異なる文書中の文間関係の特定

宮部 泰成[†]　　高村 大也[††]　　奥村 学[††]
[†] 東京工業大学大学院 総合理工学研究科知能システム科学専攻
〒 226-8503 横浜市緑区長津田町 4259
[††] 東京工業大学 精密工学研究所
〒 226-8503 横浜市緑区長津田町 4259
[†] miyabe@lr.pi.titech.ac.jp,[††]{takamura,oku}@pi.titech.ac.jp

一つのトピックについて書かれた複数の新聞記事に対し，異なる記事中の文間が同じ内容を述べているか (同等関係) を特定する機械学習に基づく手法を提案する．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に分け，分けられたクラスに合った特徴で学習することによって，データを分けずに学習するより，優れた結果を得られることを示した．また，2 つの文の類似度があまり高くない文間ペアのクラスにおいて,「同等」の数は，クラス内の全ての文間ペア数と比べて，大変少ない．このため，2 つの文が同じ内容を述べていても,「同等」関係であると特定できないときがある．この問題を解決するために，異なる記事中の文間に存在する，同じ内容を簡潔に述べたり，詳しく述べたりする「同等と似た関係」を利用する．最初に「同等」と「同等と似た関係」を一つの粗いクラスにまとめて特定し，次に粗いクラスから「同等」のみを特定する，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ることによって，高い正解率が得られることを示した．

キーワード：文間の類似度, 文書間構造理論, 談話構造解析

Identifying a cross-document relation between sentences

Yasunari Miyabe[†]　　Hiroya Takamura [††]　　Manabu Okumura[††]
[†]Department of Computational Intelligence and Systems Science, Interdiciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
4259 Nagatsuta Midori-ku Yokohama, JAPAN, 226-8503
[††]Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259 Nagatsuta Midori-ku Yokohama, JAPAN, 226-8503
[†] miyabe@lr.pi.titech.ac.jp,[††]{takamura,oku}@pi.titech.ac.jp

We propose a machine learning based method that identifies an equivalence relation between sentences in different newspaper articles on a topic. We showed that our method, which divided the corpus into several classes by sentence similarity and learned a classifier, yielded a superior result than without dividing it. In addition, compared with the number of total sentence pairs, the number of sentence pairs in an equivalence relation is too small in a relatively less similar class. Therefore, the classifier sometimes cannot identify equivalence relations. To solve this problem, we use "relations similar to equivalence" that describe a same content more briefly or in more detail in different newspaper articles. We also propose a two-stage method that first identifies a coarse class that includes both an equivalence relation and "relations similar to equivalence", and then identifies an equivalence relation from a coarse class. We showed that high accuracy was yielded by combining these two methods.

Keywords: sentence similarity, cross-document structure theory, discourse structure analysis

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，単一文書内の構造を解析する (横山, 2003; Marcu, 2000; Marcu, 2002) 研究であったが，Radev(2000) は複数文書を対象とした異なる文書間の構造に着目し，CST 理論を構想した．また Zhang ら (2003) は異なる文書間の構造を解析する手法を提案した．異なる文書間の構造解析は，自然言語処理の複数文書要約や QA，情報抽出等の分野において，有用である．例えば，テキスト間で同じ内容が書いてあると認識できれば，複数のテキストから要約文を抽出するときに，内容が重複した冗長な要約となるのを避けることができる．

本研究は，異なる文書間の構造解析の第一段階として，異なる文書中の 2 つの文が同じ内容を述べている「同等」関係の特定を機械学習を用いて行なう．しかし,「同等」関係と一口に言っても，その中には，文間で表層的な文字列が大変類似しているクラスもあったり，文字列が全く類似せず意味だけが類似しているクラスもある．このような様々な「同等」を一様に扱って学習するより，それぞれの「同等」に合った特徴で

学習した方が良い結果が得られると考えられる．また，2 つの文の類似度があまり高くない文間ペアのクラスにおいて，クラス内の全ての文間ペア数と比べて,「同等」の数は大変少ない．そのため，2 つの文が同じ内容を述べていても,「同等」と特定できないときがある．この問題を解決するために，異なる記事中の文間に存在する，同じ内容を簡潔に述べたり，同じ内容を詳しく述べたりする,「同等と似た関係」を利用する．最初に,「同等」と「同等と似た関係」を一つの粗いクラスにまとめて特定し，次に特定した粗いクラスから「同等」のみを特定する，2 段階の特定手法を提案する．これら 2 つの手法を組み合わせることによって，優れた結果が得られることを示す．

以下，2 章で異なる文書間の構造解析について説明し，3 章で本研究で用いるコーパスの説明をする．4 章で異なる文書中の文間が「同等」かどうかを特定する手法の説明をし，5 章でその手法の実験結果を示し，考察を行なう．6 章で本研究のまとめと今後の課題について述べる．

2　文書間構造理論

ここでは，文書間構造理論について簡単に紹介する．なお，以後出てくる文とは，句点で区切られた文字列のことである．また，本研究のトピックの例としては,「1998 年ノーベル賞受賞」や「AIBO 発売」などがある．

Radev(2000) は，一つのトピックについて書かれた複数の新聞記事を集め，異なる新聞記事中の文間には様々な関係があると考え，24 個の関係を定義した．衛藤 (2004) は Radev の 24 個の関係を基に，日本語の新聞記事集合に対して 14 個の関係を定義した．14 個の関係は大きく分けて,「同一性に基づく関係」と「差異性に基づく関係」に分けられる．14 個の関係について以下に記述する．また，図 1 は,「1998 年 11 月 27 日の小渕首相の行動に関するトピック」の異なる文書中の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，国会に関する事柄の時間が経過しているので,「推移」関係が成り立つ．また，文書 A の文 2 と文書 B の文 2 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．

同一性に基づく関係

- 同等
  同じ事柄を述べたもの．全く同一の表現である場合もあれば，異なる言い回しで同じ事柄が表現されている場合もある．
- 簡略
  同じ事柄を要約したり，一部を省略したりしたもの．時系列的に前の記事を後の記事が簡略化するという関係である．
- 詳細
  同じ事柄をより詳しく述べたもの．時系列的に前の記事を後の記事が詳細化するという関係である．
- 例示
  同じ事柄を，具体的な例を挙げて説明する場合．
- 並列
  同じタイプの事柄をただ単に並べたもの．
- 参照
  同じ事柄を別の視点から述べるような場合．別の視点とは，例えば，過去を振り返る視点であったり，予測する視点であったり，理由づけをする視点であったり，さらには別の人物の視点であったもよい．別の人物の視点の特別な場合として，引用の関係もこれに含める．
- 補足
  ある事柄に関することを補足的に説明する．一連の記事の流れから見ると本筋ではないが，その事柄を何らかの意味で補うような場合である．

差異性に基づく関係

- 対照
  同じ事柄について違ったこと，対立すること，矛盾することを述べる場合である．
- 追加
  ある事柄について，前の文にはなかった新しい情報を付け加える場合である．
- 背景
  ある事柄の歴史的・非歴史的な背景を述べる．
- 推移
  ある事柄が時間の経過とともに変化する場合である．例えば，ある事件が発生し，警察が捜査し，容疑者が逮捕され，起訴され，判決が下る，というような，一連の経過が述べられる場合である．また事柄の変化には，数量に関わるものがあり，株価の変動や年平均気温の変動のように，その都度の数値が意味を持ち，数値の変動がそのまま経済や気候の変動を表すものを,「推移」に含める．
- 更新
  ある事柄の変化のうち，特に数量的に変化する場合である．災害の被害者数のように，数値の変動が被害者の数そのものの変動を表すのではなく，それまで判明していなかった正確な数値が判明しただけで，最後の数値だけが意味を持っているものである．
- 継起
  同タイプの事柄が時間を置いて生起する場合．
- 因果
  ある事柄を原因として，後続の記事でその結果が述べられるような場合．

3　コーパス

ここでは，本研究で用いたコーパスの説明をする．まずデータの説明をし，次に各関係の数を説明する．

3.1　コーパスに用いたデータ

本研究のコーパスは，テキスト自動要約タスク TSC2, 3(Text Summarization Challenge)(難波, 奥村, 2001) で使用されている新聞記事に，衛藤 (2004) が定義した関係を，文レベル，段落レベル，文書レベルで付与したものである．また文レベルの組合せは,「1 文対 1 文」の対応だけでなく,「1 文対複数文」,「複数文対 1

表 1: 各関係の数 (cos はコサイン類似度を表す)

cos	文間ペア	同等	簡略	詳細	例示	並列	参照	補足	対照	追加	背景	推移	更新	継起	因果
$(0, 0.1]$	65388	12	5	3	1	7	0	0	2	2	2	13	0	0	0
$(0.1, 0.2]$	60827	14	6	3	1	30	2	3	2	3	0	46	3	3	0
$(0.2, 0.3]$	27063	19	21	12	0	52	3	3	5	11	1	101	5	6	0
$(0.3, 0.4]$	10524	21	18	9	0	37	8	1	5	3	1	189	12	9	2
$(0.4, 0.5]$	3881	21	15	8	0	17	8	0	2	8	4	219	6	5	1
$(0.5, 0.6]$	1399	42	12	11	0	3	6	1	1	4	0	197	4	7	1
$(0.6, 0.7]$	501	57	13	5	0	0	1	0	3	5	1	108	2	4	0
$(0.7, 0.8]$	163	37	6	7	0	2	0	0	1	2	0	50	1	1	0
$(0.8, 0.9]$	85	42	4	2	0	1	0	0	0	2	0	22	0	0	0
$(0.9, 1]$	345	342	0	2	0	0	0	0	0	0	0	0	0	0	0
合計	170176	607	100	62	2	149	24	8	21	40	9	945	33	35	4

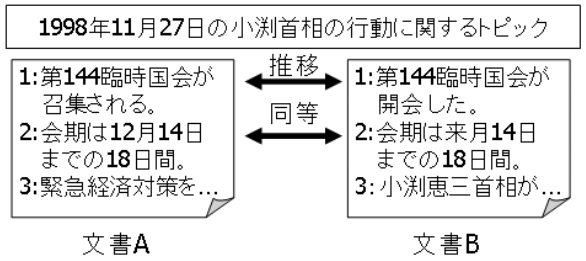

図 1: 文書間構造理論の例

文」,「複数文対複数文」で対応している組合せもある．本研究では 1 文対 1 文で対応している組合せのみを扱うことにした．可能な組み合わせ数 (文ペア数) は 170,176 個である．トピック数は 85 個で，1 トピックにつき平均 14 文書，1 文書につき平均 9.5 文が確認された．

3.2 各関係の数

各関係の数を 2 つの文のコサイン類似度別で調査した (表 1)．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．ただし，簡単のため以降では S1，S2 を明示せずに，コサイン類似度を cos で表すことにする．

4 「同等」関係の特定

ここでは,「同等」関係の特定モデルについて説明する.「同等」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．それをふまえ本研究では，規則の獲得に機械学習手法を用いることにした.「同等」関係の特定は，文ペアが与えられたとき，そのペアが「同等」か否かの特定を行なう 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)(Vapnik, 1998) を学習器として使用する．ただし，文ペア (2 つの文を S1，S2 とする) は，S1 を含んだ新聞記事の掲載日が，S2 を含んだ記事の掲載日より，時間的に前となるように与える．

4.1 類似度別でのクラス分け

表 1 より,「同等」は類似度が高い「同等」だけでなく，類似度が低い「同等」も多数存在している．本研究で使用する SVM は教師付き学習器であり，類似度が高い「同等」と類似度が低い「同等」を一緒に学習することは，類似度が高い「同等」，低い「同等」どちらにとっても学習時のノイズになると考えられる．それをふまえ，本研究では,「同等」関係を類似度別のクラスに分けて学習を行なうことにする．

クラスの分け方は，3 段階とする．まず，文間ペアのうち半分以上の個数が「同等」関係となっているクラスとそれ以外のクラスで分ける．すなわち,「表層的に大変似ているクラス」($0.7 < cos \leq 1$) と「それ以外のクラス」($0 < cos \leq 0.7$) に分ける．更に「それ以外のクラス」を類似度が半分に満たないかどうかで「意味的にしか類似していないクラス」($0 < cos \leq 0.5$) と「文字列がある程度類似しているクラス」($0.5 < cos \leq 0.7$) で分けた．以下に各クラスの「同等」関係の例を示す．

表層的に大変類似しているクラス

S1 また、国連安保理も今月半ば、ラディン氏を引き渡さない場合、経済制裁を科すと通告している。

S2 国連安保理は１４日を期限にラディン氏を引き渡さなかった場合、経済制裁を発動すると通告している。

文字列がある程度類似しているクラス

S1 修理は来年 9 月にも終える予定。

S2 来年 9 月には修復工事が完了する予定。

意味的にしか類似していないクラス

S1 武装解除を「恥ずべき降伏」と見なしてきたＩＲＡが、武装解除にかかわる動きを見せたのは初めて。

S2 対立するプロテスタント側の難局打開に向けた呼びかけに応じたもので、ＩＲＡが公式に武装解除

問題で交渉の意思を示したのは初めて。

4.2 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「同等」か否かの特定を行なうので，文ペアが一事例になり，そこから素性値を算出することになる．

また素性ベクトルの各要素は 0 か 1 の 2 値となるように定義した．2 値とならない素性は，[0,1] の値に正規化し，正規化した値が [0,1] を 10 分割した区間 $[0.0, 0.1), [0.1, 0.2), \cdots, [0.9, 1.0]$ のどこに属するかを表す 10 次元のベクトルに変換した．例えば，ある素性の値が，0.65 であれば，0000001000 に変換される．また形態素解析は，ChaSen[1]，文節の抽出は CaboCha[2]を用いた．この節の 17 種の素性を基本素性とする．

類似度

2 つの文の類似度に，コサイン類似度，bigram の類似度，trigram の類似度，文節の類似度，段落間類似度，文書間類似度を使用する．

bigram，trigram の類似度は，(1) 式の U1，U2 を形態素の bigram，trigram の頻度ベクトルに置き換えた場合とし，文節の類似度は，U1，U2 を文節の頻度ベクトルに置き換えた場合とする．

段落間類似度と文書間類似度は，(1) 式の U1，U2 を，2 つの文 (S1，S2) を含んだ段落間，文書間の単語 (名詞，動詞，形容詞) の頻度ベクトルに置き換えた場合とする．

文の文字数

文字数が少ないと，2 つの文が表層的に似ていても，類似度は低いことがある．このことから各文の文字数を素性として使用する．各文 (S1，S2) の文字数 ($len(S1)$，$len(S2)$) を以下の (2) 式で定義する:

$$len(S) = \frac{str(S)}{TopicMax(S)}. \quad (2)$$

str(S) は，文 S の文字数，TopicMax(S) は，S を含んだトピック内にある全ての文において，最も大きかった文字数とする．

2 文の形態素数，文節数の関係

2 つの文 (S1,S2) の形態素数，文節数が似た数であればあるほど，1 に近い値を返す素性 ($relM$，$relC$) を，以下の (3)，(4) 式で定義する:

$$relM = 1 - \frac{|CountM(S1) - CountM(S2)|}{Max(CountM(S1), CountM(S2))} \quad (3)$$

$$relC = 1 - \frac{|CountC(S1) - CountC(S2)|}{Max(CountC(S1), CountC(S2))}. \quad (4)$$

ここで，$CountM(S1)$ は，S1 の形態素の個数とし，$CountC(S1)$ は，S1 の文節の個数とし，$Max(a, b)$ は a，b のうち大きい方の値を返す関数とする．

日付の一致

2 つの文 (S1，S2) に対し，S1 を含んだ新聞記事の掲載日と S2 を含んだ新聞記事の掲載日が一致すれば 1，しなかった場合に 0 を取る 2 値の素性を定義する．

文の位置

各文 (S1，S2) が新聞記事のどこにあるかを表す素性 ($Pos(S1)$，$Pos(S2)$) を以下の (5) 式で定義する:

$$Pos(S) = \frac{BDocStr(S)}{Docstr(S)}. \quad (5)$$

$Docstr(S)$ は S を含んだ文書の総文字数，$BDocStr(S)$ は，S を含んだ文書の先頭から S が出てくるまでの文字数とする．

用言の意味

日本語語彙大系 (池原ら, 1999) を用い，各文 (S1，S2) に出現する全ての用言に対し，特定の意味カテゴリがあれば 1 を，なければ 0 を返す素性を定義する．日本語語彙大系における用言の意味カテゴリは 36 種類である．

名詞の意味

日本語語彙大系 (池原ら, 1999) を用い，各文 (S1，S2) に出現する全ての名詞に対し，特定の意味カテゴリがあれば 1 を，なければ 0 を返す素性を定義する．日本語語彙大系における名詞の意味カテゴリは 2715 種類である．

接続詞 (横山, 2003)

各文 (S1，S2) の文頭に横山 (2003) で使用された特定の接続詞があれば 1，なければ 0 をとる 2 値の素性を定義する．接続詞は 55 種類である．

文末表現 (横山, 2003)

横山 (2003) が使用した文末文字列と文末表現の対応規則を用いて，各文に特定の文末表現があれば 1，なければ 0 をとる 2 値の素性を定義する．文末表現は，過去，現在，断定，存在，推量，様態，問掛，判断，可能，理由，要望，叙述，義務，意見，継続，使役，伝聞，状態の 18 種類を用いた．

固有表現

各文 (S1，S2) に特定の種類の固有表現が存在すれば 1，なければ 0 を返す素性と，2 文の特定の種類の固有表現が一致すれば 1，一致しなければ 0 を返す素性を定義する．固有表現の抽出には，bar[3]を用いた．また，ここでの固有表現の種類は，bar が抽出する ARTIFACT，DATE，ORGANIZATION，MONEY，LOCATION，PERCENT，PERSON，TIME の 8 種類である．

助詞

特定の固有表現が係る 2 つの文の助詞が一致すれば 1，一致しなければ 0 を取る 2 値の素性を定義する．助詞は「格助詞一般」と係助詞 (「は」,「も」) の計 13 種類とする．

4.3 類似度の高くない「同等」の特定

表 1 より $0.5 < cos \leq 0.7$，$0 < cos \leq 0.5$ のクラスは，文間ペア数と比べて,「同等」の数が少ない．そのため，2 つの文が同じ内容を述べていても,「同等」と特定できないときがある．この問題を解決するために，異なる文書中の文間に存在する，同じ内容を簡潔

[1] http://chasen.naist.jp/hiki/Chasen/

[2] http://chasen.naist.jp/~taku/software/cabocha/

[3] http://chasen.naist.jp/~masayu-a/p/bar/

に述べる「簡略」と同じ内容を詳しく述べる「詳細」の 2 つの関係 (「同等と似た関係」) を利用する．最初に「同等」と「同等と似た関係」を一つの粗いクラス (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，2 段階の特定を行なう (図 2)．

このように，最初に粗く特定し，候補を絞った中から精細に特定する手法は，coarse-to-fine といい，画像認識のテンプレートの探索において，計算時間を削減する手法として使用されている (Vanderburg, 1977; Rosenfeld, 1977)．以下では，この手法を「coarse-to-fine 特定法」と呼ぶことにする．

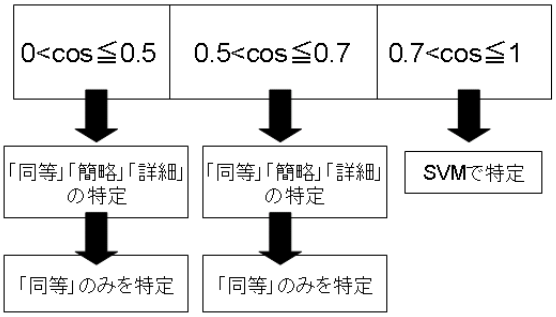

図 2: 本研究の手法

4.4 fine クラスの特定における素性

coarse クラスで特定した,「同等」と「同等と似た関係」から,「同等」のみを特定する素性を以下で説明する．

文節の個数の関係

4.2 節で使用したものと同様

固有表現

4.2 節で使用したものと同様

主動詞の一致 (畑山, 2002)

畑山 (2002) による主動詞抽出規則を用いて，2 つの文の主動詞を抽出し，主動詞の文字列が一致，主動詞の意味が一致，動詞無し，それぞれに当てはまれば 1 を，当てはまらなければ 0 を返す素性を定義する．

数の一致と単位 (助数詞) の一致 (福島, 2005)

2 文の「数」と「単位」それぞれに対し，一致すれば 1 を，しなければ 0 を返す素性，福島 (2005) が使用した相対表現を表す特定の単位があれば 1 を，なければ 0 を返す素性を定義する．

2 文間で異なった単語の意味カテゴリ

日本語語彙大系 (池原ら, 1999) を用いて 2 文間で異なった単語の特定の意味カテゴリが一致すれば 1 を，しなければ 0 を返す素性を定義する．以下の S1，S2 の 2 文を例に説明する．

S1 修理は来年 9 月にも終える予定．

S2 来年 9 月には修復が完了する予定．

上記の 2 文間で異なった単語は，修理，終える，修復，完了するの 4 つである.「修理」と「修復」,「終える」と「完了する」それぞれの意味カテゴリが共に一致しているので，これらの意味カテゴリを素性に加える．

5 実験

提案手法の性能を, 実験により示す．まず各クラス毎での結果を示し，次に各クラスの結果を統合した全データ ($0 < cos \leq 1$) の結果を示す．結果は，クラスに分けて学習し coarse-to-fine 特定法を使用する場合 (divctf) と使用しない場合 (divNctf)，クラスに分けずに $0 < cos \leq 1$ の全データで学習し，coarse-to-fine 特定法を使用する場合 (Ndivctf) と使用しない場合 (NdivNctf) の 4 つのモデルで示す．また coarse-to-fine 特定法を使用しないときの SVM のパラメータは，2 次の多項式カーネルを用い，ソフトマージンパラメータは 10 を用いた．coarse-to-fine 特定法を使用する場合の SVM のパラメータは，coarse クラスの特定では，最も多くの数を再現できた線形カーネルのソフトマージンパラメータ 1 を用い，fine クラスの特定では，2 次の多項式カーネルのソフトマージンパラメータ 10 を用いた．評価は 5 分割交差検定で行なった．

5.1 表層的に大変類似しているクラス ($0.7 < cos \leq 1$) の結果

$0.7 < cos \leq 1$ のクラスの結果を表 2 に示す．比較は，4 つのモデルが 4.2 節の基本素性を使用した場合と効果的な素性を使用した場合両方で行なう．効果的な素性とは，次の手順によって求めた素性とする．17 種の基本素性から 1 つの素性を省いて残りの 16 種の素性を使用して実験し，そのときの F 値が基本素性を使用した場合より高ければ，省いた素性を効果的でない素性とした．そして，効果的でない素性の全ての組み合わせを求め，基本素性からそれらを省いた素性で実験し，最も F 値が高かったときの素性を効果的な素性とした．各モデルの効果的な素性を以下に示す．ただし，簡単のため，基本素性から効果的でない素性 a，b を省いた素性を，[基本素性 - (a+b)] と表す．

- divctf [基本素性 - 名詞の意味]
- divcNtf [基本素性 - (名詞の意味+文の位置+文書間類似度)]
- Ndivctf [基本素性 - 日付の一致]
- NdivcNtf [基本素性 - 文書間類似度]

表 2 よりいずれの場合も coarse-to-fine 特定法を使用しない方が F 値は良い．これは，$0.7 < cos \leq 1$ の「簡略」や「詳細」の数が少ないため coarse-to-fine 特定法を使用する必要がないことが原因と考えられる．また，クラスに分けて学習，分けないで学習の比較については，基本素性を使用した場合は，クラスに分けない場合の方が F 値は良いが，効果的な素性を使用した場合は，分けた方が良くなった．すなわち，データを分けて学習することによって，そのデータに合った最適な素性を選択し精度が良くなるといえる．

5.2 文字列がある程度類似しているクラス ($0.5 < cos \leq 0.7$) の結果

$0.5 < cos \leq 0.7$ のクラスの結果を表 3 に示す．比較は，5.1 節と同様に，4 つのモデルで 4.2 節の基本素

表 2: $0.7 < cos \leq 1$ の結果

基本素性			
	精度	再現率	F 値
divctf	90.20	83.91	86.94
divNctf	87.07	91.64	89.30
Ndivctf	94.27	79.04	85.99
NdivNctf	94.30	89.91	92.05
効果的な素性			
divctf	90.51	85.53	87.95
divNctf	93.41	92.32	92.86
Ndivctf	95.08	79.19	86.41
NdivNctf	94.75	90.34	92.49

性を使用した場合と効果的な素性を使用した場合両方で行なう．各モデルの効果的な素性を以下に示す．

- divctf [基本素性 - 文の位置]
- divNctf [基本素性 - (文の位置+文節の個数の関係+助詞+文末表現+文の文字数)]
- Ndivctf [素性素性 - 日付の一致]
- divNctf [基本素性 - 文書間類似度]

表 3 では，基本素性を使用した場合も，効果的な素性を使用した場合も，coarse-to-fine 特定法を適用した方が F 値は大幅に良くなっている．すなわち，$0.5 < cos \leq 0.7$ のクラスのように「同等と似た関係」が存在している場合，最初に「同等」と「同等と似た関係」を一つの粗いクラスにまとめて特定し，次に粗いクラスから「同等」のみの特定を行なうことが望ましいといえる．

表 3: $0.5 < cos \leq 0.7$ の結果

基本素性			
	精度	再現率	F 値
divctf	52.22	17.91	26.67
divNctf	10.11	1.05	1.90
Ndivctf	37.42	9.13	14.68
NdivNctf	28.09	6.90	11.08
効果的な素性			
divctf	48.56	19.33	27.65
divNctf	42.00	6.13	10.70
Ndivctf	48.00	9.05	15.23
NdivNctf	38.33	8.16	13.45

5.3 意味的にしか類似していないクラス $(0 < cos \leq 0.5)$ の結果

$0 < cos \leq 0.5$ のクラスの結果は，どのモデルの場合も精度が無し (全ての文間ペアに対し「同等」と特定しなかった) となり，再現率が 0 となった．現在のモデルでは，このクラスは特定できない．原因を 5.5 節のエラー解析で述べる．

5.4 各クラスの結果を統合した全データ $(0 < cos \leq 1)$ の結果

各クラスの結果を統合した $0 < cos \leq 1$ の全データについての結果を表 4 に示す．表は効果的な素性での結果であり，Mix モデルとは，$0.5 < cos \leq 0.7$, $0 < cos \leq 0.5$ のクラスで coarse-to-fine 特定法を使用し，$0.7 < cos \leq 1$ クラスでは使用しない本研究の手法 (図 2) である．また本研究のベースラインは，コサイン類似度 0.8 以上を「同等」と見なしたときであった．

表 4 より，クラスに分けた方がどの場合も F 値は良くなった．すなわち，各クラスに分けて，各クラスに合った素性を選択し学習する方がクラスに分けないで学習するより，最適な素性選択ができ，F 値が良くなるといえる．また本研究の手法である，Mix モデルの F 値が最も良かった．Mix モデルによって，$0.7 < cos \leq 1$ の表層的に類似している「同等」は，NdivNctf の手法より多く再現できるようになり，コサイン類似度がそれほど高くない $0.5 < cos \leq 0.7$ の「同等」も，他の手法より多く再現できるようになった (表 5)．

表 4: 統合結果

	精度	再現率	F 値
ベースライン	89.08	62.12	73.20
divctf	88.10	61.66	72.54
divNctf	92.51	64.20	75.80
Ndivctf	91.35	55.98	69.42
NdivNctf	92.72	63.41	75.31
Mix モデル	91.03	66.25	76.69

表 5: 類似度別再現数

	Mix モデル	divNctf	NdivNctf
$0.5 < cos \leq 0.6$	5	2	0
$0.6 < cos \leq 0.7$	14	4	8
$0.7 < cos \leq 0.8$	12	12	10
$0.8 < cos \leq 0.9$	36	36	33
$0.9 < cos \leq 1$	342	342	341

5.5 エラー解析

本手法の Mix モデルによる，エラーの解析を各クラス毎に行なう．

$0.7 < cos \leq 1$ のクラス

このクラスでは，bigram の類似度や文節類似度の素性が効果的であるので，単語が連続している場合に誤って「同等」と特定し，逆に単語が連続せず，順番が入れ替わっている場合に「同等」と特定できなかった．

- 誤って「同等」と特定した例

  S1 藤波被告は判決後、弁護人を通じて「事件全体を私が行ったことにされた。信じられない結果です」とする手記を発表した。

  S2 藤波被告は判決後、弁護人を通じて「事件全体を私が行ったことにされた。信じられな

い結果です」などとする手記を発表し、死刑判決に強い不満の気持ちをあらわにした。

- 「同等」と特定できない例

  S1 ウィンドウズ９８の日本語版は、アップグレード版が１万３８００円、通常版が２万４８００円の予定。

  S2 推定小売価格は通常版が２万４８００円、アップグレード版が１万３８００円。

$0.5 < cos \leq 0.7$ **のクラス**

coarse クラスと fine クラスのそれぞれでのエラーについて述べる.

- coarse クラスのエラー
  coarse クラスの「同等」は 99 個あるが，ここで特定できているのは 35 個である．ここで「同等」と特定できなかったエラーは、主に文節類似度が低く，省略によって固有表現が片方の文にしかなく一致しない場合であった．以下の例の文 2 では,「関大」が固有表現となるが，文 1 ではない (同一パラグラフ内にはある).

  S1 一方、本来は不合格だった百四人が合格になっており、これについては合格扱いとした。

  S2 また、本来は不合格だが、既に関大に入学している学生については、合格扱いとする措置を取った。

- fine クラスのエラー
  このクラスの「同等」と特定できなかったエラーは主に主動詞が一致せず，固有表現の不一致が見られた場合であった．また，逆の場合だと誤って「同等」と特定してしまった.

  1. 誤って「同等」と特定した例
     下記の例では,「実施する」という主動詞が一致し，固有表現の LOCATION、ORAGANIZATION、ARTIFACT が一致しているため,「同等」と特定してしまった.

     S1 国際系新電電の国際デジタル通信（ＩＤＣ）の 買収をめぐって、 英通信大手のケーブル・アンド・ワイヤレス（C＆W）が、連休明けの五月上旬にも株式公開買い付け（ＴＯＢ）を実施することが、二十八日明らかになった。

     S2 英大手通信会社ケーブル・アンド・ワイヤレス（C＆W）は六日、国際デジタル通信（ＩＤＣ）の株式を一株（額面五万円）あたり十万七千三百七十二円で買収する株式公開買い付け（ＴＯＢ）を実施すると発表した。

  2. 「同等」と特定できなかった例
     下記の例では,「取る」という主動詞が抽出できず主動詞が一致せず,「アップル社」と「米アップル社」の固有表現が一致しなかったため,「同等」と特定できなかった.

     S1 一方、アップル社は「ソーテックの新製品が仮処分決定に違反しないか慎重に検討した上、ｉＭａｃを模倣していると判断した場合、直ちに必要な裁判上の措置を取る」としている。

     S2 色違いの新製品に対して米アップルの日本法人は、「新製品への判断はこれから行うが、仮処分決定に違反する場合は、必要な裁判上の措置を取る」とのコメントを発表し、新たな法的措置を検討する姿勢を見せている。

$0 < cos \leq 0.5$ **のクラス**

このクラスでは，1 つも特定できなかった．以下の例のように，このクラスの「同等」の特定は，照応，省略表現や文の主要情報の抽出や知識の獲得などが必要である.

- 主語が省略されている (又は前文や後の文に出てくる) 場合
  S1:東海道・山陽では五代目で，アヒルのくちばしに似た車両先頭部が特徴。
  S2:アヒルのくちばしに似た独特の先頭形状が人気を呼びそう。
- 文の主要な情報が等しく,「同等」となる場合
  S1:１９８１年に肉親捜しの訪日調査が始まって以来，最少の訪日人数となった。
  S2:総勢２８人で過去最少となった。
- 背景知識が必要な言い換えの場合
  S1:武装解除を「恥ずべき降伏」と見なしてきたＩＲＡが，武装解除にかかわる動きを見せたのは初めて。
  S2:対立するプロテスタント側の難局打開に向けた呼びかけに応じたもので，ＩＲＡが公式に武装解除問題で交渉の意思を示したのは初めて。

6 結論と今後の課題

一つのトピックについて書かれた複数の新聞記事に対し，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 つの文の類似度でクラスに分け，各クラス毎で学習し特定した．各クラスに分けて学習することによって，クラスに分けないで学習するより優れた結果を得られることを示した．また,「同等と似た関係」が存在するクラスでは，最初に,「同等」と「同等と似た関係」を一つの粗いクラスにまとめて特定し，次に粗いクラスの中から「同等」のみを特定する 2 段階の特定を行なう手法も提案し，この手法とクラスに分ける手法を組み合わせることによって，優れた結果が得られることを示した.

今後の課題としては，以下のものが考えられる.

- coarse クラスのモデルの改善
  coarse クラスでの「同等」と「同等と似た関係」の特定において,「同等」と特定できないエラーが多い．特に，文の省略表現が固有表現である場合が多く，そのため「同等」と特定できない．よって，前後の文の固有表現を素性として試す必要がある.

- fine クラスのモデルの改善
  fine クラスでの「同等」と特定できなかったエラーは，主に，表記揺れのため固有表現が一致しない場合と主動詞が一致しない場合である．よって，表記揺れの固有表現に対し緩和すること．また主動詞が抽出できてない文に対し抽出ルールを作る必要がある．
- 意味的にしか類似していないクラス ($0 < cos \leq 0.5$) のモデルの作成
  現在では，$0 < cos \leq 0.5$ のクラスの「同等」を特定できていない．エラー解析でも述べたように，このクラスの「同等」の特定は，照応，省略の解決や知識の獲得などが必要である．それらを解決後，新たなモデルを作成することが必要である．

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．

References

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.

Daniel Marcu. The Rhetorical parsing of unrestricted texts A surface-based approach. *Computational Linguistics*, Vol. 26, No. 3, pp. 395–448, 2000.

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.

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*, 2000.

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.

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，2005, pp. 482–485, 2005.

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，NL-144，pp. 143–150，2001.

Vladimir N. Vapnik. Statistical Learning Theory. A Wiley-Interscience Publication，1998.

G. J. Vanderburg and A. Rosenfeld. Two-stage template matching. *IEEE transactions on computers*, Vol. 26, No.4, pp. 384–393, 1977.

A. Rosenfeld and G. J. Vanderburg. Coarse-Fine Template Matching. *IEEE transactions Systems, Man, and Cybernetics*, Vol. 7, pp. 104–107, 1977.

平尾 努，賀沢 秀人，磯崎 秀樹，前田 英作，松本 裕治．機械学習による複数文書からの重要文抽出. 自然言語処理 Vol. 10，No. 1 pp. 81–108, 2003.

池原 悟，宮崎 正弘，白井 諭，横尾 昭男，中岩 浩巳，小倉 健太郎，大山 芳史，林 良彦．日本語語彙大系．岩波書店．1999.

畑山 満美子，松尾 義博，白井 諭．重要語句抽出による新聞記事自動要約. 自然言語処理 Vol. 9，No. 4 pp. 55–73, 2002.

福島 志穂．文書間構造解析コーパスの分析．広島市立大学 情報科学部 知能情報システム工学科 自然言語処理学講座 平成 16 年度卒業論文.