文書横断文間関係の特定

宮部 泰成 † 高村 大也 †† 奥村 学 ††
東京工業大学大学院総合理工学研究科 † 東京工業大学精密工学研究所 ††
miyabe@lr.pi.titech.ac.jp {takamura,oku}@pi.titech.ac.jp

1 序論

一般にテキストは，文という意味単位に分割され，その単位間には様々な関係が成立することが分かっている．このようなテキストの文間の関係を解析し，テキストの構造を明らかにすることを，談話構造解析という．従来の談話構造解析は，修辞構造理論（RST）に基づいた単一文書内の談話構造を解析するもの [1, 2] であった．しかし，Radev[3] は，RST を異なる文書中の文間関係に拡張し，CST（Cross-document Structure Theory）理論を構想した．衛藤ら [4] は，Radev の定義した関係を基に，日本語の新聞記事集合に対し，14 個の文書横断文間関係を定義した．文書横断文間関係としては，異なる文書中の 2 つの文が同じ内容を述べている「同等」や文間で数値が変化する「推移」などがある．図 1 は,「i-モードのサービス開始に関するトピック」の文書横断文間関係の例である．文書 A の文 1 と文書 B の文 1 の間では,「i-モード」の契約件数が変化しているので,「推移」関係が成り立ち，文書 A の文 3 と文書 B の文 3 の間では，同じ内容を述べているので「同等」関係が成り立つ．

文書横断文間関係を特定することは，複数文書要約や情報抽出等において有用である．例えば，文書中の文間で同じ内容が書いてあると認識できれば，複数文書要約において，冗長な要約を避けることができる．また，文書中の文間で数値の変化を述べていると認識できれば，動向情報の抽出が可能となる．

本研究では，文書横断文間関係の「同等」と「推移」という 2 つの関係に着目し，これらの関係を機械学習を用いて特定することを目的とする.「同等」関係の特定では，2 つの文の類似度でデータを複数のクラスタに分ける手法と，2 段階の特定手法を提案する．この 2 つの手法を組み合わせた手法により，優れた結果を得られることを示す.「推移」関係の特定では,「Aが１０万円となった。」という文のA (以後,「数値を値として持つ名詞句」と呼ぶ) を係り受け情報を用いて抽出し,「数値を値として持つ名詞句」の類似度を用いることで従来手法より優れた結果になることを示す．

2 文間関係を特定する関連研究

文書横断文間関係を特定する研究に，Zhang ら [5] がある．Zhang らは，2 文間で一致した品詞の数などを素性として使用し，文書横断文間関係の特定を行なった．しかし，Zhang らの手法では，各関係の特定で，同じ分類器を用いており，各関係に適した手法で特定していないため，精度が良くない．

また,「推移」関係を特定する関連研究に，難波ら [6] がある．難波らは,「推移」と「更新」のみのデータから，ルールを用いて,「推移」の特定や「更新」の特定を行なった．しかし「推移」と「更新」以外を含んだデータから,「推移」や「更新」を特定する精度は十分ではない．

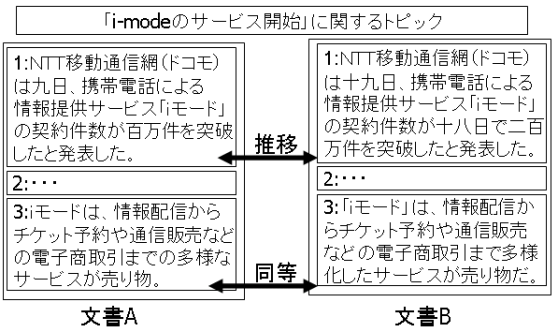

図 1: 文書横断文間関係の例

3 「同等」関係の特定

ここでは,「同等」関係の特定手法を説明する．関係の特定規則を人手で作成することは，コストがかかるため，本研究では，規則の獲得に機械学習手法を用いることにした．関係の特定は，文ペアが与えられたとき，そのペアが「同等」関係か否かの 2 値分類問題と考えることができる．よって，2 値分類問題において，高い汎化能力がある Support Vector Machine(SVM) を学習器として使用する．

3.1 「同等」特定モデル

「同等」関係の特定は，[7] の手法と同様に「2 つの文のコサイン類似度でクラスタに分ける手法」と「2 段階の特定手法」を組み合わせた手法で特定する．ここでの 2 つの文 (S1，S2) のコサイン類似度は，以下の (1) 式で計算される:

$$cos(S1, S2) = \frac{U1 \cdot U2}{|U1||U2|}. \quad (1)$$

U1，U2 は文 S1，S2 に現れる単語 (名詞，動詞，形容詞) の頻度ベクトルを表す．

クラスタは，図 2 のように「表層的に大変類似しているクラスタ」と「ある程度類似しているクラスタ」に分ける.「表層的に大変類似しているクラスタ」とは，下記の例 1 のように名詞や動詞だけでなく，助詞や助動詞などの機能語も類似しているクラスタとし,「ある程度類似しているクラスタ」とは，下記の例 2 のように機能語は似ておらず,「簡略」や「詳細」という 2 つの「同等と似た関係」が存在するクラスタとする．ただし,「ある程度類似しているクラスタ」は，非常に負例が多く，特定が困難になることが予想される．よって,「ある程度類似しているクラスタ」もある閾値で,「ある程度類似しているクラスタ」と「負例が多く特定が困難なクラスタ」に分ける．クラスタに分ける類似度の閾値は 4.1 節で説明する．

例 1：表層的に大変類似しているクラスタ
S1 成果は二日付の英科学誌「ネイチャー」に掲載される。
S2 この成果は 2 日発行の英科学誌「ネイチャー」に掲載される。

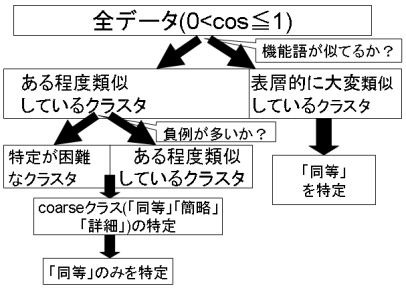

図 2: 本研究の「同等」特定モデル

例 2：「簡略」関係である 2 文

S1 ９６年６月に加藤さんの家族から明美さんとの結婚準備金の名目で１００万円を借りており「金の返済を迫られたので殺した」と供述しているという。

S2 同容疑者は９６年６月に加藤さんの家族から結婚準備金の名目で１００万円を借りていたという。

また，2 段階の特定手法とは，図 2 のように最初に「同等」と「同等と似た関係」を一つの上位クラス (coarse クラス) にまとめて特定し，次に coarse クラスから「同等」のみ (fine クラス) を特定する手法とする．以下では，この 2 段階の特定手法を「coarse-to-fine 特定法」と呼ぶことにする．

クラスタに分けて学習する手法と coarse-to-fine 特定法の 2 つの手法を組み合わて,「同等」を特定する (図 2).

3.2 「同等」特定で使用する素性

学習時に用いる素性の説明をする．本研究は，2 つの文がある関係か否かの特定を行なうので，文ペアが 1 事例になる．素性は，[7] とほとんど同じ種類の素性を使用した．どのクラスタの特定でも使用する 14 種の基本素性と coarse-to-fine 特定法の fine クラスで使用する素性を下記で説明する．

1. 14 種類の基本素性
   - 2 つの文のコサイン類似度，bigram 類似度，trigram 類似度，文節間類似度，意味的類似度
   - 2 つの文を含んだ段落間類似度，文書館類似度
   - 各文の接続詞，各文の文末表現，各文の位置，各文の文字数
   - 固有表現類似度，固有表現が係る格助詞の一致
   - 2 つの新聞記事の掲載日の差
2. fine クラスの特定で使用する素性 (20 種類)
   - 前述した基本素性
   - 2 つの文の形態素数の差，文節数の差
   - 主動詞，数，単位が一致するかどうか
   - 主題（初出のガ格とハ格）が一致するかどうか

4 「同等」特定の実験

前節で述べた「同等」の特定手法の性能を実験により示す．コーパスは，テキスト自動要約タスク TSC2，3(Text Summarization Challenge)[8]，Must（Workshop on Multimodal Summarization for Trend Information）[9] に，衛藤 [4] が定義した関係を付与したものである．1 文対 1 文で関係が対応している 471586 個の文ペアから 798 個の「同等」を特定する実験を行なった．評価は，10 分割交差検定で行なった．

表 1: 各閾値での効果的でない素性の結果例

閾値	効果的でない素性
0.9	格助詞，文節類似度，意味類似度
0.89	意味類似度，文末表現，**bigram 類似度**，格助詞
0.88	**bigram 類似度**
0.87	掲載日の差，文書間類似度，文の長さ，格助詞，文末表現，段落間類似度，文の位置，**bigram 類似度**

4.1 閾値の推定

本研究の「同等」の特定モデルでは，あるコサイン類似度の値でクラスタに分ける．クラスタに分けるときの閾値の推定は，訓練データを更に 10 分割交差検定する development test で推定する．

4.1.1 「表層的に大変類似しているクラスタ」と「ある程度類似しているクラスタ」の閾値

まず,「表層的に大変類似しているクラスタ」と「ある程度類似しているクラスタ」の閾値の決め方を説明する．

「表層的に大変類似しているクラスタ」の 2 文は，自立語（名詞，動詞，形容詞）+機能語（助詞，助動詞）の連接レベルで類似しているが,「ある程度類似しているクラスタ」では，連接レベルで類似していない．そのことから,「ある程度類似しているクラスタ」の「同等」の特定においては，bigram 類似度の素性を使用しない方が，bigram 類似度の素性を使用したときより，精度と再現率の 2 つの値は良くなると仮定した．

それを確かめるために，各閾値で 14 種の基本素性から，bigram 類似度の素性を省いて残りの 13 種の素性を使用して development test を行ない，そのときの精度と再現率の 2 つの値が，基本素性を使用したときの値よりも高くなるか調べる．説明のため，素性を省いて 13 種の素性を使用したときの精度と再現率の 2 つの値が，基本素性を使用したときの値よりも高くなるとき，省いた素性を効果的でない素性とよぶ．コサイン類似度の閾値を 1 から 0.01 ずつ減らしていき，各閾値での効果的でない素性を調べた結果，表 1 の例のように，bigram 類似度の素性が効果的でない素性となる閾値が見つかり，以後ほとんどの場合において閾値を減らしても効果的でない素性となっている．同様にその他の素性でも効果的な素性が無いか調べた結果，bigram 類似度のような現象は見られなかった (表 1)．よって，bigram の類似度が初めて効果的でない素性となるときを閾値とした．

4.1.2 「ある程度類似しているクラスタ」と「負例が多く特定が困難なクラスタ」の閾値

次に,「ある程度類似しているクラスタ」と「負例が多く特定が困難なクラスタ」の閾値の決め方を説明する．

本研究のデータは，類似度の値が減っていくにつれて,「同等」の数は少なくなっているが「文間関係が存在しないペア」の数は大変多くなっている．そのため，類似度を下げていくにつれて，特定精度が悪くなることが予想できる．それをふまえ，ある類似度の閾値で,「ある程度類似しているクラスタ」の F 値が一番良くなる値があると仮定した．4.1.1 節で決定した閾値から，類似度を順に 0.01 づつ減らしていき，各類似度を閾値としたときの F 値を調べた結果，表 2 の例のように，ある閾値

表 2: 各閾値での F 値の結果例

閾値	精度	再現率	F 値
0.58	55.08	16.64	25.56
0.57	**52.81**	**16.93**	**25.64**
0.56	49.15	14.45	22.34
0.55	51.51	14.84	23.04

表 3: 「同等」の特定結果

	精度	再現率	F 値
ベースライン	87.29	57.35	69.22
基本素性			
div	81.98	59.40	68.88
Notdiv	86.10	59.49	70.36
旧 Mix モデル [7]	86.67	60.32	71.14
新 Mix モデル	94.96	62.27	75.22
最適な素性の組合せ			
div	80.93	59.74	68.63
Notdiv	86.11	60.16	70.84
旧 Mix モデル [7]	86.31	60.56	71.18
新 Mix モデル	94.99	62.65	75.50

で，F 値が最も高くなる点があった．よって，F 値が最も高くなるときを閾値とした．

4.2 「同等」特定の実験結果

提案手法の性能を, 表 3 に示す．比較は，提案手法と以下の 4 つのモデルでおこなう．結果は，各モデルで 1 4 種の基本素性を使用した場合と，最適な素性の組合せを使用した場合（development test で最も F 値が良かったときの素性の組み合わせ）の両方で示す．

ベースライン (1) 式のコサイン類似度を閾値とし，閾値以上を「同等」と見なしたモデル．閾値は，最も F 値の良かった 0.84 とする．

Notdiv クラスタに分けずに $0 < cos \leq 1$ の全データで学習したモデル．

div 4.1 節の閾値でクラスタに分けて，分けたクラスタで学習したモデル．

旧 Mix モデル ([7] の手法) コサイン類似度 0.7 と 0.5 を閾値としてクラスタに分けて学習する手法と coarse-to-fine 特定法を合わせたモデル．

新 Mix モデル 4.1 節の閾値でクラスタに分けて学習する手法と coarse-to-fine 特定法を合わせた本研究のモデル (図 2)．

表 3 より，基本素性を使用した場合も，最適な素性の組合せを使用した場合も，本研究の新 Mix モデルが一番良い F 値となった．最適な素性の組合せを使用したときの，新 Mix モデルと Notdiv モデルや旧 Mix モデルの F 値に有意差があるか，ウイルコクソン符号付順位和検定で検証した．Notdiv との結果は，有意確率 $p \leq 0.037$ で，旧 Mix モデルとの結果は，有意確率 $p \leq 0.037$ で，共に有意水準 5%において有意差が存在した．

5 「推移」関係の特定

次に,「推移」関係の特定手法を説明する．文間で数値が変化してしても,「Aが１０万円となった。」という文と「Bが２０万円になった。」いう文の場合は,「推移」関係とならない．それをふまえ，本研究では係り受け情報を利用し,「数値を値として持つ名詞句」（例文の「A」や「B」）を抽出する．

図 3: 係り受け解析結果の例

5.1 「数値を値として持つ名詞句」の抽出

下記の手法で,「数値を値として持つ名詞句」を抽出する．

1. 文に出てくる日付表現を除いた数値表現のある句 (A とする，以後「数値句」と呼ぶ) を探す．
2. 「数値句」A が係る用言のある句 (B とする，以後「述句」と呼ぶ) を探す．
3. 「述句」B に係る名詞句を探し，これを「数値を値として持つ名詞句」とする．
4. 「数値を値として持つ名詞句」に係る句を抽出し (用言から係る句と日付表現は除く),「数値を値として持つ名詞句」に係る句と「数値を値として持つ名詞句」合わせて「数値を値として持つ名詞句」とする．

例えば,「６月末の携帯電話の加入台数は３４０７万７０００台となった。」という文の係り受け解析結果が図 3 のとき 「数値句」は,「３４０７万７０００台と」になり,「述句」は,「なった。」となる．そして,「数値を値として持つ名詞句」は,「加入台数は」となる．日付表現を除いた「加入台数は」に係る句は,「携帯電話の」となり,「数値を値として持つ名詞句」は合わせて,「携帯電話の加入台数は」となる．

5.2 「推移」特定の素性

推移で用いた 20 種類の素性を簡潔に説明する．「同等」の特定と同様に，2 つの文ペア (S1，S2 とする) が 1 事例になる．

- 「同等」のときに用いた素性 (11 種類)
  基本素性から，段落間類似度と文書間類似度と格助詞の一致を除いた 11 種
- 「数値を値として持つ名詞句」の名詞の類似度
- 「数値を値として持つ名詞句」の名詞の bigram, trigram 類似度
- 抽出した「数値句」において，数値が変化しているかどうか
- 数値が変化した単位
- 文に出現する用言
- ハ格，ガ格，モ格のある句に存在する名詞の類似度
- 相対表現が 2 文にあるかどうか，S1 のみにあるかどうか，S2 のみにあるかどうか

  難波ら [6] は,「推移」関係の特徴として，以下の例の太字のように，数値の相対的な差異を表す表現や数値の変動を表す表現 (以後，相対表現とよぶ) が出現することが多いと報告している．

  例 3： 日経平均株価は前日終値比 218 円**安**の...

  例 4： 97 年度末に比べ 36 万 4000 台**減**の...

  本研究は，難波が提示した相対表現 26 種に以下の 2 つの相対表現を追加した．下記の (A｜B) は，A または B どちらの表現でも可能なことを表し，$[0-9]+$ は，数値の繰り返しを表す．

  1. (前年｜昨年｜前月｜前日｜先月｜先週｜先日｜同月｜前年同月)(より｜比｜に比べ)$[0-9]+$
  2. $[0-9]+$(年｜ヶ月｜日)(連続｜ぶり)

表 4: 「推移」特定の結果

	精度	再現率	F 値
ベースライン	27.44	41.26	32.96
難波らの手法	19.85	45.96	27.73
NotUseEqResult	42.41	47.06	44.61
UseEqResult	43.13	48.51	45.67
UseMan	43.06	48.55	45.64

- 報告者の類似度
  「〜が発表した 1998 年の完全失業率は...」などの表現の〜の部分を報告者とし，それの類似度をを素性とした．報告者は,「発表する」「まとめる」に係るガ格のある句と文頭に出現する「によると」に係る句から抽出した．

5.3 「同等」特定結果の利用

「数値を値として持つ名詞句」が似ていると,「推移」である可能性は高い．しかし,「同等」関係もほとんどの場合で「数値を値として持つ名詞句」は似ている．よって，実際は「同等」である文ペアを，誤って「推移」と特定することを避けるため，本研究の「同等」特定モデルが「同等」とみなした文ペアをデータから除いて特定する．

6 「推移」特定の実験

前節で述べた「推移」の特定手法の性能を，実験により示す．2 文に数値表現が含まれていない場合と本モデルが「同等」とみなした場合を除き，55209 個の文ペアから 562 個の「推移」を特定する実験を行なった．

提案手法の性能を表 4 に示す．比較は，提案手法と以下の 4 つのモデルでおこなう．

ベースライン 「数値を値として持つ名詞句」の類似度を閾値として，2 文で数値が変化していて，2 文に相対表現が存在するときを「推移」とみなした場合．閾値は，最も F 値の良かった 0.7 とする．

難波 ([6] の手法) コサイン類似度を閾値として，2 文の単位が等しく，2 文に相対表現が存在するときを「推移」とみなした場合．閾値は，最も F 値の良かった 0.42 とする．

NotUseEqResult 「同等」結果を利用しないで，学習して特定したモデル．文ペア数は，55547 個である．

UseEqResult 本研究の「同等」モデルで特定した「同等」結果を評価データから除いて特定した本研究のモデル．

UseMan 人手で付与された正解データを用いて「同等」結果を評価データから除いて，特定したモデル．文ペア数は，55067 個である．

表 4 より，先行研究の難波らの手法よりもベースラインの方が良い F 値となっている．難波らの手法とベースラインの手法の違いから,「数値を値として持つ名詞句」の類似度と数値の変化に着目したことで，従来手法よりも良い結果になったといえる．

「同等」結果を評価データから省くことの有効性においては，NotUseEqResult と UseEqResult のモデルを比較すると，本研究のモデルの方が良い F 値となった．更に UseMan と比較すると，ほとんど変わらない F 値となった．実際は「同等」であるのに「推移」とみなしたエラーが，NotUseEqResult では 7 事例存在したが，UseEqResult では 3 事例に減っている．

また,「推移」のエラー解析の結果，複文によって構文解析エラーが起こり,「数値を値として持つ名詞句」が正しく抽出できず,「推移」と特定できなかったエラーが，69 事例あった．仮に，構文解析エラーがなく,「数値を値として持つ名詞句」が正確に抽出できていれば，精度はもう少し改善されると考えられる．

7 結論

一つのトピックについて書かれた異なる記事中の文間で同じ内容を述べているかを機械学習を用いて特定する手法を提案した．提案手法では，2 つの文の類似度でデータを複数のクラスタに分ける手法と，coarse-to-fine 特定法の 2 つの手法を組み合わせて特定し，優れた結果になることを示した．

また，異なる記事中の文間で数値が変化しているかを特定する手法を提案した．提案手法では，係り受け情報を利用して,「数値を値として持つ名詞句」を抽出し,「数値を値として持つ名詞句」の類似度などの情報を用いて「推移」を特定し，従来手法よりも優れた結果となった．

謝辞

本研究を進めるにあたり，ご指導頂いたＮＴＴコミュニケーション科学基礎研究所の平尾努氏に深く御礼申し上げます．また，ランゲージウェア社の衛藤純司氏には本研究の文書間構造解析コーパスを作成して頂き，コーパスの仕様についての貴重な意見も頂きました．深く御礼申し上げます．

参考文献

[1] 横山憲司, 難波英嗣, 奥村学. Support vector machine を用いた談話構造解析. 情報処理学会 自然言語処理研究会 NL-155, pp. 193–200, 2003.

[2] Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 368–375, 2002.

[3] Dragomir Radev. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, pp. 74–83, 2000.

[4] 衛藤純司, 奥村学. 文書横断文間関係タグ付コーパスの構築. 言語処理学会第 11 回年次大会, pp. 482–485, 2005.

[5] Zhu Zhang, Jahna Otterbacher, and Dragomir R.Radev. Learning cross-document structural relationships using boosting. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, pp. 124–130, 2003.

[6] 難波英嗣, 国政美伸, 福島志穂, 相沢輝昭, 奥村学. 文書横断文間関係を考慮した動向情報の抽出と可視化. 情報処理学会研究報告 自然言語処理研究会 NL-169, pp. 67–74, 2005.

[7] 宮部泰成, 高村大也, 奥村学. 異なる文書中の文間関係の特定. 情報処理学会研究報告 自然言語処理研究会 NL-169, pp. 35–42, 2005.

[8] 難波英嗣, 奥村学. 第 2 回 ntcir ワークショップ自動要約タスク (tsc) の結果および評価法の分析. 情報処理学会研究報告 自然言語処理研究会 NL-144, pp. 143–150, 2001.

[9] 松下光範, 加藤恒昭, 平尾努. 動向情報の要約と可視化に関するワークショップの提案. 情報処理学会研究報告 自然言語処理研究会 NL-164, pp. 89–94, 2004.