社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－NL－166 (13)
2005／3／11

依存構造を用いたテキスト間の対応箇所の同定

伊達 貴裕　　山村 毅

愛知県立大学 〒480-1198 愛知県愛知郡長久手町大字熊張字茨ヶ廻間 1522-3

あらまし 近年，インターネットを通じて電子化されたテキストが増加している．それらのテキストの中には同様の内容のものがあるため，何度も無駄に閲覧してしまう問題がある．本研究では，複数テキストを対象とした要約においてのテキスト間の対応箇所の同定を行うため，対応文のデータベースの作成および調査を行った上で，対応する文，および文中の対応箇所を文の依存構造を用いて同定する手法を提案する．

Identification of the correspondence part between texts based on dependence structure

Takahiro Date and Tsuyoshi Yamamura

Aichi Prefectural University. 1522-3 Ibaragabasama, Kumabari, Nagakute-cho, Aichi-gun, Aichi, 480-1198, JAPAN

abstract Recently, electronic text increases through the Internet. There is a problem that a user reads similar contents many times in those texts. In this paper, first of all, we built a databese of the sentence correspondence and developed a method to identify the correspondence between texts in multidocument summarization. In Addition, we propose a technique for identifying the corresponding sentence and the corresponding phrase in that sentences by using the dependence structure of the sentence.

1 はじめに

近年，日々インターネットを通じて電子化されたテキストが増加している．このため，読み手が同様の内容を含むテキストを閲覧してしまうことも少なくない．情報を取得するにあたって同様の内容を何度も閲覧することは時間の大きな損失となるだろう．このような問題を避けるために複数テキストを対象とした要約が大きな助けになる．一般に複数テキストを対象とした要約の処理はいくつかのポイントに分けることができる [1]．その中で，複数のテキストから抽出した内容を要約する際，内容が重複して冗長な要約にならないように，同様の内容である（対応する）箇所の判断が必要となる．

2 文書間の照合としては，対訳コーパスを用いたテキストアライメントに関する研究が多く行われている [2] ～ [5]．これらは，有限長文字列間の照合をとるという意味では基本的に本研究と同様の問題を扱っているが，対訳コーパスを用いたテキストアライメントでは，文間の照合から語や句の対訳表現を得ることを目的としている．2 つのテキストは異なる言語で書かれているが，同じことが書かれていることが保証されている．

一方，要約を目的とした対応文や対応箇所の同定は，対訳コーパスのように 2 つのテキスト間には同じことが書かれている保証がなく，対応文の中に対応先のない異なる情報を含む場合が多い．

高橋ら [6] は，質問文と情報源の構文的照合を目的として Tree Kernel を拡張した構文木間の類似度評価方法を提案し，対応ノード間の類似度の定量化やノードの飛び越えを許す照合を実現している．しかし，構文的類似度が高いからと言って意味的に含意しているとは限らず，また部分的な一致を発見しているので，必ずしも同様の意味を持った文が見つかるわけではない．

本研究では，複数テキストを対象とした要約のための対応付けを目標としている．そのため対応文を同様の意味を持つ文として定義する．また，要約の手法として，文もしくは句単位で重要度判定をして抄録することを想定しているので，対応箇所を対応文間で意味的に対応関係があり，かつ表層的にその箇所を置き換えても文の構造を壊すことのない箇所であるとして定義するものとする．これらの対応の同定を実現するために文の依存構造を考慮することで文の構造，意味的な繋がりを意識した手法を提案する．

文の組合せ	組合せ数	割合
1 対 1	369	(72.1 %)
1 対 2	79	(15.4 %)
1 対 3	6	(1.2 %)
1 対 4	2	(0.4 %)
2 対 2	33	(6.4 %)
2 対 3	9	(1.8 %)
2 対 4	2	(0.4 %)
3 対 3	11	(2.1 %)
3 対 4	1	(0.2 %)
計	512	

表 1: 対応文の組合せパターンの分類

2 対応データベース作成，対応文の調査

同様の意味を持った同言語間の対応文についてのコーパスが一般に利用されるものとして存在しないので，本研究では，まず同トピックの記事を収集し，対応関係を表すデータベースを作成した．このデータベースより，対応関係についてどのような特徴があるか調査を行った．

2.1 データの収集

朝日，日経，毎日，読売新聞社の同トピックテキストの記事を WWW で収集した．記事数は 191 で，1271 文である．このデータを人手で解析して対応箇所の同定を行い，対応文の組合せのパターン，対応文の情報量の違い等の調査を行った．

2.2 対応文の組合せ

テキスト間で対応関係にある文の組合せについて調べた結果を表 1 に示す．

表より「1 対 1」以外の文の組合せが約 3 割ほどを占めている．冗長でない複数テキストを対象とした要約を目指していくには，「1 対多」や「多対多」の組合せも考慮した対応付けも考慮していく必要がある．

2.3 対応文の情報量の違い

対応関係にある文中の情報がすべて対応先があるとは限らない．異なる情報を文中に含む場合も考えられる．そこで，「情報量が等しいとみなせる対応関係」，「一方に異なる情報が含まれると考えられる対応関係」，「双方に異なる情報が含まれる対応関係」の 3 つに分類してそれらがどのような割合で存在しているかを調査した．結果を表 2 に示す．これより，対応関係にある組の過半数が双方に異なる情報を含んでおり，一方のみに異なる情報を含む場合とあわせて全体の約 90 %を占めている．異なる情報の量にもよるが，これは単純な文間の距離計算で対応文を見つけることが困難なケースが多く存在していることを示している．

情報量が同じ	49 組
一方に異なる情報を含む	180 組
双方に異なる情報を含む	283 組
計	512 組

表 2: 対応文の情報量の違い

単語のバリエーション		総数	割合
同一単語		890	75.6 %
異表記	類義語	125	10.6 %
	省略	65	5.5 %
	時間	40	3.4 %
	人物，場所等の呼称	22	1.9 %
	照応	17	1.4 %
	数量	12	1.0 %
	揺れ（カタカナ語）	6	0.5 %

表 3: 単語レベルでの対応のバリエーション

2.4 対応箇所を構成する単語のバリエーション

対応箇所をより詳しく調べていくと，文のレベルで対応しているものからより細かい句のレベル，最終的には単語のレベルで対応しているものまで存在する．このうち，単語レベルで対応しているものは，同じ単語が，それぞれの対応箇所で使われているものがほとんどであるが，中には異なった単語が使われていることがある．このバリエーションにどのようなものがあるかを調べたのが表 3 である．

同一単語が 75.6 %と大半を占めているが，類義語などの異表記も存在する．より正しく対応箇所を同定するには，これらも考慮する必要がある．

3 依存構造を用いた対応箇所の同定

本研究では，同様の意味を持つ文を対応文と定義し，対応関係にある箇所を意味的に対応関係を持ち，かつ表層的に置き換えが可能である箇所としてその同定を行う．これを実現するために，文の依存構造を求めて対応をとることにする．

この際，文の中心要素となる文末の文節が一致する文は大まかな枠組みが一致していると考えられるので，文末の文節が一致する文を探索し，それらを

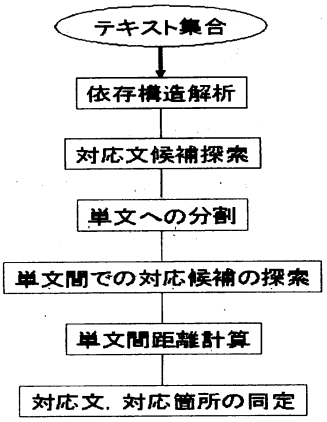

図 1: 対応文，対応箇所同定の流れ図

対応文候補として扱うものとする．

図 1 に対応箇所同定の処理の流れを示す．この図で，依存構造の解析には係り受け解析器 CaboCha [7] を利用している．以下，まず 2.4 節で述べた単語のバリエーションに対応するために，異表記単語の同定について述べ，次いで，対応文候補探索以下の各処理について述べる．

3.1 異表記単語の同定

同様の内容を示す異表記単語の同定をするために EDR 概念辞書 [8] を利用した 2 つの単語の意味的な類似性を定量化し，閾値処理をすることで類義語，同義語の判断を行う．また，省略や表記の揺れに対しては，CaboCha による解析で得られる固有表現ラベルを用いて単語の同定を行う．

3.1.1 概念辞書を用いた類義語，同義語の判別

EDR 概念辞書は，語の上位下位関係を中心において木構造の内部のノードにも語が対応するシソーラスである．ただし，シソーラスは実際に単純な木構造とはならず，あるノード（語）に複数の親ノード（上位語）が存在したり，単語の語義が複数ある場合が存在するために類似度が一意に決まらないことがある．

深谷ら [9] は類義語，同義語の判定で，単語間の類似度を単語 i，j の語義 x，y のルートからの距離（深さ）を d_{ix}，d_{jy}，それらの共通の上位語（ノード）の深さを d_{xy} とし，以下の式のように類似度の最大値をとり，経験的に $R_{ij} > 0.85$ となる単語同士を類義語，同義語として良好な結果を得ている．

$$R_{ij} = \max_{xy}\left(\frac{d_{xy}\times 2}{d_{ix}+d_{jy}}\right) \tag{1}$$

本研究もこれに習い，類義語，同義語の判定を行う．

3.1.2 固有表現を利用した異表記単語の同定

省略や表記の揺れ等にも対応できるように単語の固有表現を利用する．固有表現は固有名詞的表現，時間的表現，数値的表現に分けられる．本研究では，固有名詞的表現となる単語を対象とし，同様の属性である単語を文字列の一致度によって判定する．単語 w_x，w_y を構成する文字集合を $w_x = \{x_1, x_2, \cdots, x_n\}$，$w_y = \{y_1, y_2, \cdots, y_m\}$ とするとして，以下の式で求められる単語 w_x と w_y の包含関係の値がそれぞれ 0.5 以上を満たすときに同一の単語であると判断する．ここで $N(w)$ とは，単語 w における文字集合の要素数である．

$$sim(w_x|w_y) = \frac{N(w_x \cap w_y)}{N(w_y)} \tag{2}$$

$$sim(w_y|w_x) = \frac{N(w_x \cap w_y)}{N(w_x)} \tag{3}$$

なお，時間的表記に関しては，対応単語の調査時に西暦を示す場合に 4 桁で表記されているものと 2 桁で表記されているもの（例えば「2000 年」と「00 年」）が多くあった．これらに関しては，「年」の直前の 2 桁の数字の一致をみて同定の判断を行っている．

3.2 対応文候補探索

日本語文では基本的に文の中心的成分が文末の文節に現れる．そこで，依存構造解析を行って文末の文節が同等なものだけを対応文の候補として扱う．より具体的には，ある 2 つの文について，文末文節内の助詞，助動詞などの機能語を除いた単語について一致するものが 1 つ以上あれば，それら 2 つの文を対応文の組の候補として扱う．これは，文末が体言止めになる文や，文末表現が「発表した」や「明らかにした」等と意味的に一致するが，文末文節が部分的にしか一致しない文が存在するからである．ただし，これによって対応文の候補としてふさわしくないものも候補として挙げられてしまうことがあるが，以下で説明する対応文を決定する過程で対応文としてふさわしくないものは概ね除去可能である．

3.3 単文への分割

部分的に対応する箇所が存在する場合に対処するため，対応文の組候補となった各文を単文に分割する．依存構造解析の結果を用いて，文の係り元（葉）から辿って用言を含む文節，もしくは文末までを 1 つの単文として分割を行う．

例えば，「部長は佐藤さんに報告書を提出することを命じた。」という文ならば，図 2 のような係り受け関係となり，用言である「する」で分割を行うことに

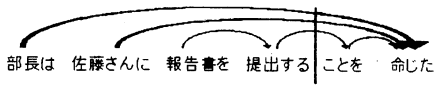

図 2: 係り受け関係と分割境界

なる．この文は，「部長は命じた」「佐藤さんに命じた」「報告書を提出する」「ことを命じた」の4つの単文に分割される．

3.4 単文間での対応候補の探索

対応文の組の候補の各文を単文に分割したあと，それら単文間で対応している可能性のあるものを対応単文の組の候補として求める．これは，3.2 節で述べたのと同様の方法で，文末文節に一致している単語があるかどうかで判断する．

3.5 単文間距離計算

単文の対応の組候補に対して，単文間の距離計算を行い，閾値以下であればそれらを対応する単文であると判定する．

距離は深谷ら [9] と同様に，2つの文に含まれる助詞，助動詞を除いた単語の頻度が一致する度合いによって計算する．

いま，単語の助詞，助動詞等の機能語を除いた単語集合をそれぞれ W_A，W_B とする．次に W_A，W_B の和集合を求めることにより，文 A，B のうち少なくともどちらかに現れる単語集合 $W_A \cup W_B$ を作成する．そして，この単語リストに対して 3.1 節で述べた方法で同様の単語として判定された異表記単語をまとめていくことにより，異表記単語クラスタ $c_1, c_2, \cdots, c_n$ を求める．具体的には，単語リスト $W_A \cup W_B$ に含まれる単語を $w_i (i = 1, 2, \cdots, m)$ とするとき，異表記単語クラスタ集合 $W_A \cup^* W_B = \{c_1, c_2, \cdots, c_n\} (n \leq m)$ を以下のアルゴリズムで求める．

(1) 各単語 w_i に対して1つの異表記単語クラスタ c_i を割り当て，クラスタの初期集合 $W_A \cup^* W_B$ を構成する．初期クラスタ数 n は単語リスト $W_A \cup W_B$ に含まれる単語数 m に等しい．すなわち，$c_i = \{w_i\}, w_i \in W_A \cup W_B (i = 1, 2, \cdots, m), W_A \cup^* W_B = \{c_1, c_2, \cdots, c_n\}$．

(2) $W_A \cup^* W_B$ 中のクラスタ c_i, c_j に対して $(1 \leq i < j \leq n)$，c_i 中の単語と c_j 中の単語で，同様の単語であると判定できるものがあれば c_i, c_j をまとめて1つのクラスタにする．すなわち，$c_i \cup c_j \rightarrow c_i$．そのような2つのクラスタが見つからなければ終了する．

(3) c_j を $W_A \cup^* W_B$ から取り除き (2) へ．

以上の手順で異表記単語クラスタ $c_1, c_2, \cdots, c_n$ を求めた後，それらクラスタに含まれる単語の頻度を，文 A，B それぞれについて求める．そして文 A, B の特徴ベクトル f_A, f_B を次のように定義する．

$$f_X = (h_X(c_1), h_X(c_2), \cdots h_X(c_n)) \tag{4}$$

ここで $h_X(c)$ は文 $X(= A, B)$ における異表記単語クラスタに含まれる単語頻度の和を $\sum_{k=1}^{n} h_X(c_k) = 1$ となるように全体で正規化したものを表す．この f_X を用いて，文 A, B 間の距離 $Dis(A, B)$ を，f_A, f_B の差の1ノルムで定義する．すなわち，

$$\begin{aligned} Dis(A, B) &= |f_A - f_B|_1 \\ &= \sum_{k=1}^{n} |h_A(c_k) - h_B(c_k)| \end{aligned} \tag{5}$$

と計算される．$\sum_{k=1}^{n} h_X(c_k) = 1$ であるため，$0 \leq Dis(A, B) \leq 2$ である．この文間の距離 $Dis(A, B)$ は文 A，B での同様の単語であると判断された単語が同じ回数出現した場合，最小値 0，使われている単語が全く異なる場合に対しては最大値 2 となる．

このようにして単文間の距離を求めたあと，距離が1より小さいものを対応単文の組とする．ただし1つの単文が複数の単文と対応しないようにするため，以下の手順で対応単文の組を求めていく．

(1) 単文間の距離をすべての組合せについて計算．
(2) 距離が 1 未満となる組合せを昇順のリストに並べる．
(3) 距離が最小となる組合せを対応単文の組と判定．この単文を含む組合せをリストから除く．
(4) リストが空になるまで (3) を繰り返す．

3.6 対応文，対応箇所の同定

以上の方法により，対応文の組の候補の各々について，そこに含まれる単文の対応がとれたことになる．最終的な対応文の組を以下の手順で求める．

(1) 対応文の組の候補の各々について，対応がとれた単文の数をカウントする．
(2) 対応がとれた単文の数が最大となる対応文の組の候補以外を対応文の組の候補から除去する．
(3) 対応文の組の候補が2つ以上存在するならば，対応文の組の候補の文間距離を計算し，距離が最も 0 に近い組，対応文の組の候補が1つならば，その組を対応文，対応のとれた単文を対応箇所として同定して終了する．

精度	0.681
再現率	0.784

表 4: 本研究手法による精度，再現率

精度	0.820
再現率	0.400
F 値	0.538

表 5: ベースラインの精度，再現率，F 値

4 評価実験

4.1 実験データについて

本研究では，文末の文節が一致する単語を持つ対応文のみを対象として行っているので，実験データは，2 節で調査した「1 対 1」の対応関係を持つ文集合のうちで文末の文節が一致する単語を持つ対応文 164 組を正解データとして用いた．また，単文に分割せずに，そのまま，文間距離を出現単語の 1 ノルムで計算し，距離が最も近い文を対応文と判定したものをベースラインとして本研究の提案手法と比較を行う．

4.2 対応文同定の評価

提案手法で対応文同定を行ったときの精度，再現率を表 4 に示す．

表より，文末の文節に一致する単語をもつ対応文に対して比較的良好な結果が得られていることが分かる．しかし，文末文節の一致条件を緩和させたために誤対応をしてしまったケースもあった．例えば，以下に示す 2 文は，対応文ではないが本研究手法では対応文であると判定してしまった．

- ドコモは 2 日までに FOMA で使う OS を、従来のトロンからリナックスなど複数のオープンソースを採用する**方針を決めた**。
- ＵＦＪは１９日、傘下のＵＦＪ信託銀行を９月をめどに住友信託銀行に売却する**方針を決めた**。

この 2 文は，文末の文節が一致しているので，対応文の組の候補に挙がり，単文に分割した際に 2 文とも「方針を決めた」という単文が作られてしまう．これらの単文が対応単文であると判断されたために対応文として判定されてしまった．本手法では，対応文候補の中から，最も対応単文の数が多いものを対応文であると判定しているために，ふさわしい対応候補が他に見つからなかった時にこのような誤対応が行われてしまう．最終的な対応文を同定する方法には，一考の余地がある．

4.3 ベースラインとの比較

本研究の有効性を検証するために，単純に文間類似度だけを用いた対応文抽出方法と比較を行った．ベースラインによる結果を表 5 に示す．

提案手法よりもベースラインでの手法の方が精度は高いが再現率は低い．ベースラインで正しく選択された正解の 90 %は提案手法でも正解と判定されたが，提案手法で正しく選択された正解の約 53.1 %はベースラインで正解として選択されなかった．実験に用いた文は，その多くが異なる情報を含んでいたので提案手法の方が，対応文に含まれる異なる情報に左右されにくいといえる．

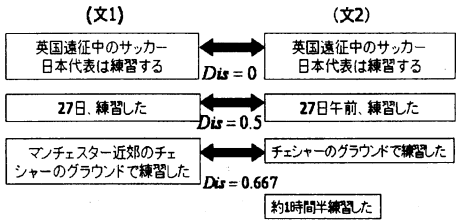

図 3: 対応箇所同定例

4.4 対応文の対応箇所同定例

本研究の手法により，文の文末文節が一致する対応文の同定については比較的良好な結果を得ることができた．対応文と判定された以下の文での対応箇所を図 3 に示す．

(1) 英国遠征中のサッカー日本代表は２７日、マンチェスター近郊のチェシャーのグラウンドで練習した。
(2) 英国遠征中のサッカー日本代表は２７日午前、チェシャーのグラウンドで、約１時間半練習した。

5 むすび

本研究では，対応文データベースを作成し，対応のバリエーションについて調査を行い，文の依存構造を利用することで対応文の検出と文内の対応箇所の同定を行った．文の文末文節が一致する文を対応文の候補として，単文単位での文間距離計算により対応文の同定，および対応箇所の同定手法を提案した．本研究は以下のような特徴がある．

- 文末文節，すなわち文の中心的な成分の一致を条件とすることで，文の構造を考慮した対応文

検出が可能となる．

- 単文間の距離計算のみを行っているので，対応文中に異なる情報を含んでいたとしてもその影響を受けない．
- 文レベルでの対応にとどまらず，さらに詳細な対応箇所の同定が行える．
- 対応文同定の条件を変えることで，容易に 1 対多の対応文も容易に扱うことが可能となる．

そして，対応文集合を用いて，文末文節が一致する対応文に対しては比較的良好な結果を得ることができた．しかし，文の依存構造を考慮しているが，助詞，助動詞を除いて処理を行っているので「太郎が次郎をたたいた」「太郎を次郎がたたいた」といったような文は誤対応となる．これらの問題を対処する方法として，用言に係る格情報を調べ，一致するかの判断を対応箇所同定の条件として加える方法が考えられる．今後これらの方法についても検討し問題に対処していきたい．また，複文や受身文などに対する対処方法も今後検討していく必要があるだろう．

謝辞

本研究の一部は文部科学省科学研究費補助金（課題番号 16200001）の支援による．

参考文献

[1] 奥村学，難波英嗣，”テキスト自動要約に関する最近の話題”，自然言語処理,Vol.9, No.4, 97-116,2002

[2] Ohara, M. Matsubara, S. and Inagaki, Y. ”Automatic Extraction of Translation Patterns from Bilingual Legal Corpus”, Proceedings of 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2003), pp. 150-157, Beijing, China, Oct. 2003.

[3] 熊野明，平川秀樹，”対訳文書からの機械翻訳専門用語辞書作成”，情処学論，vol.35, no.11, pp.2283-2290,1994.

[4] M.Haruno, S.Ikehara, and T.Yamazaki, ”Learning bilingual collocations by word-lelvel sorting”, The 16th International Conference on Computational Linguistics (COLING 1996), vol.1, pp.525-530,1996.

[5] 北村美穂子，松本裕治，”対訳コーパスを利用した対訳表現の自動抽出”，情処学論，vol.38, no.4, pp727-736, 1997.

[6] 高橋哲郎，乾健太郎，松本裕治，”テキストの構文的類似度の評価方法について”，情報処理学会自然言語処理研究会，NL-150-24, 2002.

[7] CaboCha/南瓜:Yet Another Japanese Dependency Structure Analyzer, http://chasen.org/ taku/software/cabocha/

[8] 日本電子化辞書研究所，”EDR 電子化辞書使用説明書”，日本電子化辞書研究所，1995.

[9] 深谷亮，山村毅，竹内義則，工藤博章，松本哲也，大西昇，”単語の頻度統計を用いた文章の類似性の定量化——部分的類似性の考慮——”，電子情報通信学会論文誌 D-II，Vol. J87-D-II，No. 2, pp.661-672 (2003)