3I1-R-9-3

SVMを用いた接続関係の同定

Identifying Discourse Relations Using Support Vector Machine

若山 裕介*1　　内海 彰*1
Yusuke Wakayama　Akira Utsumi

*1電気通信大学大学院 情報理工学研究科 総合情報学専攻
Department of Informatics, The University of Electro-communications

In this paper, we propose a method for identifying discourse relations between two sentences using Support Vector Machine. We use as features for SVM contextual information before and after the target sentences, semantic information such as functional expressions, and syntactic information such as part-of-speech tags. The result of the evaluation experiment showed that our method obtained a higher F-measure score (42.5%) than the existing example-based method.

1. はじめに

文章を作成する能力というのは正確に物事を伝える上でとても重要である．その際，自然な文章にするためには適切な接続詞が必要である．しかし近年，大学生等でも接続詞を上手に扱えていない人が増えている．そこで，計算機が文間の結束関係を正しく表す接続詞が用いられているかを自動的に判断して，可能ならば正しい接続詞を指摘する添削システムが必要となる．

言語処理において接続詞は文間の結束関係を同定するための手掛かり語として使用されることが主であり，接続詞自体を同定する研究は少ない．しかし，先述した添削システムを実現するには接続詞を同定することが必要であるが，同じような関係を表す接続詞どうしの違いが曖昧であることから計算機で同定するのはとても困難である．そのため，接続詞を大まかに分類した接続関係を同定することが求められる．日本語において接続関係を同定する数少ない研究の一つに，用例利用型の手法 [1] がある．この手法では，まずコーパスから接続詞を挟む前後 2 文の述語とそれに係る格要素を収集して，それらをクラスタリングする．次に，生成されたクラスタを用いて判定対象の文と大量に用意された正解文との類似度を計算し，その類似度が最も高かった正解文に使われている接続関係をシステムの解とする．しかし，この手法では扱っている情報が述語と述語にかかる格要素の 2 種類のみであることや接続詞を挟む前後各 1 文のみしか扱わないことから接続関係が一意に決まらないことも多々ある．

英語においては手掛かり語として接続詞を用いずに結束関係を同定する研究は存在する．Pitler ら [2] の研究では，機械学習手法の一つである Naive Bayes を用いて，接続関係（比較，随伴，展開，時間）に有効な素性について調査している．ここで用いている素性は動詞，単語のペア，既存の言語モデルにあるかなどの表層的な素性と極性判定，質問形かどうか，数値的情報があるか，モダリティなどの文法や意味的な素性である．

Marcu ら [3] の研究では，大規模なテキストデータから Naive Bayes を用いてセグメント間の 4 種類（逆説，因果・並列，条件，累加）の接続関係と 2 種類（同じテキストから取り出したもの，違うテキストから取り出したもの）の接続関係を持たないものの計 6 種類を用いて，6 種類の中から 2 つの関係を選び，すべての組み合わせについて二値分類を行っている．ここで用いている素性は，2 つのセグメントから取り出した単語対であり，大量のコーパスから取り出した単語対の情報が分類性能を向上させることを示している．さらに，コーパスの量が同じなら単語対に用いる品詞を限定した方が精度が良くなることを示している．

Sporleder ら [4] の研究では Marcu の研究を受けて表層的な情報だけでなく，対象とする文の出現位置，文の長さ，単語の bigram，品詞，時制などを用いて機械学習手法の一つである boosting による同定を行っている．ここでは，5 種類の接続関係（contrast, explanation, result, summary, continuation）を対象とし多値分類を行っている．また，有効な素性についての検証も行っている．

しかし，これらの研究では扱っている文の数が接続関係を挟む前後 2 文や 2 つのセグメントであるため，文脈等を考慮していないという問題点がある．

そこで，本研究では文脈を考慮するために扱う接続詞を挟む前後各 1 文よりも多くの情報（前後各 2 文，前 1 文後 2 文等）を用いている．さらに，接続関係を一意に決めるために，多値分類手法の一つである one vs. rest 法を用いる．接続関係の推定には種々の分類問題において精度が良いと知られている機械学習アルゴリズムの一つである Support Vector Machine（SVM）を用いる．

連絡先: 若山 裕介，電気通信大学大学院 情報理工学研究科 総合情報学専攻，〒 182-8585 東京都調布市調布ヶ丘 1-5-1，ywganjin@utm.se.uec.ac.jp

2. 文間の接続関係

表 1: 接続詞の分類

種類	接続詞の例
並列	一方，もしくは，あるいは，つまり，…
例示	例えば
因果	だから，ゆえに，なので，すると，…
累加	また，そして，さらに，しかも，まずは，…
転換	さて，ところで，では，…
逆説	しかし，だが，でも，なのに，ところが，…

先述したように計算機で接続詞単体を判定することはとても困難である．そこで文献 [1] で使用されている表 1 に示す接

続詞の分類を用いる．なお，この分類に入らない接続詞（例：「おしむらくは」「ついては」）も存在するが，それらの接続詞は文章を書く上で使用する頻度が他の接続詞に比べると低いため，本研究では対象外とする．

3. SVM を用いた分類器の構築

3.1 二値分類

SVM は二値分類のための教師有り学習アルゴリズムの一つである．学習データは式 (1) に示す特徴ベクトルとして表すことができる． $\mathbf{x_j}$ を各事例の特徴ベクトル，y_j を事例 j が正例であるときに $+1$，負例であるときに -1 となる教師信号とすると式 (3) の判別関数を得る．

$$(\mathbf{x_1}, y_1), \cdots, (\mathbf{x_u}, y_u), \quad \mathbf{x_j} \in \mathbf{R^n}, \quad y_j \in \{+1, -1\} \tag{1}$$

SVM は正例・負例間の距離 (マージン) が最大となるような分離平面を決定する．学習事例が線形分離不可能な場合にはスラック変数 ξ_j を導入する．このとき以下のように定式化される．

$$\min_{\mathbf{x},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{j=1}^{u} \xi_j, y_j(\mathbf{w}\cdot\mathbf{x_j} + b) \geq 1 - \xi_j \tag{2}$$

$$g(x) = \sum_{j=1}^{u} \alpha_j y_j (\mathbf{x_j}\cdot\mathbf{x}) + b \tag{3}$$

ここで，α_j は Lagrange 乗数である．式 (3) の符号を用いてテストデータを分類することができる．

3.2 多値分類への適用

元々，SVM は二値分類のための学習手法であるが，多値分類に適用する方法の一つに one vs. rest 法がある．one vs. rest 法では k 個の各クラスに対して，あるクラスか，それ以外かという二値分類器 $g_c(x)$ を，k 個構築する手法である．未知事例 x' に対して分類を行う場合は，$g_c(x')$ の値が最大となる分類器に対応するクラスに決定される．

文献 [5] によると，one vs. rest 法が他の手法（k 個のクラスから任意の 2 つのクラスに対する二値分類器を ${}_kC_2$ 個構築する pairwise 法など）に比べて計算時間がかかるが精度が良いのでこの方法を採用した．

3.3 素性

表 2: 特徴ベクトルに用いた素性

種類		重み	次元数
品詞	名詞-サ変接続	文中に現れる頻度	8865
	動詞	文中に現れる頻度	7600
	助詞	文中に現れる頻度	313
	形容詞	文中に現れる頻度	958
	接続詞*1	文中に現れる頻度	208
文間類似度		類似度 (実数)	1
機能表現の意味情報		1	94
係り受けのパタン情報		1	551888

[1] 判定しようとしている接続詞は除く

まず，接続詞を挟む前後各数文（この数は変更可能）を情報として入手する．これらの文すべての集合を 1 単位とする．各文に形態素解析・係り受け解析・パターンマッチ等の各処理を行い，表 2 に示した素性に基づき 1 単位ごとに特徴ベクトルを構成する．なお形態素解析には MeCab[*2]，係り受け解析には CaboCha[*3]を用いる．

与える情報として，接続詞を挟む前後の文数と素性ごとに前後の区別も付与できるようになっている．例えば，前後 2 文を情報として用いる時，前一文目と前二文目をそれぞれ区別する場合，前文と後文の 2 つに分ける場合，さらに前文と後文の区別もしないという 3 通りが考えられる．

以下に表 2 の素性の詳細を述べる．品詞の情報としては，名詞，動詞，助詞，形容詞，接続詞（判定しようとしている接続詞は除く）を用いた．また，名詞に関してはシソーラス[*4] を用いて，名詞をある階層に分類することも考える．

文間類似度は接続詞を挟む前後の文に含まれる単語の $tf\cdot idf$ によるベクトル（素性による特徴ベクトルとは関係ない）の類似度のことである．文ベクトルを d_i, d_j とすると，類似度 $S(d_i, d_j)$ は以下のようになる．

$$d_i = (\omega_{i1}, \omega_{i2}, \cdots, \omega_{iM})^T, \omega_{ij} = tf_{ij}\cdot idf_j \tag{4}$$

$$tf_{ij} = \text{文 } i \text{ 中に出現する単語 } j \text{ の頻度} \tag{5}$$

$$idf_j = \log \frac{\text{総文数}}{\text{単語 } j \text{ が出現する文数}} + 1 \tag{6}$$

$$S(d_i, d_j) = \frac{\Sigma\omega_{1j}\omega_{2j}}{\sqrt{\Sigma\omega_{1j}^2}\sqrt{\Sigma\omega_{2j}^2}} \tag{7}$$

「機能表現の意味情報」とは，機能表現辞書「つつじ」[6] に存在する,「立場」「許可」などの意味情報のことである．これらの意味情報は全部で 94 種類ありそれらに該当する機能表現が文に存在した場合，該当する意味情報の素性が 1 となる．

係り受けの構文パタンとは，文を係り受け解析しその文のつながりをパタン化したものである．具体的には以下の手順で構文パタンを求める．

1. 文を係り受け解析し，各文節ごとに文節末が助詞，助動詞，すべての品詞で「非自立」であるものと動詞の「ある」についてのみ抽出する．その他の品詞についてはワイルドカードにする．
2. 上記で求めた文節パタンのすべての係り受けの組み合わせを求め，それらを構文パタンとする．

4. 評価

4.1 使用したコーパス

毎日新聞の 3 年分のデータを使用した．まず，接続詞を挟む前後数文を入手しその接続詞を表 1 のように接続関係として分類した．学習用データとして表 1 の接続関係ごとに 3000 文の計 18000 文用意し，評価用データとして各 100 文の計 600 文用意した．

4.2 有効な素性の検証

3.3 節で示した素性が接続関係において重要であるかを調べる実験を行った. ただし，素性の組み合わせを総当たりで実験することは実質不可能であるため, 素性の種類ごとの傾向を調べた．

実験結果を以下に示す.

品詞

*2 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
*3 http://code.google.com/p/cabocha/
*4 日本語語彙大系，岩波書店 (1997)

表 3: 組み合わせた素性一覧（○：前後で完全に区別する，△：前後で大別する，□：前後で区別しない，空欄：用いない）

種類	扱う文数	品詞					文間類似度	機能表現の意味情報	係り受けの構文パタン
		名詞-サ変	動詞	助詞	形容詞	接続詞			
素性集合 1	前後各 2 文	△	△	△	△	△	△	□	○
素性集合 2	前 1 文後 2 文	△	△	△	△	△	△	□	○
素性集合 3	前 1 文後 2 文	△	△	△	△	△	△	□	△
素性集合 4	前後各 1 文	△	△				△	□	
素性集合 5	前後各 2 文	△	△				△	□	

品詞の種類を初期値として「サ変名詞，動詞」とし「接続詞，形容詞，助詞」をそれぞれ追加した．結果として，品詞の種類を追加すればするほど精度の向上が見られた．

考慮する文数の変更

判定する接続詞を挟んで，前後各 1 文，前後各 2 文，前後各 3 文，前 1 文後 2 文，前 2 文後 1 文を試した．結果として，前後各 1 文と前後各 3 文，前 2 文後 1 文での精度は悪かったが，前後各 2 文や前 1 文後 2 文では精度の向上が見られた．

機能表現の意味情報

判定する接続詞に対する機能表現の位置情報（前/後）を考慮するかしないかを試した. その結果，位置情報を考慮しない方が精度の向上が見られた．

シソーラスによるサ変名詞の分類

シソーラスを用いてサ変名詞を分類した方が精度が高くなるかどうかと，分類した際ある階層に統一するときにどの階層が精度が良いかを検討した．また，ある単語が統一する階層よりも上位にある場合は単語自体を 1 単位として考えた．結果としては，シソーラスを考慮しない方が精度が高いことがわかった．

係り受けの構文パタン

係り受けの構文パタンを追加するかしないかと追加した場合は以下の 4 つの組み合わせを試した．

- 前後に完全に区別する
- 前後で大きく区別する
- 前後で区別しない
- 扱っている文数に関係なく前 1 文後 1 文のみ

結果として，接続関係の種類により精度が良い組み合わせが違うことがわかった．

4.3 評価方法

4.2 節で得られた傾向を基に，精度が良くなる組み合わせの中で各接続関係において最も精度が高くなった素性の組み合わせを表 3 に示す．

評価方法としては，まず表 3 の素性を用いて学習用データのベクトルを作成し，SVM を用いて各接続関係の分類器を構築した．構築した各分類器を用いて評価用データを 6 種類の接続関係に分類し，もとの接続関係と一致するかどうかの判定を行った．また，比較手法として，用例利用型による分類手法 [1] でも分類を行った．ただし，この手法では扱っている素性が少ないため，接続関係が一意に決まらないことがある．そこで，一意に分類できたデータのみを評価対象とする場合と一意に分類できなかったデータの接続関係をランダムに決める場合の評価を行った．一意に決まらなかったデータは 600 件中 335 件であり，一意の場合はそれらを除外した計 265 文で評価を行った．

なお，SVM のプログラムは TinySVM を使用した．

5. 結果と考察

結果を表 4 に示す．表中の値は F 値である．F 値は再現率と適合率の調和平均のことであり，以下の式で表せる．

$$F\text{ 値} = \frac{2PR}{P+R} \tag{8}$$

$$\text{適合率 } P = \frac{\text{システム正解数}}{\text{システム出力数}} \tag{9}$$

$$\text{再現率 } R = \frac{\text{システム正解数}}{\text{正解数}} \tag{10}$$

表 4 を見るとすべての接続関係で提案手法の方が精度が良かった．また，接続詞の種類によって精度が良くなる素性が異なることがわかる.「例示」や「因果」では前後各 2 文を素性として用いるより前 1 文後 2 文とした方（素性集合 2，素性集合 3）が良い結果が得られている．これは「例示」の場合，例を述べている文は接続関係の後であるため前の文の情報はあまり必要ではないと考えられる.「因果」の場合でも理由を述べる部分は接続関係の後に詳しく述べるからであると考えられる．そのことから，これらを組み合わせた方法を考えた.「素性集合 1+2」は「例示」かどうかの判定に関してのみ素性集合 2 の情報を用い，その他の接続関係は素性集合 1 の情報を用いたものであり,「素性集合 1+2+3」は「例示」に関しては素性集合 2 を「因果」に関しては素性集合 3 を用い，その他の接続関係は素性集合 1 の情報を用いたものである.「例示」に対してのみ異なる素性を用いる素性集合 1+2 は素性集合 1 を比べると平均の精度が高くなり，さらに，すべての手法の中で精度が最も高くなった．つまり「例示」の場合，二つ前の文が判断の際にノイズとなってしまい精度が低くなると考えられる．

しかし,「因果」に対しても追加した場合（素性集合 1+2+3）は素性集合 1 を比べると因果自体の精度は良くなったものの平均の精度は悪くなった．これは，素性集合 3 では「因果」以外の分類器の精度が悪いことから相対的に「因果」の精度が良くなっただけであり，素性集合 1 では他の分類器の精度が高いことから分類器の精度が相対的に下がってしまった言える．

また，全体の接続関係では前後各 1 文のみ学習させた素性集合 4 と前後各 2 文に情報量を追加したのみの素性集合 5 を比較すると，これだけの情報のみで精度がとても上がっていることが分かる．つまり，接続関係は少なくとも前後各 1 文のみでは判定が難しいということが言える．

ここで，素性集合 1 では判定できなかったが素性集合 1+2 では判定できた例を図 1 に示す．この新聞記事では，前一文

表 4: 提案手法と用例利用型（比較手法）[1] での結果（F 値）

手法	並列	例示	因果	累加	転換	逆説	平均	判定可能な文書数（割合）
提案手法 素性集合 1	**0.529**	0.368	0.356	0.402	**0.509**	**0.447**	0.437	600/600 (100%)
提案手法 素性集合 2	0.505	**0.466**	0.357	0.304	0.437	0.420	0.418	600/600 (100%)
提案手法 素性集合 3	0.371	0.398	**0.447**	0.384	0.447	0.410	0.413	600/600 (100%)
提案手法 素性集合 4	0.212	0.193	0.296	0.188	0.266	0.204	0.230	600/600 (100%)
提案手法 素性集合 5	0.414	0.371	0.204	0.262	0.269	0.277	0.300	600/600 (100%)
提案手法 素性集合 1+2	**0.529**	0.453	0.376	**0.416**	0.500	0.441	**0.452**	600/600 (100%)
提案手法 素性集合 1+2+3	0.527	0.420	0.371	0.386	0.440	0.413	0.426	600/600 (100%)
用例利用型 一意	0.517	0.337	0.300	0.409	0.205	0.418	0.365	265/600 (44.1%)
用例利用型 ランダム	0.350	0.230	0.241	0.273	0.232	0.267	0.266	600/600 (100%)

注）太字は各接続関係ごとに精度が一番良いものを表す.

前一文目 死闘を繰り返す彼らのどっちの側にカードを切るか、多様な選択肢を独占し、もてあそぶ「永田町」の影がちらついている。

前二文目 最近のドラマは権力の構造を遠近法でみつめ、そこに想像力を賭ける作品がめだつ。

後一文目 例えば『踊る大捜査線』（フジ）や『ケイゾク』（TBS）、先夜の『刑事たちの夏』（NTV 系）などの手法がそれだ。

後二文目 複雑で分かりにくい筋立てを RPG（ロールプレイング・ゲーム）ふうなゲーム感覚で読み取ることを求めるドラマといっていい。

図 1: 素性集合 1 と素性集合 1+2 における判定精度向上の例

目をみるとドラマのことを表していないように思われる．しかし，前一文目のみを見ると，前二文目からのドラマの話とはあまり関係がないように思われる．このように例示に関しては前一文目は考慮しない方が精度が向上することがわかる．

また，英語における Sporleder ら [4] の研究の結果では，explanation（説明）と continuation（継続）の精度が良い．しかし，contrast（対照），result（結果），summary（要約）では比較的精度が良くない．本研究でも並列や例示の精度は比較的良い結果が得られており，因果では精度があまり良くないことから，日本語と英語でも同じような傾向があることがわかる．

さらに Spolder らの研究では素性ごとに有効な情報も調べている．そこでは，word や stem といった表層的な素性や接続関係を挟んで左側の情報が右側の情報よりも重要であると述べている．本研究では接続関係を挟んだ前後情報を広げることにより精度が向上することを示している．

6. おわりに

本研究では，接続関係の同定を機械学習手法の一つである SVM によって行う手法を提案し，実験によってその有効性を評価した．結果として，全体としての精度は良くないが，従来の用例利用型による手法 [1] よりも優れた結果が得られた．

本研究の結果から接続関係によって素性の種類が異なることが分かったため，今後の課題としては各接続関係に特化した手法を考えることが必要となる．また，計算機による自動添削を実現するためには，接続詞のより詳細な分類に対応することも課題の一つである．

参考文献

[1] 山本和英，斎藤真美：用例利用型による文間接続関係の同定，自然言語処理，Vol.15，No.3, pp.21–51 (2008).

[2] Emily Pitler, Annie Louis and Ani Nenkova : Automatic sense prediction for implicit discourse relations in text, *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp.683–691 (2009).

[3] Daniel Marcu and Abdessamed Echihabi : An unsupervised approach to recognizing discourse relations, *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp.368–375 (2001).

[4] Caroline Sporleder and Alex Lascarides : Exploiting Linguistic Cues to Classify Rhetorical Relations, *Proceedings of Recent Advances in Natulral Language Processing (RANLP-05)*, pp.532-539 (2005).

[5] 山田寛康，松本裕治：Support Vector Machine の多値分類問題への適用法について，情報処理学会研究自然言語処理研究会報告，pp.33–38 (2001).

[6] 松吉俊，佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え，自然言語処理，Vol.15，No.2，pp.75–99 (2008).