4S-3

WordNetを用いた機械学習による談話構造解析

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

談話構造解析は，入力された2文の間に存在する関係を同定する問題である．これまでに，談話構造解析の手法として，コーパスによる機械学習を用いた手法が提案されている[1]．

しかし，未だ多くの手法において，無作為に抽出した単語のペアを素性やスコアに用いている．単語のペアは，談話構造解析において重要な素性である一方，出現するすべての単語のペアを素性とすることで，冗長さや悪影響が大きくなる．特に，機械学習においては，分類の際に計算量を増大させ，overfitting の問題を引き起こす．

そこで，我々は，日本語 WordNet[1]を用いて，素性抽出を行う手法を提案する．提案手法は，出現する単語が意味する概念を用いる．各概念は，上位の概念に抽象化され，素性として用いられる．概念の上位下位関係は，日本語 WordNet 内に定義されている階層構造 (Taxonomy) を利用して得られる．抽象化した素性を用いることで，素性数を削減することが可能である．

2 関連研究

2.1 学習データを作成する手法

日本語において，大規模な談話構造の付与されたコーパスは存在しない．このため，横山らは人手で学習データを作成し，教師あり学習を行う手法を提案している[2]．この手法は，談話構造理論の一つである修辞構造を SVM を用いて学習する手法であるが，人手で作成したデータは，小規模なために高い精度が得られないといった問題が残っている．

このように談話構造の付与されたコーパスが入手できない場合において，学習データを自動で作成する方法として，Marcu らの手法がある[3]．この手法では，まず大規模なテキストから，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談話関係 (この場合は「contrast (逆接)」) があると仮定する．

2.2 用例利用型を用いた手法

Marcu らの手法を日本語に適用した手法として，山本ら[4]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，談話関係の同定をしている．これは，Web 文書から収集した約120万件の文の中から，入力文に最も類似する文を選び，その類似文の談話関係により，入力文の談話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となる文を絞り，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．

3 機械学習を用いた談話構造解析

我々が行う談話構造解析の手順を，図1に示す．入力された2文から，形態素解析器 MeCab[2]による出力結果の基本形を取得する．これらを単語とし，前後の文から得られた単語のペアを生成する．その後，我々は，得られた単語を WordNet の概念と結び付け，同時に Taxonomy を用いて概念を抽象化する．

分類器は，得られた単語および概念から成る素性を用いて，入力された2文がどの談話関係に属しているかを推定する．

3.1 談話関係

我々は，MeCab の辞書 (IPADIC，ver.2.7.0) に登録されている接続詞170語中152語を，6種類の談話関係に分類した．談話関係と対応する接続詞の一部を表1に示す．ある接続詞の前の文が，文頭に接続詞を持つ場合，この前文は，さらに1つ前の文ともペアを構築する．このため，1つの文が

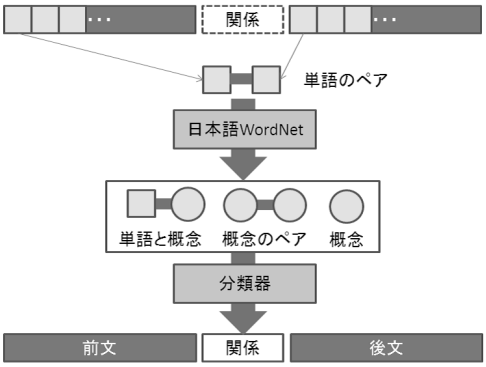

図 1: 処理手順

表 1: 談話関係と接続詞の例

談話関係	接続詞の例
順接	すると，なので，そうしたら
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

複数の談話関係に含まれる場合がある．また，「なかんずく」「追って」「恐れながら」などの対応付けが難しい接続詞や，「てか」「さて」といった話題の転換を表現する接続詞は用いていない．

3.2 WordNet を用いた語義曖昧性解消

単語は，WordNet 内で複数の概念 (意味) をもつ場合がある．このため，我々は，2種類の手法を用いて，単語と1つの概念を結び付ける語義曖昧性解消 (WSD: word sense disambiguation) を行う．

1つ目の手法は，WordNet に定義されている頻度点 (frequency score) を用いる手法である．頻度点は，単語がある概念をもって文章に出現しやすいという情報を，概念ごとにスコア付けしたものである．我々は，頻度点の最も大きい概念を，その単語の概念としている．

もう1つの手法は，2つの概念の定義文における単語の重なり (gloss overlap) を用いる手法である．我々は，定義文を MeCab により解析し，同一の単語 (名詞，動詞，形容詞，および副詞の基本形) が，2つの概念の定義文に同時に出現した回数を，gloss overlap としている．ある単語がもつ概念の中で，同一文内の他の単語がもつ概念との gloss overlap の合計が最大となる概念が，その単語の概念となる．

3.3 Taxonomy を用いた素性抽出

WordNet においては，名詞と動詞において，上位下位関係の Taxonomy が定義されている．我々は，前後の文から得られた単語を，Taxonomy を利用して抽象化する手法を提案する．

3.3.1 Minimum path CS

Minimum path common subsumer (MPCS) は，2つの概念の Taxonomy における共通の祖先の中で，概念間のパス長が最小となる経路上に存在する祖先を用いる手法である．MPCS は，2つの概念を同時に1つの概念 (祖先) に抽象化することになる．また，MPCS においては，単語と概念を結び付ける必要はない．これは，各単語がもつすべての概念の組み合わせに対して，パス長を計算するためである．

A supervised approach for discourse analysis using WordNet
†Graduate School of Information Science and Technology, The University of Tokyo
[1]http://nlpwww.nict.go.jp/wn-ja/
[2]http://mecab.sourceforge.net/

3.3.2 相対抽象化

相対抽象化 (RA: relative abstraction) は，図 2 のように，ある概念を Taxonomy において n 段階上位の概念に抽象化する手法である (A→A')．ただし，上位 0〜$(n-1)$ 階層にある概念は，その Taxonomy の根 (root) となる概念に抽象化される．

3.3.3 絶対抽象化

絶対抽象化 (AA: absolute abstraction) は，図 3 のように，ある概念を Taxonomy において根から一定の距離 n にある概念に抽象化する手法である．ただし，n より近い位置，すなわち上位 0〜$(n-1)$ 階層にある概念は，抽象化されない．

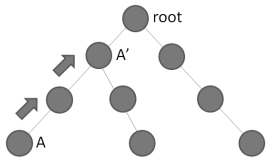

図 2: RA の例

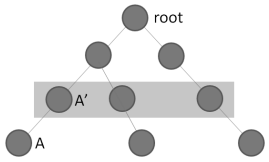

図 3: AA の例

4 実験

4.1 データセット

用いたテキストデータは，Wikipedia 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，MeCab 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文頭に持つ文が抽出される．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する．

各談話関係毎 1000 ペア，合計 6000 ペアのデータを用いて，10 分割交差検定により性能を評価した．

4.2 分類モデルの比較

まず我々は，複数の分類モデルを比較した．実験に用いたモデルは，SVM(線形カーネル，RBF カーネル) およびロジスティック回帰である．RBF カーネルは，データセットを用いて，あらかじめ c および g のパラメータ調節を行っている．値はそれぞれ，$c = 2048.0$，$g = 3.0517578125 \times 10^{-5}$ である．各モデルは，ツール LIBSVM[3]，LIBLINEAR[4]，Classias[5]を用いて実装した．素性は，前後の文の単語のペアである．

結果を表 2 に示す．F-measure は，各談話関係の Recall と Precision のマイクロ平均を関係毎に計算し，これらの調和平均を全関係で平均したものとする．

RBF カーネルは，性能がパラメータの設定により約 17%〜37% の間で大きく推移した．山本ら [4] は比較手法として SVM を用いているが，性能が 3 割に満たないと報告している．これは，パラメータの設定が不十分であったのではないかと考えられる．

LIBLINEAR は非常に高速であり，また，性能面で最も良かったもの (Classias: 40.7，LIBLINEAR: 40.3) と大きな差がない．このため我々は，以降の実験で LIBLINEAR を用いることとした．

4.3 提案手法の評価

次に我々は，Taxonomy を用いた抽象化を評価した．前後の文の単語のペアを取得し，各単語および単語のペアが，提案手法により抽象化が可能な場合，これらを抽象化した．

表 2: 分類モデルの比較

Model	Tool(kernel)	F-measure
SVM	LIBSVM(Linear)	38.7
	LIBSVM(RBF)	37.0
	LIBLINEAR(Liner)	40.3
Logistic Regression	Classias	40.7

[3]http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[4]http://www.csie.ntu.edu.tw/ cjlin/liblinear/
[5]http://www.chokkan.org/software/classias/

表 3: 提案手法の分類性能と素性数

Method	n	F-measure	Dimension
Random	-	16.7	-
Word pair	-	**40.3**	1,722,605
MPCS	-	**40.7**	1,328,869
RA(Freq)	1	39.1	1,298,969
RA(Overlap)	1	38.1	1,424,688
AA(Freq)	5	**39.3**	1,160,442

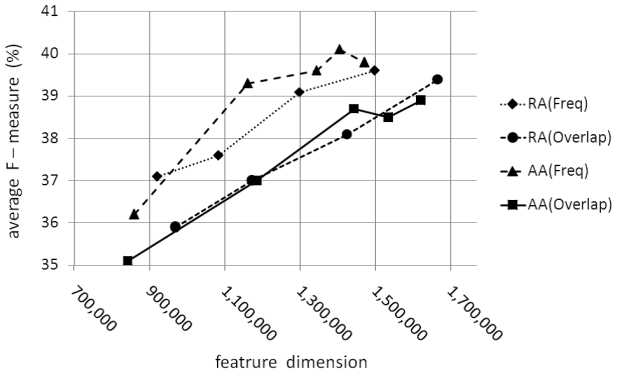

図 4: n を変化させた時の RA および AA の性能

結果を表 3 に示す．本表において，Freq は，頻度点による WSD を用いた場合を表し，Overlap は，gloss overlap による WSD を用いた場合を表す．また，太字は，有意水準 5%($\alpha = 0.05$) の両側検定において有意差がない結果であることを表す．

各手法は，ランダムに談話関係を選択した場合 (16.7%) に比べて，いずれも優位な結果を示した．

相対抽象化は，他の手法に比べて，性能が低下した (RA(Freq): 39.1，RA(Overlap): 38.1)．相対抽象化においては，2 つの単語の階層が異なっていた場合，抽象化後の階層も異なってしまう．このため，相対抽象化は，素性数が削減されず，図 4 に示すように，素性数の減少に伴う性能の低下が著しい．

MPCS および絶対抽象化は，単語のペアを用いた場合と比べて，性能が同等であることが分かる (MPCS: 40.7, AA(Freq): 39.3)．一方で，素性数を比べると，提案手法は，単語のペアに比べて，素性数が少ないことが見てとれる (MPCS: 1,328,869，AA(Freq): 1,160,442)．出現頻度による WSD を用いた絶対抽象化 (AA(Freq)) は，素性数を 32.6% 削減しつつ，単語のペアを用いた場合と，同等の性能を示している．

5 おわりに

本稿では，機械学習による談話構造解析において，日本語 WordNet を用いて単語を概念に結び付け，その概念を抽象化することで，素性を削減する手法を提案した．実験により，提案手法は，無作為に単語のペアを利用する場合に比べ，性能を維持しつつ，素性を削減していることが確認された．

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] D. Marcu and A. Echihabi, “An unspervized approach to recognizing discourse relations,” Proc. of the Association for Computational Linguistics, pp.368–375, 2002.

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