# 類似用例による文間接続関係の推定

齋藤真実　山本和英
長岡技術科学大学電気系
{saito, ykaz}@nlp.nagaokaut.ac.jp

## 1 はじめに

以前我々は自然言語処理の持つ問題点の洗い出しを目的とし、小学 2 年生を対象に国語問題の自動解答を試みた[2]。さらにそこで挙げられた問題点の中から文間接続関係の推定問題に着目した。Marcu ら[1]は単語情報を用いてナイーブベイズ分類器による機械学習手法を提案している。我々は入力文から得られる単語情報に加え、構文情報を用いて接続関係を大量のコーパスから統計的に推定する手法を提案した[3]。

文間の接続関係を同定することは、談話解析、ドキュメント要約、質問応答など様々な分野で応用可能である。我々の先行研究[3]では入力のテキストに依存してしまう問題があり、当初扱おうとした小学 2 年生の国語問題には対応できなかった。そこで本稿では、大量の WEB 文書から与えられた二文に類似した用例を探すことで二文間の接続関係を推定するシステムを提案する。本稿では接続関係を推定するためのルールを作るのではなく、用例利用型（example-based）の手法をとる。

## 2 本手法の概要

入力文と類似した二文の組を探すといっても、単純に文が似ているものを探せばよいというものではない。例えば「雨が降った。試合は中止になった。（因果）」と「雨が降った。試合は中止にならなかった。（逆接）」は、単語の一致率などを用いた一般的な類似度計算によって非常によく似た二文の組とされるものであるが、それぞれの文間の接続関係はまったく逆である。反対に、「本を読んだ。つまらなかった。（逆接）」と「評判の映画を観た。僕には退屈で眠くなった。（逆接）」では、一致する単語や一般的に類義語や上位語、下位語とされるものが存在しないにもかかわらず、人間が見ると直感的にこの二つの例文は似ている文であり、接続関係も同じであるといえる。本稿ではこのように、入力に対して同じ接続関係を持つと思われる類似用例文を大量のコーパス中から探し、その用例によって接続関係を推定する手法を提案する。以下に大まかな処理の流れを示す。

**Step1.** 入力の二文から構文情報を用いてパタンを生成し、大量の WEB 文書から構文的に似た用例を候補として抽出する。この結果得られた文を候補文と呼ぶ。ここで候補文に対するパタンスコアも計算する。

**Step2.** クラスタリング用のコーパスから接続関係を決定すると考えられる主要な単語を抽出し、GETA[3]を用いてクラスタリングする。「本を読んだ。つまらなかった。」と「評判の映画を観た。僕には退屈で眠くなった。」の例では、「本をー映画を」、「読んだー観た」、「つまらないー眠い」といった文節内の主要素でクラスタリングされることが理想である。さらに入力文と候補文を比較し、候補文に対して単語スコアを計算する。

**Step3.** パタンスコアおよび単語スコアを用いて抽出した候補文に対して入力文との類似度を計算し、類似度の高い順に出力する。

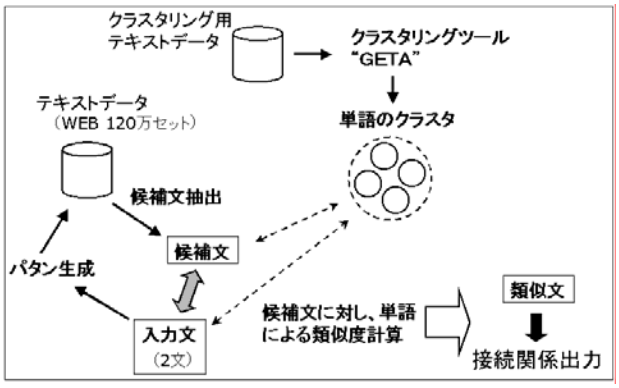



図１　類似用例による接続関係の推定

## 3 構文パタンによる候補文の抽出

本章では入力文からの構文パタン生成手法および候補文の抽出について説明する。また、この構文パタンは 4 章で述べる単語のクラスタリングの素性としても使用する。

### 3.1 構文パタンの生成

まず入力の一文目、二文目からそれぞれパタンを作成する。入力の各文から生成されるパタンを基本パタンと呼ぶ。本節では「未知語」は「名詞」と同等に扱っている。また、「。」以外の「記号」は全て無視している。例 1 を用いて構文パタンの生成について説明する。

例 1）入力の二文の例
一文目：「上海の新生活はサンディにとって心地よいものとなるはずだった。」
二文目：「最愛の母親の死は彼女に大きな打撃と計り知れない心痛を与えた。」

ｉ）文末文節以外の文節からのパタン要素の抽出

パタンを構成する要素を文節単位で抽出する。構文解析器には「南瓜[1]」を用いた。文節内の助詞、助動詞を抽出し、さらに全ての品詞で「非自立」であるものと動詞の「ある」を無条件にパタン要素として採用する。

文節末が名詞または動詞の場合はそれぞれ“NOUN”と“VERB”に一般化し、それで一要素とする。ただし NE（固有表現）タグがついているものは NE タグに変換する。

また、「～の」といった連体修飾節はパタンの要素には採用しない。ただし、NE タグの要素を含む場合を除く。例えば、例 1 の「最愛の」、「母の」はパタンの要素として採用しないが、「上海の」は「LOCATION の」という形で採用する。

ⅱ）文末文節からのパタン要素の抽出

文末文節に関しては他の文節とは異なり、複数のパタン要素が生成される。文末文節の末尾から一形態素ずつ付与して複数のパタン要素を作成する。ここで、抽出対象となるものは助詞、助動詞、感動詞および全ての品詞で「非自立」であるもの、動詞の「ある」である。ただし、文末の助詞または助動詞の連続は切り離さない。例1の文末文節の「なるはずだった。」からは「はずだった。」と「だった。」の二種類のパタン要素が生成され、これらをそれぞれ末尾として構文パタンを生成する。つまり、ここでは助動詞「た。」の要素は作成されない。また文末が形容詞ならば、「形容詞（出現形）＋『。』」とし、文末が名詞または動詞の場合はⅰ）の「文末文節以外の文節からのパタン要素の抽出」と同じくそれぞれ“NOUN”と“VERB”に一般化し一要素とする。

このようにして各文節から取り出したパタン要素とそれぞれ係り受けを図 2 に示す。入力文中で形態素が離れている場合は間に0個以上の任意の文字列を意味する「＊」を入れる。

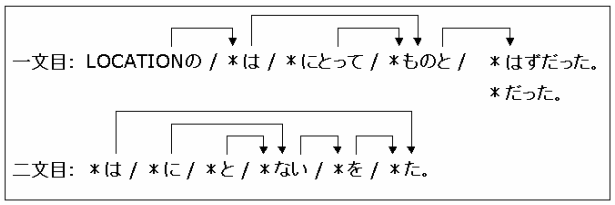



図2　各パタン要素と係り受け

一文目、二文目でそれぞれ各文節から係り先文節のパタン要素を連結させ基本パタンを生成する。文末については要素が複数存在するので、それぞれの要素について構文パタンを作成する。

生成した一文目と二文目の基本パタンの全ての組み合わせを、入力に対する構文パタンとする。図 3 に例 1 から生成される構文パタンの例を示す。

**PT1:** LOCATIONの＊は＊ものと＊はずだった。＊に＊ない＊を＊た。
**PT2:** ＊は＊ものと＊はずだった。＊に＊ない＊を＊た。
**PT3:** LOCATIONの＊は＊ものと＊はずだった。＊は＊た。
**PT4:** LOCATIONの＊は＊ものと＊だった。＊は＊た。
⋮

図3　例1から生成される構文パタン例

例1の二文目の文末が「与える。」だった場合は、文末文節から抽出される要素が「た。」ではなく「与える」を一般化した“VERB”となるため、例えば図3の **PT1** は「LOCATIONの＊は＊ものと＊はずだった。＊に＊ない＊を＊VERB。」といったパタンになる。

## 3. 2 候補文の抽出

パタン生成後はパタンをコーパス中で照合し、二文の組を探す。WEB 文書から接続詞でつながった二文の組を約 120 万セット抽出し、これをパタンの照合に使用した。生成したパタンを長いものから順に同じ要素数のパタンを一組として照合させる。ここでは実験的に、抽出された候補文が 100 セットを超えたとき照合を終了するとした。ここで抽出した候補文に対して 5 章で示すスコア付けによって入力に最も近い文を探す。

# 4 GETA による単語のクラスタリング

本章では、単語のクラスタリングについて説明する。本章で生成した単語のクラスタは入力文と候補文との類似度を測る際に使用する。

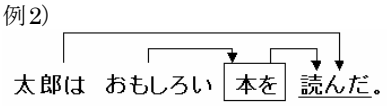
## 4. 1 クラスタリングに用いた素性

一文目、および二文目の述語と、一文目、二文目それぞれの述語に係る格要素について接続関係が同じ文で用いられやすい単語のクラスタを作成する。すなわち、ここでは 4 つのクラスタリングを行うことになる。単語のクラスタリングでは WEB 文書から抽出した二文の組を使用したが GETA[3]の処理時間との関係から、二文の組 1 万セットを用いて分類を行った。ここで使用した 1 万セットは候補文抽出の際の構文パタンの照合で用いた 120 万セットとは別のものを使用している。クラスタと単語は一対一ではなく、ある単語が複数のクラスタに属す場合も存在する。

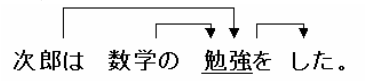
ⅰ）述語の同定

本節で述べる述語とは、茶筌[2]の品詞体系で「動詞」（基本形が「する」,「ある」,「なる」,「せる」,「れる」,「られる」であるものを除く）と「名詞－サ変」および「形容詞」となるもので、文末文節中で最も文末に近いものをいう。ここで、品詞が動詞であるものは全て基本形にしている。また、品詞が動詞で、基本形が「する」,「ある」,「なる」,「せる」,「れる」,「られる」となるもののみが文末文節にある場合はその係り元文節から同様にして探す。例 2 では「読む」が述語となり、例 3 では「勉強」が述語となる。

例2)




例3)




ⅱ）述語に係る格要素の抽出

述語に係る格要素とは、ⅰ）の「述語の同定」で抽出した述語を含む文節に係る各文節で、文節末が「名詞+助詞」となるもの全てを指す。ただし、ここでいう名詞は茶筌[2]の品詞体系で「名詞－一般」および「名詞－サ変」となるもののみを扱っている。さらに名詞が連続している場合は末尾の名詞のみを使用する。「名詞+の」は格要素として使用していない。また、述語が抽出されなかった場合は文末文節に係る格要素をここでいう述語に係る格要素として使用している。

例 2 では述語「読む」に係る「本を」が格要素となる。例 3 では「勉強」に係る文節が「数学の」のみであるが、これは「名詞+の」の形であるため、この例からは格要素は抽出されない。

## 4. 2 クラスタ数の決定

本稿では単語のクラスタを単語の汎化の目的で使用するため、一種類の単語しか存在しないクラスタは汎化の意味をもたないと考えた。そこで一つのクラスタに複数の単語が存在するという条件の下で、最も多いクラスタ数を採用した。

## 5 候補文に対するスコア付け

本章では 3 章で抽出してきたコーパス中に存在する候補文に対してスコア付けを行い、入力文に最も類似した候補文を探す。

### 5.1 構文パタンによるスコア

入力文 $i$ が与えられたときの構文パタンによる候補文 $c$ のスコアを、$S_{PT}(i, c)$とする。候補文を抽出する際に使用したパタンをコーパス中で照合し、得られた二文の組（候補文）の数の逆数をパタンスコアとして用いた。つまり、あるパタンAを用いて抽出した候補文が 5 文あったとき、それら全ての候補文に対してパタンスコア 1/5 が付与される。

### 5.2 各単語スコアの計算

単語による候補文 $c$ の単語スコアの計算について説明する。まず、一文目の述語、二文目の述語、一文目の格要素、二文目の格要素をそれぞれ $V1$, $V2$, $N1$, $N2$ とする。この 4 種類の単語それぞれから候補文 $c$ の単語スコアを計算する。個々の単語 w によって計算された、入力文 $i$ が与えられたときの候補文 $c$ のスコア $s_w(i, c)$を 5.3 節でまとめ、候補文 $c$ の単語スコア $S_W(i, c)$としている。候補文 $c$ から得られる単語 w を $w_c$、入力文 $i$ から得られる単語 w を $w_i$ とする。また、ある単語 $\alpha$ が属すクラスタを $G(\alpha)$としている。ここで単語 $w_i$, $w_c$ は述語とそれに係る格要素であり、ともに同種の単語でなければならない。例えば一文目の述語による単語スコア $s_{V1}(i, c)$を計算する場合、$w_i$ と $w_c$はどちらも一文目の述語$(V1_i, V1_c)$となる。$s_w(i, c)$の初期値をそれぞれ0.001 とし、式(1)のように各単語スコアを計算する。式（1）の$|G(w_i \cap w_c)|$ は $w_i$ と $w_c$ を含むクラスタの数を意味する。

$$s_w(i,c) = \sum_{w_i, w_c \in i,c} score(w_i, w_c), \qquad w \in \{V1, V2, N1, N2\}$$

$$score(w_i, w_c) = \begin{cases} 1 & if \quad w_i = w_c \\ \dfrac{1}{|G(w_i \cap w_c)|} & if \quad G(w_i \cap w_c) \neq \phi \end{cases} \tag{1}$$

$w_i$ と $w_c$ を含むクラスタが複数存在する場合、$w_i$ と $w_c$ の語の意味が広いものであると考え、ペナルティの意味を含め、その逆数をスコアに加算することとした。

### 5.3 候補文に対するスコア計算

パタンスコアおよび単語スコアを用いて入力文 $i$ と候補文 $c$ の類似度 $Sim(i, c)$を式（2）のように計算した。

$$\begin{cases} Sim(i,c) = S_{PT}(i,c) \times S_W(i,c) \\ S_W(i,c) = \{s_{N1}(i,c) \times s_{V1}(i,c)\} \times \{s_{N2}(i,c) \times s_{V2}(i,c)\} \\ \qquad\qquad \times \{s_{V1}(i,c) \times s_{V2}(i,c)\} \end{cases} \tag{2}$$

入力文と候補文の一文目の述語が同一もしくは類似であったとしても、二文目の述語がまったく異なるものでは入力文と候補文が類似であるとはいえないため $s_{V1}(i, c)$と $s_{V2}(i, c)$を掛け合わせている。また、格要素の類似性は述語と一組で考えるべきであるため、一文目と二文目でそれぞれ述語と格要素のスコアを掛け合わせている。それらを全て掛け合わせた $S_W(i, c)$を最終的な単語スコアとし、それに 5.1 節のパタンスコアを掛けたものを候補文に対する類似度としている。この類似度が最も高い候補文の接続関係を入力の二文間の接続関係として出力する。

## 6 評価実験および考察

本実験で、入力文に対して類似した二文の組を探すために 120 万セットの WEB コーパスを使用した。

WEB 文書を入力としたテストでは、入力として WEB 文書から接続詞でつながった二文を 6 種類の接続関係に対して 50 セットずつランダムに抽出した。これらをシステムの入力とするが、形態素解析のミス等により入力文から構文パタンが生成されず、候補文がひとつも得られないものが 19 セット（累加 : 2、逆接 : 3、転換 : 8、例示 : 6）あったため、これらを除いた合計 281 問に対して実験を行った。

実験では二文目の文頭の接続詞を除いた形で二文を入力としている。ここで、正解は元の接続詞が属す接続関係としている。本章では、複数の接続関係に属す接続詞も対象としているが、それらはテストセットには含んでいない。

### 6.1 WEB 文書からの入力に対する評価

WEB 文書を入力としたときの評価結果を表 1 に示す。どの候補文に対してもスコアが加算されず差が出ない場合、コーパス中で最も頻度の高い「累加」を答えとして出力するものとした。ここでは接続関係ごとの正解率と各接続関係の出現頻度を考慮した場合（weighted）で正解率を求めた。また weighted との比較で、使用したコーパス（120 万セット）中で最も多く存在した「累加」の割合をベースラインとした。

表 1　WEB 文書を入力とした評価結果

| 接続関係 | 問題数 | 正解率 |
|---|---|---|
| 累加 | 48 | 0.516 |
| 逆接 | 47 | 0.572 |
| 因果 | 50 | 0.342 |
| 並列 | 50 | 0.562 |
| 転換 | 42 | 0.406 |
| 例示 | 44 | 0.470 |
| 合計 | 281 | 0.479 |
| Weighted | – | 0.508 |
| ベースライン | – | 0.425 |
| 人手による評価 | | 0.754 |

実験の結果、同スコアで一位となる候補文が複数出力される場合が多く存在した。そこで本実験では一位の候補文の接続関係を全て出力として、出力された接続関係の種類数に対する正解の割合をその入力に対する正解ポイントとし、累計を問題数で割り、正解率とした。つまり、例えばある入力に対して最も高い類似度を持つ二文の組が 4 セット得られたとする。それぞれの二文間の接続関係が「因果」、「累加」、「因果」、「逆接」ならば、システムは「因果」、「累加」、「逆接」の 3 つを出力する。正解が「因果」であった場合、この入力に対する正解ポイントは 1/3 となる。正解率はこれらの合計を問題文の総数で割って求める。

人手による評価では、三人の被験者にシステムが出力した接続関係と入力の二文を提示し、システムが出力した接続関係で正し

いと思うものを判断してもらった。二人以上の被験者が正しいと判断した接続関係を正解として同様に正解率を求めた。

評価結果からベースラインよりも高い正解率が得られた。人手による評価では自動評価に比べて高い正解率となっている。これは、入力を二文に限定した場合で正解の幅が広がっているためと考えられる。本研究においては文間の接続関係そのものの曖昧性[4]が大きいため、正解を何とするかが問題となる。自動評価では「元の二文をつなぐ接続詞が属す接続関係」を正解と定義しているが、応用分野によっても正解の幅は異なる。しかし、一般的に人間が見て適切であると判断できる範囲であれば実用に耐え得ると考える。人手による評価では 75.4%の正解率となり、良好な正解率が得られたと考えている。

また今回述語とそれに係る格要素のみに限定することで、文中の主題間の接続関係を正確に把握できると考えた。しかし本手法での単語の抽出方法では、一文目の述語（$V1$）、二文目の述語（$V2$）、一文目の述語に係る格要素（$N1$）、二文目の述語（$N2$）に係る格要素の 4 種類全てが取り出せる二文の組は本実験でのテストセット中で 1 割程度しか存在しなかった。表 2 に実験で用いた 281 セットの中で取り出せた単語の種類数とその問題数を示す。

表 2　抽出された単語の種類数と問題数

| 単語の種類数 | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 問題数 | 23 | 53 | 88 | 90 | 27 |

入力文から $V1$, $V2$, $N1$, $N2$ のどれも抽出できなかった問題は全部で 23 問あった。これらの入力文は主に「AはBだ。」といった“ダ文”からなるものであった。本手法では述語を「動詞」、「サ変名詞」、「形容詞」に限定しており、「一般名詞」は対象としていない。一般名詞を対象としなかったのは、体言止の文章を考慮したためであるが、「AはBだ。」といった“ダ文”に対しても別に対処する必要がある。

一般に文間の接続関係はその全ての要素で決まるわけではなく、ある一つの単語によって関係が定義される場合もある。だが、抽出する単語を限定していることで必要な情報を落としている場合も多く存在する。かといって入力文が与えられたときにどの単語に注目すべきなのかを自動的に判断するのは難しい。述語とそれに係る格要素以外の部分に関しては、係り受けや品詞では重要部分を同定できない。今後これをどうシステムで対応するかは重要な課題である。

### 6. 2 小学生の文章を入力としたときの評価実験

小学 2 年生の国語の問題集に含まれる接続関係を選ぶ問題から 78 セットの二文の組を取り出し、それらを入力とした場合の実験も同様に行った。表 3 に結果を示す。ベースラインは 78 問中で最も多い接続関係（「因果」）の割合とした。また、我々の先行研究[3]のシステムで同様に実験を行った結果も示す。

表 3 より小学生レベルの入力文に対しても WEB 文書を入力とした場合と同程度の正解率が得られた。先行研究[3]では、入力文から得られる構文的特徴により WEB 文書からの入力では適切な統計量が得られたが、小学生用に用いられる文章はシンプルな文体であるものが多いため構文的特徴がほとんど見られない。さらに文が短いため、得られる単語情報も少なく、入力文に出現した単語をそのまま用いただけではスパースネスの問題があった。本手法では入力文から取り出せる単語が少ない場合でも GETA[3]を用いた単語の汎化によって単語の意味範囲を広げ、類似文を探せるようにした。また先行研究[3]に比べ、より一般的な形にした構文パタンを候補文の抽出目的に限定して使用した。これらによって、スパースネスの問題を解消し、構文的特徴の少ない文に対しても正解を探すことができるようになったと考える。

表 3 小学生の文章を入力とした場合の評価結果

| 接続関係 | 問題数 | 先行研究[3] | 本システム |
|---|---|---|---|
| 累加 | 19 | 0.316 | 0.509 |
| 逆接 | 24 | 0.458 | 0.660 |
| 因果 | 25 | 0.160 | 0.350 |
| 並列 | 6 | 0.167 | 0.450 |
| 転換 | 3 | 0.333 | 0.067 |
| 例示 | 1 | 0.000 | 0.000 |
| 合計 | 78 | 0.295 | 0.476 |
| ベースライン | − | 0.321 | |
| 人手による評価 | − | − | 0.628 |

## 7 おわりに

本稿では用例利用型（example-based）の文間接続関係推定手法を提案した。文から抽出する単語情報を文中の述語とそれに係る格要素に限定し、それらを用いて入力の二文に最も「近い」接続関係をもつ二文の組をコーパス中から探し、その類似用例によって文間の接続関係を推定している。クラスタリングツール GETA[3]を用いて同じ接続関係をもつ二文の中で同様の使われ方をする単語をクラスタリングし、それを用いて単語の汎化を行った。これにより、表層的な情報が少ない小学 2 年生の問題に対して、先行研究[3]の統計的手法では 29.5%だった正解率を 48.7%まで上げることができた。本手法により、文の難易度や文体によらずに接続詞を推定し、小学生の文章にも対応できるようになったといえる。さらなる正解率の向上を目指すためには、入力文に対して文の意味を理解するための必要最小限の単語の抽出と単語のクラスタリングの精度の向上が必要不可欠であると考える。

## 使用した知識およびツール

(1) 構文解析器, “南瓜”, Ver.0.50, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/~taku/software/chabocha/

(2) 形態素解析器, “茶筌”, Ver.2.3.3 , 奈良先端科学技術大学院大学　松本研究室, http://chasen.org/~taku/software/ChaSen/

(3) クラスタリングツール, “汎用連想計算エンジン（GETA）”, 第二版, http://geta.ex.nii.ac.jp

## 参考文献


[1] D. Marcu, and A. Echihabi, “An Unsupervised Approach to Recognizing Discourse Relations,” Proc. of ACL-02, pp.368-375, 2002.

[2] M. Saito, K. Yamamoto, S. Sekine, and, H. Isaharra, “A system to Solve Language Tests for Second Grade Students,” Proc. of IJCNLP-05, pp.45-50, 2005.

[3] M. Saito, K. Yamamoto, and S. Sekine, “Using Phrasal Patterns to Identify Discourse Relations,” Proc. of HLT-NAACL 2006, pp.133-136, 2006.

[4] 齋藤真実、山本和英、関根聡, 「文間接続関係の自動同定のための人間による同定分析」, NL174-12, pp.65-70, 2006.