同一事象を報告する複数文書の比較に基づく個々の情報伝達の差異に対する一考察

坂梨 優　(指導教員：小林 一郎)

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．

このことを踏まえ，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　類似文書判定

類似文書判定は，一般に ‘文字列の一致’，‘単語の一致’，‘構文の一致’，‘意味・内容の一致’ などの観点から行われる．本研究では，文字列の一致として tri-gram を用いた手法，単語の一致として Jaccard 係数を用いた類似文書判定手法，さらに，意味・内容の一致として，Jaccard 係数を求める際の単語の一致度に日本語 WordNet[1] に基づく単語間の類似度を加えた判定手法の 3 つ用意し，どの手法が類似文書判定に適しているかについての予備実験を行った．その結果，Jaccard 係数に WordNet を組み込んだ手法の判定精度が他の 2 つの手法より良いと判断し，その手法を採用した．以降，その手法を ‘Jaccard+WordNet’ と呼ぶことにする．

2.1　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 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，以下の式で表わされる．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．

$$Jaccard(A,B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}$$

2.2　WordNet

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 [1]．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[2] を使用する．

2.3　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 により求められる類似度は，以下の式で表される．

$$Jaccard + WordNet = \frac{\mid (A \cap B) + sim(A,B) \mid}{\mid A \cup B \mid}$$

3　実験

本実験では，2 つの文書内の文の類似性を判定することにより，各々の文書において伝達している情報の差異から，それぞれの主張の違いを把握する．

3.1　実験仕様

今回の実験において，比較対象テキストとして，表 1 に示すニュース記事を利用した．

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	「日本海側の大雪」	2011 年 2 月 1 日
朝日新聞	読売新聞	「天然ウナギの卵発見」	2011 年 2 月 2 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

提案手法に基づく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する．

3.2　Jaccard+WordNet に基づく類似度算出

単語の一致による類似文判定を行う．判定の手順は以下のとおりである．

step1. 文書 D_1 と文書 D_2 のすべての文章を MeCab 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. WordNet を利用して単語間の類似度を求める．

step5. Jaccard 係数の分子に step4 で求めた類似度を加え，$Jaccard + WordNet$ の値を求める．

3.3　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ため，本研究では，文書 D_1 から見た文書 D_2 の類似度の順位と文書 D_2 から見た文書 D_1 の順位をクロスチェックすることにより，類似文の抽出を行う．

表 2: 2011 年 2 月 3 日の「鳥インフルエンザ」の記事に対する実験結果

文	朝日新聞	文	読売新聞	類似度	判定	トピック
1	大分県は2日、大分市宮尾の養鶏場で、採卵鶏３８羽が死に、鳥インフルエンザの遺伝子検査で感染力の強い高病原性ウイルス（Ｈ５亜型）が確認された と発表した。	1	高病原性鳥インフルエンザ問題で、農林水産省と大分県は２日、大分市の養鶏場で感染を確認したと発表した。	0.47	○	発表
2	県は、この養鶏場から半径１０キロ内を移動制限区域とし、飼育中の約８１００羽の殺処分を始めた。	8	同県はこの養鶏場で飼う約８１００羽を殺処分する。	0.38	○	決定
3	今回の感染は、国内の養鶏場では今季１１例目となる。	3	全国で１１例目となった。	0.40	○	件数
4	県によると、２日午後2時２０分ごろ、養鶏場から「鶏がたくさん死んでいる」と連絡があった。	4	同省によると、同日午後、食用の卵を出荷する採卵用の養鶏場から「鶏が前日の２倍以上死んでいる」と同県に通報があった。	0.34	○	通報
5	遺伝子検査で、死んだ鶏６羽中５羽、同じ鶏舎の５羽中４羽で感染が確認された。	7	遺伝子検査でＨ５型のウイルスが検出された。	0.23	○	確認
6	飼育されている採卵鶏には、産卵数が少なくなるなどの症状があるという。				○	症状
7	半径１０キロ内には臼杵市、豊後大野市、津久見市の一部が入るが、養鶏場は大分、臼杵両市の１１カ所。				○	地理
8	計３２万２６１０羽がいる。	5	１日は１２羽だったが、２日は３８羽が死んでいた。	0.36	×	数量
9	内訳は、採卵鶏が７カ所で約２４万１２００羽、肉用鶏が２カ所で約８万１２００羽、自家用が２カ所で約２１０羽となっている。				○	内訳
10	大分県では２００４年２月、九重町で飼育されていたチャボが高病原性鳥インフルエンザ（Ｈ５Ｎ１型）に感染した事例がある。	11	昨年１１月以降、島根、宮崎、鹿児島、愛知県の計１０養鶏場で高病原性鳥インフルエンザの感染が確認されている。	0.31	×	病名
		2	この冬の家畜では、大分県で初めて。		○	記録
		6	県で死んだ鶏を含む１１羽を簡易検査したところ、８羽で陽性反応。		×	確認
		9	また、半径１０キロ圏内の鶏や卵の移動を禁止した。		×	決定
		10	同圏内には、１１戸の養鶏場があり、約３２万羽が飼育されているという。		×	数量

抽出の手順は以下のとおりである．

step1. 文書 D_1 中の文それぞれに対して最も類似する，文書 D_2 中の 1 文を抽出する．

step2. 文書 D_2 中の文それぞれに対して最も類似する，文書 D_1 中の 1 文を抽出する．

step3. 双方ともに順位が 1 位のものをそ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．

3.4 実験結果

表 2 に 2011 年 2 月 3 日の「鳥インフルエンザ」の記事に対する実験結果を示す．

また表 3 に各記事の類似文判定の正答率を示す．

表 3: 正答率

記事のトピック	正答率
「民主党代表選」	0.65
「日本海側の大雪」	0.71
「天然ウナギの卵発見」	0.77
「鳥インフルエンザ」	0.64
「新燃岳の噴火」	0.75

3.5 考察

それぞれの記事に対して，6 割以上の正答率を得ており，提案手法の有効性が示せたと考える．表 2 の結果を詳細に眺めると，朝日新聞では，採卵鶏の症状について（朝:文 6），読売新聞では大分県での今季の感染は今回が初であること（読:文 2）が独自の内容となっており，各新聞社における報道の視点の差異を抽出することができた．一方，提案手法では，類似度が最大の文のみを対象としているため，同じ記事内に同じトピックに分類される文がある場合，正しい文が最も類似する文として抽出されないことや，片方の記事では 1 文で述べていることを，もう片方の記事では 2 文にわたって述べている場合にも，2 文の内どちらかが一致しない文と判断されてしまう場合もあった．また，文章中にキーワードとなるような同じ単語が使われている場合，時制の情報を無視して類似文と判定してしまうことや，数字を扱った文章に誤った判定がされてしまうこと，また辞書内にない単語が関係のない単語と一致してしまうことから，時制，数量など数字による情報の考慮や，病名などの辞書登録されていない単語の考慮も必要であると考えられる．

4 おわりに

本稿では，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類似文判定を行い，さらに，クロスチェックを行うことで，文書の情報伝達内容の差異を抽出した．Jaccard+WordNet での類似文判定では比較的高い精度が得られたが，語彙の一致と語彙体系の知識のみのため完璧に判定できたとは言えない．そこで，今後，人の名前，キーワードの重要度，時間や場所の情報など，イベントを捉える特徴となる情報のヒューリスティックな知識を判定に取り入れることで，さらに精度の高い類似文判定が可能となり，文書間の情報伝達の差異が抽出しやすくなると考えられる．

参考文献

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

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

[3] 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.

[4] 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.

[5] 「新 s あらたにす」 http://allatanys.jp/