S_1)を $S_1=\{a_0,a_1,\cdots,a_n\}$, 文 S_2(以下 S_2)を $S_2=\{b_0,b_1,\cdots,b_m\}$とする. S_1 は n 単語, S_2 は m 単語から構成されており, $a_i(0\leqq i\leqq n)$は S_1 中の i 番目の単語 , $b_j(0\leqq j\leqq m)$は S_2 中の j 番目の単語であるとする.ここで,2 文 S_1,S_2 間の類似度を $Sim(S_1,S_2)$と表す.その取り得る値の範囲は[0,1]とし,1 に近づくほど 2 つの文は類似しているとする.

3.1 編集距離の拡張

編集距離とは 2 つの文字列がどの程度異なっているかを示す距離である.本来,編集操作(置換,削除,挿入)は 1 文字に対して行われるが,提案手法では 1 単語に対しての操作に拡張を行う.以下に編集距離の一種であるレーベンシュタイン距離を用いた手法について述べる.

長さ n,m の S_1,S_2 に対し,2 文間のレーベンシュタイン距離 $L(n,m)$は以下の漸化式で定義される.

$$L(i,j)=\max(i,j) \quad \text{if } \min(i,j)=0,\ 0\leqq i\leqq n,\ 0\leqq j\leqq m$$

$$L(i,j)=\min\begin{cases}L(i-1,j-1)+c(a_i,b_j)\\ L(i,j-1)+1\\ L(i-1,j)+1\end{cases} \quad \text{otherwise.} \qquad (2)$$

ここで置換操作のコスト $c(a_i,b_j)$は次のように定義される.

$$c(a_i,b_j)=\begin{cases}0 & (a_i=b_j)\\ 1 & (a_i\neq b_j)\end{cases} \qquad (3)$$

3.2 語同士の意味的類似度の導入

上記編集距離の拡張のみでは,文同士の距離は単語同士が一致か不一致のみで計算され,2 つの単語が同じか類似した意味を指していても,表記が異なれば全く異なるものとして判断されてしまう.

そこで提案手法では,式(3)の代わりに単語間の類似度(式(1))を用いて式(4)に置き換え,意味的な異なり度合を表す指標を用いる.

$$c(a_i,b_j)=(1-s(a_i,b_j)) \qquad (4)$$

さらに正規化を行い, 最終的に 2 文の類似度 $Sim(S_1,S_2)$は以下の式によって計算される.

$$Sim(S_1,S_2)=1.0-\frac{L(n,m)}{\max(n,m)} \qquad (5)$$

また,式(1)の Li[1]らの手法は WordNet において上位・下位関係を持つ名詞,動詞には適用できるが,同意・反意関係しか持たない形容詞,副詞には適用できない.そこで提案手法では,形容詞と副詞の場合の単語間の意味的類似度を式(6)のように算出する.$w_1\in synsetA$, $w_2\in synsetB$ において,単語 w_1 と w_2 が同じ synset 内に存在しているなら単語間の類似度 $s(w_1,w_2)$は 0,否なら 1 と設定する.

$$s(w_1,w_2)=\begin{cases}0 & (synsetA=synsetB)\\ 1 & (synsetA\neq synsetB)\end{cases} \qquad (6)$$

3.3 比較文に対する単語の影響度

編集距離を用いて文間類似度 $Sim(S_1,S_2)$を算出する際に,次の問題点が考えられる. “I ate an apple ”, “I hate an apple”の 2 文を考えると,この文同士は反対の意味を指す.しかし”ate”,”hate”を除いた語は一致しており,また語順も一致しているので,先述の手法では,類似度が高くなる.しかし実感覚的には,”ate”,”hate”の 2 単語の影響でこの 2 文が類似した意味か否かが決定する.そこで本提案では,比較文に対する単語の影響度をコーパス文書から条件確率を用いて求め,重みとして辞書で求めた単語間の類似度(式(1),(6))に付与する.

S_1 の各単語の S_2 に対する影響度を P(I| S_2), P(ate| S_2), P(an| S_2), P(apple| S_2)を計算することで求め,文 S に対する単語 w の影響度 $weight(w)$は以下の式(8)で定義する.ここ

で S^*を文 S 内の,名詞,動詞,形容詞,副詞の単語群とする.文 S が”It is a pen”なら S*={“is”,”pen”}となる.

$$P(w \mid S) = \frac{w \text{と} S^* \text{が共起している文書数}}{S^* \text{が含まれる文書数}} \quad (7)$$

$$weight(w) = \frac{1}{(1+e^{-\gamma * P(w|S)})} \quad (8)$$

γは 5.0 に設定する.これより式(4)を次のように変更する.

$$c(a_i, b_j) = (1 - s(a_i, b_j) * weight(a_i)^{-1} * weight(b_j)^{-1}) \quad (9)$$

単語が比較文と多く共起している場合,影響度は低く,共起が少ない場合,影響度が高くなる.

4. 実験

4.1 データセット及び評価方法

SemEval2014 Task10[3]の Subtask ,STS English の 6 つのデータセットを利用する.表 1 に各データセットの概要を示す.各データセット毎に 750 の文のペアが含まれている.正解データとして,各データセット内の 750 文のペアに対して人手でスコアを付与したデータ(gold-standard,以降 gs スコア)を用いる.類似度は表 2 に示すとおり範囲は[0,5]で評価する.提案システムの評価には,出力結果と gs スコアのピアソン相関をとって精度とする.本実験では前処理として形態素解析し,記号の削除を行った.形態素解析には Stanford CoreNLP の POS Tagger を使用した.語の影響度を学習するコーパスとして Reuters Corpus[4]の全 806,791 記事を用い,1 記事を 1 文書とした.

4.2 実験

S_1 と S_2 間の文間類似度 $score(S_1,S_2)$は,評価方法に則して値の範囲[0,5]で計算し, $score(S_1, S_2) = 5 * Sim(S_1, S_2)$ とした.実験は以下の 2 種類の手法で行う.

(a) 編集距離を用いる場合
(b) 編集距離+語の影響度を用いる場合

表 1 データセット

データセット	概要
images	image description (PASCAL VOC2008)
OnWN	OntoNotes and WordNet definition mappings
tweet-news	News title and tweet comments
deft-forum	Discussion forum data in the DEFT
deft-news	News article data in the DEFT project
headlines	News headlines by European Media Monitor

表 2 SemEval2014 Task10 でのスコアの付与基準

点数	説明
5	2 つの文は全く同等である.
4	2 つの文はほとんど同等であるが, 重要でない情報が異なる.
3	2 つの文はほとんど同等であるが, いくつかの重要な情報が異なる/欠落している.
2	2 つの文は同等でない. しかし部分的に共有している情報がある.
1	2 つの文は同等でないが, 同じトピックである.
0	2 つの文は異なるトピックである.

4.3 実験結果

表 3 に実験結果を示す.表 3 は提案した 2 つの手法(a)(b)を 6 種類のデータセット(表 1)に対して実験した際の各精度(ピアソン相関)である.(a)と(b)を比較すると,全てのデータセットにおいて(b)の精度が(a)の精度を上回っている.

4.4 考察

表 4,5 はデータセット headlines 内の文のペアに対し,提案手法(a)(b)で計算した文間類似度と gs スコアを示したものである.この 2 つの表に基づいて考察を行う.
まず表 4 で手法(a)と(b)による類似度の差を比較する.この 2 文は”Russia”に対し”Italy”,”plane”に対し”coach”と,イベントを特定する単語が異なる事から文全体の意味が異なると判断されている.したがって各手法の計算結果は,gs スコアの 0 に近づくほど精度は良いことになる.

表 3　実験結果

	images	OnWN	tweet-news	deft-forum	deft-news	headlines
(a)　レーベンシュタイン距離	0.4340	0.3651	0.6403	0.2964	0.4360	0.4524
(b)　(a)+語の影響度	0.5249	0.4703	0.6520	0.3251	0.4746	0.5389

表 4　手法(a)と(b)との比較

	(a)	(b)	gs スコア
Death toll rises in Russia plane crash	2.9143	2.5165	0.0
Death toll rises to 39 in Italy coach crash			

表 5　課題を示す例

	(a)	(b)	gs スコア
Cuba's Castro assumes CELAC presidency	1.4286	2.6519	5.0
Cuba's Castro to Take Over CELAC Presidency			

この 2 文では”Death”,”toll”,”rises”,”in”,”crash”の 4 単語が一致し,かつ語順も変わらない.そのため編集距離を用いた手法(a)では,類似度が高く算出されている.

一方提案手法(b)は文に対する語の影響度を考慮するため類似度が低く算出され,(a)に比べて gs スコアに近くなっている.国名は固有名詞であり,本研究では考慮されていないが,”plane”と”coach”それぞれの文に対する影響度を加味したことにより(a)の精度を上回る結果を出すことができた. このことから,語の影響度を考慮に入れるほうが性能が高いと考えられる.

表 5 には,両手法とも良好な結果が得られなかった例を示す.文のペアの gs スコアは 5.0 であり,類似性が高いが,両手法とも低い類似度になっている.これは”assume”と”take over”が同じ意味を指すと判定されていないことが問題点と考えられる.本研究では 1 単語(単語 uni-gram)に対して処理を行っているため,熟語や連語の考慮ができていない.実際は”assume”と”take”の比較になってしまっているため表 5 のような結果が算出されている.

5. まとめ

本稿では,文間類似度を求める際に,編集距離に語の意味的距離や語の影響度を利用する手法を提案した.また実験により,単なる編集距離に比べて提案手法が良い精度を達成できることを確認した.

今後は 1 単語単位だけでなく,熟語表現にも注目する必要がある.

参考文献

[1] Yuhua Li,David McLean,Zuhair A.Bander,James D.O'Shea, and Keeley Crockett(2006). Sentence Similarity Based on Semantic Nets and Corpus Statistics . *IEEE Transactions on Knowledge and Data Engineering　VOL.18 No.8 August 2006*

[2] Mehwish Aziz, Muhammad Rafi(2010). Sentence based semantic similarity measure for blog-posts *Digital Content, Multimedia Technology and its Applications (IDC), 2010 6th International Conference*

[3] Eneco Agirre,Carmen Banea,Claire Cardie,Daniel Cer,Mona Diab,Aitor Gonzalez-Agirre,Weiwei Guo,Rada Mihalcea,German Rigau,Janyce Wiebe.(2014) SemEval2014 task10: Multilingual Semantic Textual Similarity.*2014 SemEval*

[4] Reuters Corpus English Language, 1996/08/20-1997/08/19 http://about.reuters.com/researchandstandards/corpus/