2F4-1

文書の類似性判定による情報伝達の差異に関する一考察

An Approach to Comparison among Message Contents based on Document Similarity

坂梨　優[*1]　　小林　一郎[*2]
Yu Sakanashi　　Ichiro Kobayashi

[*1][*2]お茶の水女子大学大学院人間文化創成科学研究科理学専攻
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Even though the same event is observed, depending on the viewpoints of observers, the event is differently reported. Like this, there is the possibility that the report becomes the one influenced and prejudiced if only a particular observer provides it. Therefore, in this study, we analyze news articles reporting the same event; compare the difference among news publishers in terms of the way of reporting; and then discuss their reporting viewpoints.

1. はじめに

1 つの事象を説明するのに，複数の情報提供者による異なる観察点が存在する．そのため情報提供者によって，同じ事象に対する情報の伝達に差異が生じ，1 つの情報源からのみ情報を受けるとその情報源の影響を強く受ける可能性が考えられる．このため複数の情報源が提供する情報を比較し，自分の観察点を持ってその内容を正しく捉える必要がある．このことから本研究では，同一の事象を説明する複数の文書を比較し，それぞれの情報源の観察点を把握しながら，その内容を正確に捉える事ができる手法の構築を目的とする．具体的には同一事象に対して，複数の新聞社が報道しているニュース記事を取り上げ，それらを比較することによって，それぞれの新聞社の報道の差異を把握できるようにする．

2. 関連研究

竹元ら [1] は文書から共起グラフを生成し，グラフが Small-World 構造持つことを利用したを要語の選定をして，類似文書検索を行っている．ここでは語同士の関連性を表わすリンクの作成の際の関連性を図る指標として Jaccard 係数を使用している．

村上ら [2] は文法や辞書などの事前知識に依存せず，文章の表層的な特徴に基づいて特徴点となる単語を抽出し，それを用いて文章間における類似箇所を発見する手法を提案している．ここでは，文章中の英数字，カタカナはそのまま抜き出し，漢字はバイグラムで分割することで特徴語とし，それらの出現パターンから文書間の類似性を発見している．しかしこの手法では単語が十分に抽出できないほどに短い，あるいは，ひらがなが多いような文章に対して適用することは難しい．

熊本ら [3] は辞書の語義文や文書軍の単語の共起を基に自動作成した概念ベースを利用した方法，及び，tf-idf 方式の類似判別能力について評価，考察を行った．tf-idf 方式は文書中に単語が存在する場合に非常に効果的であり，概念ベース方式は，文書中に単語が存在しなくても同義語や連想語がある場合に効果的であることを示している．各類似検索方式にはそれぞれ特徴があり，目的に応じてどのように使い分けたり，組み合わせたりするかが今後の課題となっている．

市川ら [4] は構文木付きコーパスから，構文的に類似した文を検索する手法を提案した．これは構文的類似度の計算手法として Collims[5] の提案する Tree Kernel を基に，インデックス化を用いた検索の高速化を可能とするアルゴリズム提案している．しかし，出力された類似文を人が判定すると類似していないものも多数含まれていることが確認された．表層の語の一致や，構文構造以外の構造を用いることが今後の課題となっている．

連絡先: 坂梨優，お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース小林研究室，〒112-8610　東京都文京区大塚 2-1-1，Tel.03-5978-5708，sakanashi.yu@is.ocha.ac.jp

3. 類似文書判定

類似文書を判定する手法としては, 類似文書検索に使われる様々な手法が存在するが，本研究では文書の細部にわたる相違の比較を目指し，‘文字列の一致’，‘単語の一致’，‘単語の持つ意味の一致’ の観点から，文書の比較を行うことを試みる．

3.1 文字列の一致

文字列の一致に基づく類似文判定には，Tri-gram を使用する．Tri-gram とは，隣り合った 3 文字の文字列の並びである．Tri-gram を採用した理由は，2 文字の文字列の並びである Bi-gram では，一致する文字列が多過ぎてしまい，4 文字以上の場合では，人物名などの固有名詞の一致を取るのが難しくなってしまうことによる．これらの理由により，本研究では Tri-gram の一致度に基づき文の類似性を判定する．Tri-gram に基づく類似度判定指標を式（1）のように定義する．

$$Sim_{tri} = \frac{\text{一致するトライグラムの数}}{\text{2 文中のトライグラムの総数}} \quad (1)$$

3.2 Jaccard 係数による単語の一致

内容を構成する単語の一致度の判定に Jaccard 係数を利用し，文の類似度を判定する．文 S および文 T から規則に従って単語を抽出し，抽出した単語の集合を，それぞれ A，B とする．このとき Jaccard 係数は，式（2）で表わされる．分母は文 S と文 T の重複を取り除いたときの単語の数，分子は文 S と文 T 両方に共通する単語の数を表し，Jaccard 係数値が大きいほど，2 文の一致度が高いことを示す．

$$Jaccard(A, B) = \frac{|\, A \cap B \,|}{|\, A \cup B \,|} \quad (2)$$

3.3 単語の持つ意味の一致

単語の持つ意味の一致には WordNet を利用する．WordNet は，Princeton 大学の認知科学研究所によって，心理学者である同大学の教授，George A. Miller の主導のもとで運営されている英語の概念辞書である [5]．WordNet では英単語が synset と呼ばれる同義語集合に分類されており，ある単語の同義語集合へのアクセスや，概念間の誘導を行うことが可能である．これを利用して，同義語の取得や，単語間の類似度を求めることができる．例えば ‘賛成’ と ‘同意’ という語彙は類似度が 1.0 であり，同じ意味としてとることができる．これにより，語彙の表層的な一致ではなく，語彙が持つ類義語体系の下で類似性を判定することができる．本研究では，WordNet の日本語版として開発された日本語 WordNet[6] を使用する．

3.4 Jaccard+WordNet

WordNet を利用して得られた単語集合の類似度を $sim(A, B)$ とすると，Jaccard+WordNet により求められる，単語 a（$a \in A$）と単語 b（$b \in B$）の類似度は，式（3）で表される．

$$Jaccard + WordNet = \frac{\sum_{a \in A, b \in B} sim(a, b)}{\mid A \cup B \mid} \qquad (3)$$

これにより、単語の表層的一致だけでなく類似性を考慮した類似判定ができるようになる．

4. 実験

4.1 実験対象テキスト

今回，比較対象テキストとして，表 1 に示すニュース記事を用いた．

表 1: 実験対象テキスト

新聞社	新聞社	記事のトピック	日付
朝日新聞	読売新聞	「民主党代表選」	2010 年 8 月 26 日
朝日新聞	読売新聞	「鳥インフルエンザ」	2011 年 2 月 3 日
朝日新聞	日経新聞	「新燃岳の噴火」	2011 年 2 月 4 日

上記 3 つのテキストデータセットに対して，Tri-gram，Jaccard 係数，Jaccard+WordNet の 3 つの手法を用いた実験を行った．類似文判定における正当性の評価については，提案手法によって得られる結果と人が予め作成した正解データとの比較を行うことにより検証した．それぞれの手順について説明する．

4.2 類似文の対応関係抽出処理

文書 D_1 と文書 D_2 における類似文を判定する際，通常，閾値による判定が必要となる．しかし，閾値を設定すると，文書 D_1 に対して，対応する文書 D_2 の文がいくつも抽出されてしまうことや，閾値は対象とする文書によって異なることが容易に考えられる．一方，閾値を定めずに文書 D_1 の各文に対応する文書 D_2 の文を抽出すると，文書 D_1 のすべての文に対応する文書 D_2 の文を抽出してしまうため文書 D_1 独自の視点を見ることができない．そこで本研究では，文書 D_1 と文書 D_2 の双方から見たクロスチェックをすることにより，類似文の抽出を行う（図 1 参照）．クロスチェックによる対応文抽出の手順は以下のとおりである．

step1. 文書 D1 中のそれぞれの文に対して最も類似する，文書 D2 中の 1 文を抽出する．

step2. 文書 D2 中のそれぞれの文に対して最も類似する，文書 D1 の 1 文を抽出する．

step3. 双方ともに順位が 1 位のものをそれぞれの文書の中で真に類似している文として採用し，それ以外のものは削除する．

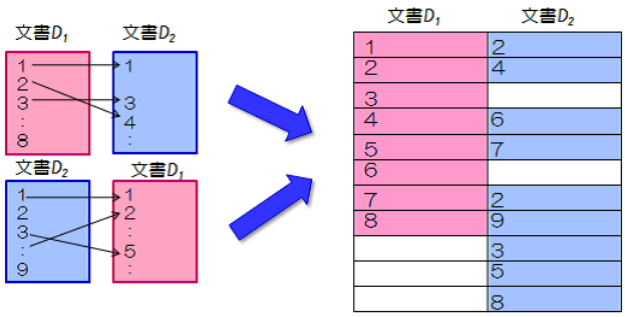

図 1: クロスチェックによる対応文抽出

4.3 文字列の一致に基づく類似文判定

文字列の一致による類似判定を行う．実験の手順は以下のとおりである．

step1. 文書 A と文書 B のすべての文のトライグラムをとり，A と B すべての文同士を比較して，何個のトライグラムが一致するかを調べる．

step2. それぞれの文章の長さが異なることを考慮し，以下の式の下，類似度 Sim_{tri} を判定する．

$$Sim_{tri} = \frac{\text{一致するトライグラムの数}}{\text{2 文中のトライグラムの総数}} \qquad (4)$$

4.4 単語の一致に基づく類似文判定

次に，単語の一致による類似文判定を行うために，実験を行った．実験の手順は以下のとおりである．

step1. 文書 A と文書 B のすべての文章を日本語形態素解析器 MeCab[8] を使い形態素解析する．

step2. 形態素解析したものの中から，文章をよく表現していると考えられる名詞と動詞を抽出し，比較対象とする．

step3. 各文ごとに抽出した単語の重複をすべて取り除く．

step4. Jaccard 係数による値を求める．

4.5 Jaccard+WordNet に基づく類似度算出

実験の手順として，Jaccard 係数による判定の step3 の後に step4 として，Jaccard+WordNet による値を求める．

ニュース記事「新燃岳の噴火」を比較対象として，上述した単語の意味の一致に基づく類似文判定処理および 4.2 節に示した類似文対応関係抽出処理の結果を表 2 に示す．表 2 における項目「トピック」には，著者が文の内容を判定することによりその内容を示すラベルを付与した．

表 2: 2011 年 2 月 4 日の「新燃岳噴火」の記事に対する Jaccard+WordNet での実験結果

文	朝日新聞	文	日経新聞	類似度	判定	トピック
1	気象庁の火山噴火予知連絡会（会長・藤井敏嗣東京大名誉教授）は3日、拡大幹事会を開き、霧島連山・新燃岳（しんもえだけ）の噴火について今後 1～2週間は「現在と同程度の爆発的な噴火が続く」との見解をまとめた。	1	活発な活動が続く宮崎、鹿児島県境にある霧島山・新燃岳（1421 メートル）について、気象庁の火山噴火予知連絡会（会長・藤井敏嗣東京大名誉教授）は3 日、緊急の拡大幹事会を開き、「当分の間、現在と同規模の爆発的噴火を繰り返す」との見解をまとめた。	0.49	○	見解
2	終息の見通しは現時点では不明で、長期的に観測を強化すべきだとしている。	10	長期的な見通しについて、藤井会長は「噴火活動がいったん鎮まった後、大きな噴火を繰り返す事もありえる」と説明。	0.15	○	予想
3	拡大幹事会では、これまでの活動を分析。	3	幹事会後の記者会見によると、新燃岳の火口付近には 700 万～1700 万立方メートルの溶岩が堆積。	0.11	×	分析
4	火山灰などを大量に噴出する初期の状況を脱し、火口にたまった溶岩に出口をふさがれた火山ガスの圧力が非常に高 くなって、頻繁に爆発を繰り返す段階に入っていると解釈した。	4	火口に蓋をする形となり、内部で火山ガスの圧力が高まって爆発的噴火を繰り返している。	0.40	○	解釈
5	このため、大量に火山灰を噴出するような初期の状況にすぐには戻らないとの見通しを示した。				○	予想
6	また、今回の噴火を江戸時代の１７１６～１７年に2年ほど続いた噴火以来「３００年ぶりの本格的なマグマ噴火」と位置付けた。				×	記録
7	藤井会長は「現在の段階を 過ぎても、江戸時代の例を考えれば、そのまま沈静化するとは考えづらい」と話した。				○	見解
8	いったん沈静化してもまた、大規模な噴火がおこる可能性があるという。	9	ただ、雨が降った場合は少量でも土石流や泥流が起こる可能性があるという。	0.20	○	可能性
9	火口の溶岩は当初、盛り上がった形の溶岩ドームだったが、藤井会長は「すでに溶岩ドームの状態ではない」と指摘。				○	見解
10	その溶岩は直径７００メートルの火口に 池のような状態でたまり、その容積は火口の半分前後に達したとみる。				×	状態
11	このため、爆発的噴火で溶岩が火口の外側に飛び出しやすくなり、火砕流を発生させる恐 れが出てきているという。	4	火口に蓋をする形となり、内部で火山ガスの圧力が高まって爆発的噴火を繰り返している。	0.36	×	見解
12	今後の監視体制について藤井会長は「地殻変動の観測を強化して地下のマグマの動向をきちんと把握することが重要」と強調した。	11	東京大や国土地理院などが地震計やＧＰＳ観測点を増設しており、「マグマの供給量を注意深く監視し続けることが重要だ」と話した。	0.18	○	見解
		2	現時点では居住地域まで達する大規模な火砕流の恐れ は少ないが、降雨時には土石流や泥流に注意が必要という。		○	見解
		5	霧島山周辺の全地球測位システム（ＧＰＳ）観測では、2009 年 12 月ごろから地面が膨張、本格的な噴火が始まった今年1月 26 日から収縮に転じた。		○	経緯
		6	地下にマグマが供給されて地面を押し上げ、噴火でマグマがはき出されて地面が縮んだとみられるが、同月 31 日からは収縮が鈍化しており、藤井会 長は「噴火に見合うだけのマグマの供給が続いている可能性がある」と話した。		○	解説
		7	爆発的噴火は少なくとも１～２週間は警戒が必要という。		○	見解
		8	爆発的噴火に伴い、火砕流が起こる可能性もあるが、過去最大とされる享保噴火（1716～17 年）の例などから最大でも火口から3キロ程度までしか広がらないと分析。		×	分析

4.6 実験結果

Tri-gram，Jaccard 係数，Jaccard+WordNet の類似度 0.5 以上，0.33 以上，0.25 以上の 5 つを比較した結果を表 3 に示す.

表 3: 3 つの手法の実験結果の正答率

記事のトピック	Tri-gram	Jaccard 係数	Jaccard+WordNet		
			閾値 0.5	閾値 0.33	閾値 0.25
「民主党代表選」	0.43	0.74	0.74	0.77	0.77
「鳥インフルエンザ」	0.30	0.50	0.50	0.50	0.50
「新燃岳噴火」	0.46	0.78	0.78	0.67	0.80

上記，正答率の算出方法は，新聞社 A のニュース記事の各文に対して，新聞社 B のニュース記事の文の対応関係が正解データと一致しているものの割合と，新聞社 B としての独自の情報として得られる文に対する正解データとの一致率の合計を，その手法による一致率としている.

4.7 考察

4.7.1 手法の一般的な特徴

表 3 において，すべての記事について，おおよそ Tri-gram，Jaccard 係数による単語の一致，Jaccard+WordNet の順に正答率が上がっていることがわかる.

また，記事のトピック「民主党代表選」における Jaccard+WordNet においては，単語間の類似度判定に用いた閾値の設定において，閾値を下げていくことにより正答率が上がっていることがわかる．これは，閾値が下がったことにより，類義語としてとられなかった単語間の類似度が加算されたことが原因であると考えられる．一方，記事のトピック「鳥インフルエンザ」では，Jaccard+WordNet においては，閾値の値の変化に対して正答率が変化していないことが確認された．これは，ニュース記事ペアにおいて類義の関係をもつ単語が無かったためと考えられる．さらに，記事のトピック「新燃岳噴火」における Jaccard+WordNet を参照してみると，閾値を 0.5 から 0.33 へ下げたことにより，正答率が減少し，さらに閾値を 0.33 から 0.25 へ下げることにより，再度，正答率が上昇している．一般的に，閾値を下げることにより，単語間の類似度が加算されることになるため単調に正答率が上がっていくのではないかと考えてしまうが，本提案手法では類似文の判定処理において類似度が 1 位のものを採用しているため，今回の実験において，閾値が 0.5 の時に順位が 2 位であった文が，閾値 0.33 の時に類似度が加算される単語を含んでいたために，順位が逆転し，正答率の低下を引き起こしていた．また，閾値を 0.25 に下げたことにより，さらに多くの単語間の類似度が加算されることになり，閾値 0.33 の時に正解では無くなった文が，再度，順位を上げると共に，他の文も同様に順位を変更し，正答率が上がったことが確認できた.

このように，閾値を下げることは類似文抽出の再現率を高めることができるが，文章比較の精度を下げる可能性があると言える.

4.7.2 比較内容

表 2 に示す新燃岳噴火のニュース記事において，新聞社間における記事を比較をする．朝日新聞と日経新聞の双方のニュース記事において，対応関係が正答しているものにおいては，共通する単語や，類義関係にあると思われる単語が多く含まれていることが分かる（例：朝日：文 1 －日経：文 1，朝日：文 2 －日経：文 10，等）.

また，朝日新聞の文 5，文 7，文 9，日経新聞の文 2，文 5，文 6，文 7 行目では，お互いに対応する文がないことが示せており，それぞれの記事独自の内容を抽出できたと言える．一方，朝日新聞の文 6 と日経新聞の文 8 においては，時間的な情報（文中において，年号が共通）が一致していることから，そ

の対応が正解とされていたにも関わらず，類似文として判定することができなかった．このことから，提案手法においては，文中における時間情報や数字による情報を正確に類似判定できていないと考えられる．

また，朝日新聞の文 3 と日経新聞の文 3 のように，内容が異なっているため，正解データでは類似文として判定されなかったものが，共通する単語を多く含んでいたことにより類似文として抽出されるケースも存在し，単語による一致は取れていても正確に類似文であるとは言い切れないことがわかる．

また，正解データの一つである，朝日新聞の文 10 と日経新聞の文 3 のペアにおいては，双方の文において共通している単語の数が多いことから類似文として判定されることが期待されるが，1 文中の語彙数が多いため，Jaccard 係数に基づく類似判定では類似度が低いと判断されてしまい，結果的に正しく類似文として判定されなかった．このように類似文判定において，1 文中の語彙数を考慮する必要があることを確認した．また，今回の例では，双方の記事に藤田会長の発言が多く出現する．しかし，双方の記事に共通した発言だけが記述されているのではなく，片方の記事には記述されていない発言があるなど，同じ人物の発言に注目することで，それぞれの新聞社が強調したい内容や重視したい点を明らかにできると考える．また，クロスチェックの問題点として，お互いの上位 1 位が必ずしも同じとは限らない点や，仮に片方の記事で 1 文で述べていることがもう片方の記事では 2 文にわたって述べられている場合に 2 文中の 1 つの文がはじかれてしまうという問題点が挙げられる．また，同じ記事に類似した文が 2 つ以上存在する場合にも同じことが起こると考えられるが，これはクロスチェックをする前にあらかじめ同文書内での類似文を 1 つに統合しておくことで解決すると考えられる．

5. おわりに

本稿では，Tri-gram を利用した文字列の一致，Jaccard 係数を利用した単語の一致，さらに Jaccard 係数による単語の一致度に，日本語 WordNet に基づく単語間の類似度を加えた判定手法である accard+WordNet による単語の持つ意味の一致を通して，複数文書の類似文判定を行い，Jaccard+WordNet が最も類似文判定に適していることを確認した．さらに，類似文判定においてクロスチェックを導入することで，閾値によらない類似文判定手法を提案した．これにより，比較対象となる双方の文書の情報伝達内容の差異を抽出することが可能になった．また，Jaccard+WordNet の正答率を閾値を変更することにより精査し，その結果，閾値を下げることにより類似文判定の再現率を上げることはできるが，文書の比較において精度が下がる可能性があることが判明した．

提案手法による類似文判定では比較的高い精度が得られたが，該当手法は，語彙の一致と語彙体系の知識のみに基づく類似文判定に基づいているため，それによって判定されない類似文を捉えることはできていない．そのことを考慮し，今後，人の名前，キーワードの重要性，時間や場所の情報など，イベント情報の特徴を捉えるヒューリスティックな知識を，類似文判定に取り入れることで，さらに精度の高い判定を実現し，文書間の情報伝達の差異を正確に抽出することを目指す．また、文の長さを考慮した類似文判定処理として、構文の一致に基づく類似判定の導入や、文章中の複数の文に 1 つの情報が分散されていたり、クロスチェックで上位 1 位の文のみをとることの不具合の解消を行う。

参考文献

[1] 竹元 勇太，沢井 康孝，山本 和英：SmallWorld による類似文検索のための重要語選定　言語処理学会　第 14 回年次大会　発表論文集　pp951-954，2008.

[2] 村上 智哉，中谷 直司，厚井 裕司，後沢 忍：辞書に依存しない文章間類似度の比較評価手法　情報処理学会　研究報告　2008(4)，pp.115-120，2008.

[3] 熊本 睦，島田 茂夫，加藤 恒昭：概念ベースの情報検索への適用～概念ベースを用いた検索の特性評価～　電子情報通信学会技術研究報告．AI，人工知能と知識処理 98(498)，9-16，1999.

[4] 市川 宙，橋本 泰一，徳永 健伸，田中 穂積：テキスト構文構造類似度を用いた類似文書検索手法　情報処理学会研究報告．情報学基礎研究会報告　2005(42)，39-46，2005.

[5] http://wordnet.princeton.edu/

[6] http://nlpwww.nict.go.jp/wn-ja/

[7] 「新 s あらたにす」 http://allatanys.jp/

[8] MeCab，http://mecab.sourceforge.net/