教師なし学習を用いた談話解析手法に関する一検討

伊藤 直貴† Hugo Hernault† 石塚 満†
† 東京大学大学院情報理工学系研究科

1 はじめに

談話解析とは，文と文の間に成り立つ接続関係を同定する手法である．これは，文章の意味を理解するために重要な要素技術であるとともに，要約技術や機械翻訳，対話生成といったアプリケーションに用いられる．

これまでに，談話解析の手法として，コーパスによる機械学習を用いた手法 [1] が提案されている．しかし，日本語において，大規模な談話情報の付与されたコーパスは存在しない．このため，独自にコーパスを作成し，教師あり学習を行う手法 [2] や大規模な Web 上のテキストを検索し，独自にスコア付けを行った用例利用型の手法 [3] が提案されている．しかし，独自に作成したコーパスは，小規模なために高い精度が得られておらず，また，大規模な Web 上のテキストをコーパスとしても，そのすべてにスコア付けをするには，膨大な計算量がかかるといった問題がある．

そこで，我々は大規模なテキストから，接続詞毎に接続関係を仮定し，それをもとに機械学習を行う Marcu らの手法 [4] を日本語に適用する．さらに，名詞のカテゴリ情報を用いて，素性数を大幅に削減させ，かつ分類性能を向上させる手法を提案する．機械学習を用いることで，分類の際に大量のコーパスを参照する必要はなく，素性数を削減することができれば比較的高速に分類できるのではないかと考えられる．本稿では，Marcu らの手法を，日本語に適用した実験の結果と提案手法を用いた実験の結果を示し，提案手法の有効性について検討する．

2 関連研究

2.1 教師なし学習を用いた手法

談話構造解析用のコーパスが与えられていない場合に，談話構造解析を行う手法として Marcu らの手法 [4] がある．この手法では，まずコーパスから “But” や “Although” といった手がかり語の前後 2 つの文のペアを取得する．その後，手がかり語を取り除き，ペアの文の間には手がかり語に応じた接続関係 (この場合は「CONTRAST」) があると仮定する．Marcu らが用いた素性は，前と後ろそれぞれの文に出現する単語対すべてである．また，Marcu らは，限られた語 (名詞，動詞，接続詞) のみを素性に用いることで，精度が向上する可能性があることを示している．

2.2 用例利用型にを用いた手法

山本ら [3] は，機械翻訳の分野で用いられる用例利用型の手法により，接続関係の同定をしている．これは，Web 文書から収集した 120 万件の候補文の中から，入力文に最も類似する文を選び，その接続関係により，入力文の接続関係を同定する手法である．山本らは，まず入力文を構文解析し，その構文パタンによる類似度のスコアが高い順に候補文を絞り，絞られた候補文の中で，構文パタンと単語による類似度のスコアと併せて，最も類似する文を選択している．

A Study on Classification of Discourse Relations with Unsupervised Method
†Graduate School of Information Science and Technology, The University of Tokyo

3 機械学習を用いた接続関係の同定

3.1 データセット

今回用いたテキストデータは，Web 上の日本語で書かれたテキストである．まず，収集したテキストを句点で分割する．分割された文の中で，日本語形態素解析器 MeCab[1] による品詞タグ付けによって「接続詞」と識別された単語を含む文が抽出される．接続詞を文頭に持つ文は，その 1 つ前の隣接する文とペアを構築し，接続詞を文中に持つ文は，接続詞の前と後ろで分割し，これらが 1 つのペアとなる．

得られた文のペアから，抽出の際に用いた接続詞を取り除き，同時に，これらのペアに取り除かれた接続詞に応じた接続関係を付与する．用いた接続詞の一部と接続関係の対応を表 1 に示す．今回は，接続詞がペアに複数含まれている場合でも，そのうちの 1 つの接続詞だけを削除し，これにより付与する接続関係を決定することとした．このため，1 つの文が複数の接続関係に含まれる場合がある．また，「なかんずく」「そーいや」「ほなら」などの対応付けが難しいものは，接続関係の付与を行っていない．

3.2 素性の抽出

サポートベクタマシン (Support Vector Machine: SVM) を用いて，前節で得られた文のペアに付与された接続関係を同定する．SVM ライブラリとして，LIBSVM[2] を用いた．SVM は機械学習手法であるため，接続関係を分類するための素性を人手で与える必要がある．

3.2.1 Marcu らの用いた素性

我々は，Marcu ら [4] の手法を参考にしており，まず MecCab によって「名詞」または「動詞」と識別された単語の組を素性とする．すなわち，文のペアに対し，「前の文の名詞－後ろの文の名詞」および「前の文の動詞－後ろの文の動詞」が素性となる．

3.2.2 名詞のカテゴリ情報による素性

テキストの意味を解析する分野では，語の関係の近さが素性として用いられる．この素性は，2 つの文の含意関係を決定するのに有効であると考えられている．そこで，我々は，語の関係の近さを，同じカテゴリに含まれているかどうかであるとして，単語にいくつかのカテゴリを付与し，その情報を素性として用いる手法を提案する．

Marcu らの手法を用いると，名詞の素性数が多くなるため，我々の手法では，文のペアに出現する名詞のカテゴリ情報を談話解析に用いて，名詞の素性数を削

表 1: 接続関係と接続詞の例

接続関係	接続詞の例
順接	すると, なので, そうしたら, こうして
対立	でも, 然しながら, なのに, 逆に
添加	及び, しかも, かつ, と同時に
対比	または, さもなければ, 若しくは
同列	すなわち, 例えば, 実は, つまり
補足	ただし, ちなみに, もっとも

[1] http://mecab.sourceforge.net/
[2] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

減する．ここで，カテゴリとは，MeCab の出力フォーマットに含まれる名詞の品詞細分類 2 である．今回は，名詞の品詞細分類 2 のうち「人名・地域・組織・助数詞・一般・特殊」のみを対象とし，これらをカテゴリと呼ぶこととする．また，カテゴリ情報が付与されている名詞をカテゴリ名詞と呼ぶこととする．文のペアに対し，「前の文のカテゴリ名詞－後ろの文のカテゴリ名詞」および「前の文に出現するカテゴリ－後ろの文に出現するカテゴリ」が素性となる．

4 実験および評価

4.1 分類性能評価

我々は，各接続関係に属する文のペアをそれぞれ 1000 ずつ収集し，前節で述べた素性による接続関係の分類性能を調べる実験を行った．各接続関係のデータのうち，900 ペアを用いて学習し，残りの 100 ペアをテストとして評価した．LIBSVM のパラメータは，デフォルトのままであり，カーネルは，線形カーネルを用いた．実験結果として，Recall，Precision，F 値について，6 種類の接続関係の平均をとった値を表 2 に示す．

カテゴリ情報を用いない場合は，名詞と動詞の両方を素性とした場合に，F 値が最大で 0.62 となる．また，名詞のみを素性とした場合と，動詞のみを素性とした場合を比較すると，動詞のみを素性とした場合のほうが F 値が高い．しかし，名詞と動詞それぞれの素性数を比較すると，動詞の素性が 18,488 個であるのに対し，名詞の素性は 1,013,272 個と膨大である．我々は，この膨大な名詞の素性数を，カテゴリ情報を用いることで削減した．

カテゴリ情報を用いた場合は，カテゴリとカテゴリ名詞のみを素性に用いるため，素性数が大幅に少なくなっている．さらに，前述の名詞のみを使った場合と比較して，F 値が 0.77 と高くなっていることも分かる．また，カテゴリ情報と動詞を用いた場合は，カテゴリ情報のみを用いた場合より，さらに F 値が高く，0.78 である．カテゴリ情報を用いない場合と比べ，素性数を削減でき，また分類性能も向上した．

ここで，カテゴリ情報と動詞を用いた場合の分類結果について，接続関係毎の値を示したものを図 1 に示す．この図より，「対比」における Precision は高く，その代わりに Recall が低くなっていることが分かる．これは，対比以外の接続関係が付与された入力に対し，対比に誤分類される割合が高いことが原因である．

4.2 素性の重み

次に，カテゴリ情報を用いることで，分類に用いられる素性の重みがどのように変化したかを調べる実験を行った．機械学習ツールの Classias[3]を用いて，再度学習し，その際の特徴量の重みが高い素性を，関係の深い接続関係とともに出力した．この際用いた学習アルゴリズムは，L1 正規化ロジスティック回帰である．各接続関係で最高の重みをもつ素性を図 3 および 4 に示す．素性のうち，“( )” で囲まれたものは，カテゴリを表す．

これより，カテゴリ情報を用いない場合は，名詞が重要な素性と見なされておらず，動詞の方が重要な素性と見なされている．一方，カテゴリ情報を用いた素性を用いた場合は，「沙也加」と「法子」といったカテゴリ名詞の組や「(一般)」といったカテゴリが重要な素性となっていることが分かる．

表 2: 分類性能 (全接続関係の平均) および素性数

素性	F 値	素性数
名詞	0.49	1,013,272
動詞	0.61	18,488
名詞＋動詞	0.62	1,031,760
カテゴリ情報	0.77	153,747
カテゴリ情報＋動詞	0.78	172,645

[3]http://www.chokkan.org/software/classias/

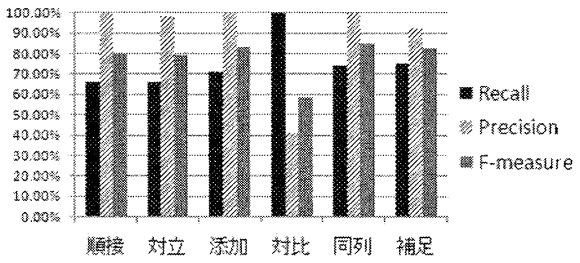

図 1: カテゴリ情報＋動詞の接続関係ごとの分類性能

表 3: 名詞＋動詞の重要素性

	素性
順接	募る，沿う
対立	及ぶ，言う
添加	止める，忘れる
対比	解す，認める
同列	恐れる，愛す
補足	する，する

表 4: カテゴリ情報＋動詞の重要素性

	素性
順接	沙也加，法子
対立	郷，ヨーコ
添加	日，食い
対比	ふう，申し出る
同列	森，大輔
補足	(一般)，(一般)

5 考察

F 値が向上するとともに，カテゴリが重要な素性となっていることから，カテゴリ情報は，接続関係を分類する際に有効な素性であると考えられる．

カテゴリとして，「人名」や「地域」といったものが一般的であると考えられるが，これらのカテゴリの分類への寄与は低かったと見られる．今回，用いた 6 種類の接続関係のすべてにおいて，人名や地域のカテゴリが前と後ろの両文に出現したペアの割合は 15% 以下であるのに対し，「一般」のカテゴリが両文に出現したペアの割合は，75% 以上であり，この点が影響したと考えられる．

6 おわりに

本稿では，まず日本語における談話解析の手法として，英語による従来手法を，日本語に適用した実験の結果を示した．そして，カテゴリ情報を付与した素性による手法を提案し，その有効性を実験により確認した．今後は，より多くのカテゴリを用いた実験をするとともに，クラスタリングによって自動でカテゴリを作成するといったことを行う．

参考文献

[1] D.A. duVerle and H. Prendinger, “A novel discourse parser based on support vector machine classification,” Proc. of the ACL and the IJCNLP of the AFNLP, pp. 665–673, 2009.

[2] 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.

[3] 山本 和英, 齋藤 真実, “用例利用型による文間接続関係の同定,” Journal of Natural Language Processing, vol. 15, no. 3, pp. 21–51, 2008.

[4] D. Marcu and A. Echihabi, “An unspervized approach to recognizing discourse relations,” Proc. of the ACL, pp. 368–375, 2002.