大規模テキストを用いた２文間接続関係の同定

齋藤真実*1　山本和英*1　関根聡*2
*1 長岡技術科学大学電気系
*2 株式会社ランゲージ・クラフト研究所 / ニューヨーク大学
saito@nlp.nagaokaut.ac.jp, yamamoto@fw.ipsj.or.jp, sekine@cs.nyu.edu

1　はじめに

機械が文章を生成したり修正したりする場合に、文の間の接続関係を認識することは非常に重要である。例えば人間の質問に対話的に答えるシステムを考えた場合、対話をスムーズに行うために、システムは伝えるべき情報を自然な発話になるように繋げなければならない。その際に、文間に適切な接続詞を補う必要が出てくる。またドキュメント要約では、文章中から重要な文を選んで列挙する重要文抽出手法が主流であるが、とびとびになっている文が選ばれた際に接続詞を適切に修正（削除、追加、変更）する必要が出てくる。

本稿では、与えられた 2 つの文の間の接続関係を同定するシステムを提案する。同定のために使う知識は、何もタグ付けされていない大量のコーパスであり、基本的にシステムは、与えられた 2 文と同じような関係を持つコーパス中の文を参考に接続関係を同定する。最終的な文生成を考えると、接続詞を選ぶことが目的となるが、接続関係が同じ場合にはその接続関係にある接続詞のどれを選んでも構わないことが考えられるため、本稿では接続詞ではなく接続関係を対象とした。

文章の意味的関係を記述する枠組みとして修辞構造理論（RST: Rhetorical Structure Theory）がある。横山ら[3]は RST を基に SVM を用いた談話構造解析を提案しているが、教師付き機械学習であるため談話構造を付与したコーパスが必要となり、訓練データの量が問題となる。本システムではタグ付きデータを必要としないので Web や新聞からの大量のテキストが利用できる点で異なる。

我々は、人が書いた文章とは書き手が文章に込める深層の意味が表出したものであり、大量のテキストを用意することで統計的に文章中に内在する接続関係を取り出せると考える。本稿では文中の単語や文の構文パタンを用いて自動的に接続関係を推定する手法を提案する。

2　接続関係の分類

接続関係の種類に関しては多くの研究者が個々の案を提示している ([1][2]) 。市川[2]は文の接続関係を、「順接」、「逆接」、「添加」、「対比」、「転換」、「同列」、「補足」、「連鎖」の 8 つの類型に分類している。そこで我々は「茶筌」の辞書（IPADIC, Ver.2.7.0）に登録されている 167 個の接続詞を、市川の分類をもとに単語や構文による偏りを考慮しつつ表 1 のように分類した。また、接続詞によっては一意に接続関係を決められないものもある。その場合は、取りうるすべての接続関係の中に分類している。例 1、2 の接続詞「したがって」はそれぞれ「なので」と同様な因果関係、「つまり」と同様な並列関係に分類可能である。

例 1)「理系の人間だって科学のごく一部しか勉強できない。したがって、文系の人がたくさんの理系の授業を受ける必要はない。」
例 2)「高気圧におおわれた地域は、天気がよくなります。したがって、雨が降らないのです。」

表 1. 接続関係の分類とコーパス中に含まれる割合

接続関係	接続詞	コーパス中の割合[%]
累加	また、そして、しかも、なぜなら、まずは、…	43.0
加反	しかし、だが、でも、なのに、ところが、…	32.2
因果	だから、ゆえに、なので、すると、そうして、…	12.1
並列	一方、もしくは、あるいは、すなわち、つまり、…	6.0
転換	さて、ところで、では、…	5.1
例示	例えば、たとえば(2 個)	1.5
その他	なかんずく、わけても、…	0.2

ここでの分類は「その他」を含む 7 種類であるが、「その他」の分類は出現頻度が低く、またその分類自体に特徴が見られないことから、以後の処理では除いている。本稿では「その他」を除く 6 分類を行う処理について述べる。

3　システムの概要

システムは、2 つの文を入力とし、1 文目（前文）と 2 文目（後文）の単語や構文の情報から、後文の頭に置くべき適切な接続詞の接続関係を推定する。システムは主に

「単語要素による判定部」と「構文パタンによる判定部」の二つのサブシステムからなる。この二つの判定部については 4 章と 5 章でそれぞれ解説する。図 1 にシステムの処理の流れを示す。

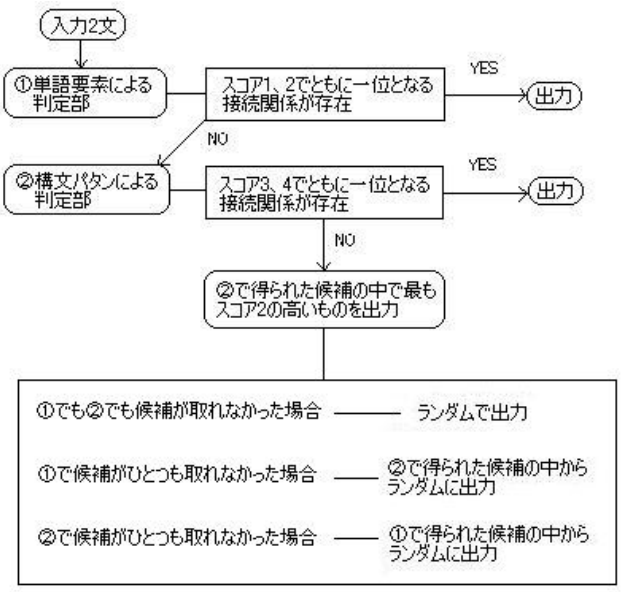

図 1. システムの処理の流れ

4　単語要素による判定

「努力」と「才能」、「夢」と「現実」など、相反する意味を持つ単語が前文と後文に含まれるとすれば、その接続関係は「加反」である可能性が高い。また、前文中の語の下位概念の語が後文中に含まれているとすれば、「例示」かもしれない。我々は、辞書やシソーラスだけでは捉えきれない、このような知識をコーパス中から統計的に得ることで、接続関係を判断することができると仮定する。このような知識獲得のため、次に示す入力の 2 文から抽出した単語要素による接続詞の判定を行った。

単語要素とは、茶筌体系での「名詞(一般)」、「名詞(サ変接続)」、「動詞」、「形容詞」である。これらを 4.1 節の①～④の組み合わせで前文と後文から一語ずつ取り出し、その単語ペアのコーパス中での頻度からスコアを計算する。単語要素を限定したのは、接続関係を推定するために必要のないと思われる単語ペアの影響を減らすためである。

コーパスは、接続詞でつながった 2 文を 1 セットとして、Web 文書から約 130 万セット、新聞記事から約 15 万セット、合計で約 145 万セットを使用した。

4.1 単語要素の抽出と組み合わせ

入力の前文と後文から取り出した単語を用いて次に示す組み合わせで単語ペアを作成する。

例 3）接続関係：「加反」（接続詞：「しかし」）

前文:「このため、**記録**は**現時点**では**分からない**。」

後文:「**感覚**としては、**満足できる**ものだったのではないかと**思う**。」

①「名詞(一般)」どうしの組み合わせ

“ 現時点＋：感覚＋ ”

②「名詞(サ変接続)」どうしの組み合わせ

“ 記録＋：満足＋ ”

③「動詞(or 形容詞)」どうし、または「動詞」と「形容詞」の組み合わせ

“ 分かる－：できる＋ ”、“ 分かる－：思う＋ ”

④「動詞」と「名詞(サ変接続)」の組み合わせ

“ 分かる－：満足＋ ”、“ 記録＋：できる＋ ”、

“ 記録＋：思う＋ ”

抽出した単語にはすべて基本形を使用する。また、各要素には肯定・否定を示す「＋」と「－」が付与されている。同じ文節中に否定の「ない」を含む場合は同文節中のすべての要素に「－」の情報が付与される。それ以外の要素はすべて「＋」となる。二重否定に関しては考慮していない。例 1 では、前文の「分からない」に否定の「ない」が含まれるので、「分かる」に「－」が付与されている。

例 3 では、“ 分かる－：できる＋ ” と “ 分かる－：満足＋ ” が有効に働き、「加反」が得られることを期待する。

4.2 スコアの計算

これらの単語ペアをコーパス中で検索し、そのときの頻度から以下の式によって各接続関係に対してスコアを計算する。スコア 1 は単語ペア i に対して、コーパス中で同じ単語ペアを持つ文章中で接続関係 conj となる割合を考え、抽出した単語ペアすべてに対して合計している。また、スコア 2 はスコア 1 の値を接続関係 conj のコーパス中での割合で割ることで、入力文を与えられたときにその接続関係が元のコーパスでの割合に比べ、何倍出現しやすいかを表している。

スコア 1：

$$S_1(conj) = \frac{\sum_i Hit\_pair_{conj}(i)}{\sum_{conj}\sum_i Hit\_pair_{conj}(i)} \quad \cdots(1)$$

スコア 2：

$$S_2(conj) = \frac{\sum_i Hit\_pair_{conj}(i)}{\left(\sum_{conj}\sum_i Hit\_pair_{conj}(i)\right) \times Rate_{conj}} \quad \cdots(2)$$

$Hit\_pair_{conj}(i)$：単語ペア i がマッチしたときに接続関係が conj であった数

$Rate_{conj}$：接続関係 conj のコーパス中の割合（表 1）

両スコアで一位となる接続関係が存在する場合、その接続関係をシステムの出力とする。ここで、一位となる接続関係がスコア 1 とスコア 2 で異なる場合、判定はその後の構文パタンによる判定部で行う。

5　構文パタンによる判定

例えば「～は～だった。～というわけではない。」という文からは「加反」が、「～は～である。～など。」という文からは「例示」の関係が読み取れる。このように、4 節で扱っているような単語情報がなかったとしても、人は構文的情報から 2 文間の接続関係を読み取ることができる。

以下では構文パタンによる接続関係の判定手法について述べる。入力の 2 文から 5.1 節に示す方法で機能語を中心とした構文パタンを作成し、コーパス中での頻度からスコアを計算する。入力の 2 文は係り受け解析器「南瓜」によってあらかじめ構文解析されている。使用したコーパスは 4 節で扱ったものと同様のものを使用している。

5.1 構文パタンの生成手法

入力文を用いて 5.1.1～5.1.3 の手順に従って単語と文節の組み合わせによってパタンを作成する。ここで、前文と後文からそれぞれ 1 文節以上、合計で 3 文節以上となる組み合わせをすべてパタンとする。

例 4）接続関係：「加反」（接続詞：「しかし」）

前文:「高齢者介護の課題は国民すべてに例外なくたち現れる課題となった。」

後文:「利用できる介護サービスは現状では充分ではない。」

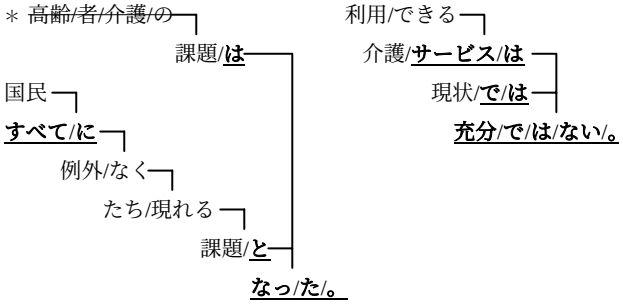

図 2. 構文解析結果

5.1.1 パタン作成に不要な文節の削除

名詞にかかる修飾部は接続関係を読み取る際の手がかりとはならないので、パタン作成の前にまず不要な文節を削除しておく。削除する文節は、文節の末尾が「助詞(連体化)」の「の」で終わっているものとその文節の係り元すべて、さらに末尾が「助詞(並列接続)」で終わっているものとした。例 4 では、「高齢者介護の」が削除される。

5.1.2 パタンの構成要素として必要な単語の抽出

残っている各文節から「(名詞|動詞|形容詞)？(助詞|助動詞|。)＊＄」または「副詞＄」にマッチする文字列を抽出し、パタンを作成する際の構成要素とする。(？:直前の文字列が 0 または 1 個，＊: 直前の文字列が 0 個以上，＄: 文末)。ただし、ここでの「名詞」は「名詞(一般)」以外のものとし、文節末に「、」がある場合はそれを無視している。

5.1.3 文節の組み合わせと単語の選択

前文と後文から 1 文節以上、合計で 3 文節以上という条件で、パタンに使用する文節の組み合わせを作り、5.1.2 で抽出した単語を用いてパタンを作成する。これまでの処理でパタンの構成要素が取り出せなかった文節は無視する。組み合わせを考える際には PrefixSpan（Ver.0.4）を利用した。以下の＊は任意の文字列が入ることを示す。パタン 1 は残ったすべての文節を採用した場合、パタン 2 は最小の 3 文節によるパタンの例である。

パタン 1:“ は＊すべてに＊と＊なった。＊サービスは＊では＊充分ではない。”

パタン 2:“ は＊なった。＊充分ではない。”

次に、各文節中の単語から①「名詞(サ変接続)」と「名詞(固有名詞)」、②「動詞」と「形容詞」、③「副詞」と残っている「名詞」を順に削除し、残った単語を用いて同様にパタンを作成する。パタン 3～5 はパタン 1 からこれらを順に削除したときの例である。

パタン 3:“ は＊すべてに＊と＊なった。＊は＊では＊充分ではない。”

パタン 4:“ は＊すべてに＊と＊た。＊は＊では＊充分ではない。”

パタン 5:“ は＊に＊と＊た。＊は＊では＊ではない。”

5.2 ランク付けのためのスコア計算

例 4 の 2 文から作成されたパタン数は 224 個であった。これら 224 個のパタンを用いてコーパス中で、各接続関係となる 2 文を探し、その頻度情報からスコアを求める。ただし、コーパス中での頻度が 1000 回を越えるパタンは「～は～た。～に～する。」など、どんな接続関係も取れる一般的構文であると考え、使用していない。

スコアは、式(1)、(2)の $Hit\_pair_{conj}(i)$ をパタン i がマッチしたときにその接続関係(conj)であった数($Hit\_pt_{conj}(i)$)に置き換えて、同様の計算をする。このときのスコアをそれぞれスコア 3、スコア 4 とする。両スコアで一位となる接続関係が存在する場合、その接続関係をシステムの出力とする。

6　実験と評価

6.1 Web 文書を入力とした評価

Web 文書から「その他」を除く 6 種類の接続関係を持つ入力 2 文をそれぞれ 50 セット、合計で 300 セット用意し、実験した結果を表 2 に示す。ここで、すべての接続関係に対して同量の入力を用意したが、表 1 に示すようにコーパス中での各接続関係の出現頻度にはかなり偏りがある。表 2 の open テストで各接続関係の出現率が表 1 の通りであったとすると、システムの精度としては 64.8%が期待される。すべての入力に対して出現頻度の最も高い「累加」を答えるシステムをベースラインとすると、その正解率は「累加」の出現率である 43%となる。本システムではこのベースラインに比べ、正解率が約 20%高くなった。

表 2．Web 文書を入力とした評価結果

接続関係	closed	open
累加	0.82 (41/50)	0.52 (26/50)
加反	0.88 (44/50)	0.86 (43/50)
因果	0.96 (48/50)	0.56 (28/50)
並列	0.98 (49/50)	0.58 (29/50)
転換	0.92 (46/50)	0.72 (36/50)
例示	0.98 (49/50)	0.60 (30/50)
合計	0.92 (277/300)	0.64 (192/300)

コーパス中での割合は「累加」と「加反」が圧倒的に高いため、スコア 1 やスコア 3 のみの判定ではこの二つが一位になりやすい。このままではベースラインとあまり変わらないが、そこにコーパス中での割合を考慮したスコア（スコア 2 とスコア 4）を並列的に用いることでコーパス中での接続関係の偏りに影響されることなく判定できていると考えられる。

6.2 人手で判定できるものに限った評価

2 つの入力文だけを与えられても、人間でもその接続関係を同定できないことがある。そこで、さらに別のテストセットとして人手で判断した場合に一意に接続関係が取れる入力 2 文を Web 文書から各 25 セット、新聞記事（毎日新聞）から各 25 セット用意し同様の実験を行った。結果を表 3 に示す。

人手で正しく判断可能なものに入力を限定した場合、正解率は Web で約 10%向上した。これは、人手でも判断のつかない原因として単語による情報の少なさが考えられる。単語要素のみでシステムが接続関係を出力する割合は、表 2 の open テストでは 52%、表 3 の Web からの入力に対するテストでは 61%であり、大部分を占めている。またその正解率はそれぞれ 78%と 96%であった。単語要素を使わずに構文パタンのみで判定すると、システムが答えを出力する割合は 32～37%、正解率は 68%前後で、人手で判断可能なものとそうでないものとでの違いはない。

人手で判断のつかないものに対しては、直前の文だけでは接続関係を判断するための情報が少なく、文章全体の流れを読む必要がある。また、本システムの単語要素による判定における単語要素と要素の組み合わせ方に関してはもう少し考える余地があるように思う。より適切に接続関係判定の手がかりとなる要素を抽出、限定し、単語要素と構文パタンの両方を考慮したスコア付けを考えることで精度が向上すると期待する。

表３．人手で判断可能な入力に対する評価結果

接続関係	Web	新聞
累加	0.64 (16/25)	0.60 (15/25)
加反	0.88 (22/25)	0.48 (12/25)
因果	0.84 (21/25)	0.20 (5/25)
並列	0.80 (20/25)	0.44 (11/25)
転換	0.80 (20/25)	0.24 (6/25)
例示	0.72 (18/25)	0.36 (9/25)
合計	0.78 (117/150)	0.39 (58/150)

7　まとめ

接続詞付きの大規模なテキストデータから、入力の 2 文間の接続関係を、単語と構文的要素の 2 つの観点から統計的に判断する手法を提案した。単語対と構文パタンを用いることで約 7 割の精度で 2 文間の接続関係を同定することができた。

謝辞

本研究の一部は、平成 17 年度 長岡技術科学大学 学長裁量経費 研究助成(A) 「小学校２年生の国語問題を解く」によって実施した。

参考文献

[1]Florian Wolf, Edward Gibson: Representing Discourse Coherence: A Corpus-Based Study, *Computational Linguistics*, Vol.31, No.2, pp.249-287, (2005).

[2]市川孝: 国語教育のための文章論概説，教育出版，pp.65-67(1978).

[3]横山憲司，難破英嗣，奥村学: Support Vector Machine を用いた談話構造解析，情報処理学会研究報告，NL154-27(2003).