談話構成素とその文脈による教師なし談話構成素構造解析

西田 典起　中山 英樹
東京大学大学院情報理工学系研究科

{nishida, nakayama}@nlab.ci.i.u-tokyo.ac.jp

1 はじめに

自然言語の統語規則および統語構造を統計的手法によって明らかにしようとする文法推定 (grammar induction) は，1990 年代から自然言語処理において盛んに研究されてきた [8, 3, 6, 7, 1]．これらの研究においては主に文の句構造や依存構造がその解析対象とされてきた．しかし，言語を言語たらしめている言語規則は文境界だけに留まらず，より広いレベルでも存在すると考えられる．

本研究は，文の境界を越え，談話レベルでの言語構造およびその傾向を発見するための手法を提案し，実験によって定量的かつ定性的に評価，分析することを目的とする．談話構造に関しては，談話構造解析において現在最も一般的である修辞構造理論 (Rhethorical Structure Theory; RST) [9] に基づいた木構造を仮定する．すなわち，与えられた談話が首尾一貫しているならば，それはその談話の主張に基づいて一つの木構造をなし，各終端ノードは最小談話ユニット (Elementary Discourse Unit; 節等) を表し，各非終端ノードは連続する談話ユニットを有向の談話関係 (逆説等) で再帰的に結合したものとする．本稿では RST の談話構造を構成する要素のうち，最小談話ユニットへの分割は既に済んでいるものとし，談話の構成素構造に焦点を当てる．

手法としては，Klein ら [5] によって提案された文法推定手法である Constituent-Context Model (CCM) に注目する．CCM の基本アイデアは言語学で用いられてきた構成素テストを単純化したものであり，CCM ではある終端シンボルの系列が構成素ならば，それは構成素的文脈において現れるはずだと考える．

本稿では CCM を教師なし談話構成素構造解析のために拡張する．ここでは便宜的にそれを Discourse Constituent-Context Model (DCCM) と呼ぶ．DCCM は各談話区間 (構成素候補) の連続的ベクトル表現を区間内と区間外 (文脈) それぞれのテキスト領域から計算し，それに基づいて各談話区間のスコア (構成素らしさ) を学習する．談話木構造のスコアはそれを構成する各構成素の区間スコアの総和として定義し，CKY と同様の動的計画法によってスコア最大となる木構造の大域探索を行う．DCCM のパラメータは Viterbi EM アルゴリズムによって逐次的に最適化する．

EM ベースの文法推定は，学習できる文法の傾向が初期のパラメータに強く依存してしまうという弱点がある．そこで本稿では，談話構造に関する事前知識に基づいてより有効な初期化と負例情報の導入を行うために，組み合わせ逐次解析器という木構造サンプリング手法を提案する．

定量実験の結果，DCCM がベースラインを上回ることを観測した．また，提案した初期化および負例サンプリングが重要であることも確認した．また，DCCM が言語的に妥当な談話構成素性をいくつか発見できていることを定性的に分析した．

2 手法

2.1 談話構成素構造 (木構造) のスコア

DCCM による談話構成素解析は，入力文書 $d = e_0, \ldots, e_{n-1}$ (最小談話ユニット e_k の系列) に対し，スコア最大となるような木構造 $\hat{T}$ を探索する問題と見なせる:

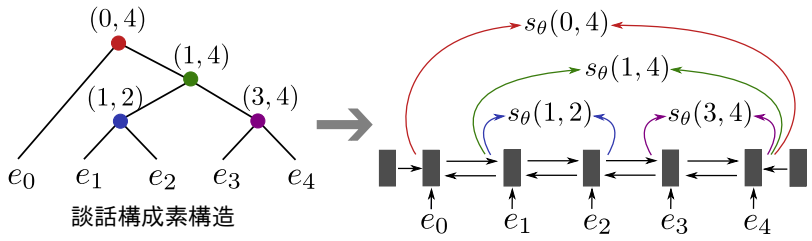
$$\hat{T} = \underset{T \in \mathrm{valid}(d)}{\mathrm{argmax}} \; s(T). \tag{1}$$

ここで $s(T) \in \mathbb{R}$ は木構造 T のスコアを表し，$\mathrm{valid}(d)$ は d の木構造として可能なものの集合とする．

木構造のスコア $s(T)$ は T を構成する各談話構成素の区間スコアの総和として定義する:

$$s(T) = \sum_{(i,j) \in T} s_\theta(i, j). \tag{2}$$

ここで s_θ は，談話区間 (i, j) のテキスト領域 $e_{i:j} = e_i, \ldots, e_j$ に対してそれの構成素らしさを計算するス



区間	構成素表現	文脈表現
$(1,2)$	$[\overrightarrow{h}_2 - \overrightarrow{h}_0; \overleftarrow{h}_1 - \overleftarrow{h}_3]$	$[\overrightarrow{h}_0; \overleftarrow{h}_3]$
$(3,4)$	$[\overrightarrow{h}_4 - \overrightarrow{h}_2; \overleftarrow{h}_3 - \overleftarrow{h}_\blacklozenge]$	$[\overrightarrow{h}_2; \overleftarrow{h}_\blacklozenge]$
$(1,4)$	$[\overrightarrow{h}_4 - \overrightarrow{h}_0; \overleftarrow{h}_1 - \overleftarrow{h}_\blacklozenge]$	$[\overrightarrow{h}_0; \overleftarrow{h}_\blacklozenge]$
$(0,4)$	$[\overrightarrow{h}_4 - \overrightarrow{h}_\diamond; \overleftarrow{h}_0 - \overleftarrow{h}_\blacklozenge]$	$[\overrightarrow{h}_\diamond; \overleftarrow{h}_\blacklozenge]$

図 1: 提案手法における談話構成素構造 (木構造) のスコア計算の概要図.

コア関数であり，θ はそのパラメータである．図 1 に $s(T)$ の計算過程の概要図を載せる．

まず，解析対象となる文書 d 中の各最小談話ユニットを独立にベクトル化する:

$$\boldsymbol{e}_k = \mathrm{ReLU}(\boldsymbol{W}_e \sum_{w \in e_k} \boldsymbol{w} + \boldsymbol{b}_e). \tag{3}$$

ここで，$\boldsymbol{w}$ は単語 w のベクトル表現とし，$\boldsymbol{W}_e$ と $\boldsymbol{b}_e$ はそれぞれ線形変換のための行列とバイアスベクトルとする．

次に，談話区間スコア $s_\theta(i,j)$ を計算するための区間表現を計算する．方法としては Gaddy ら [4] の区間ベース構文解析器にならい，まず双方向 LSTM を先に求めた最小談話ユニットベクトルの系列に対して適用し，各ステップにおける前方向，後方向ベクトル $\overrightarrow{\boldsymbol{h}_k}, \overleftarrow{\boldsymbol{h}_k}$ を計算する．そして，談話区間 (i,j) の構成素表現および文脈表現をそれぞれ次のように前方向，後方向ベクトルの結合とする:

$$\boldsymbol{r}_{i,j} = [\overrightarrow{\boldsymbol{h}_j} - \overrightarrow{\boldsymbol{h}}_{i-1}; \overleftarrow{\boldsymbol{h}_i} - \overleftarrow{\boldsymbol{h}_{j+1}}], \tag{4}$$

$$\boldsymbol{c}_{i,j} = [\overrightarrow{\boldsymbol{h}_{i-1}}; \overleftarrow{\boldsymbol{h}_{j+1}}]. \tag{5}$$

CCM [5] にならい, DCCM では談話区間のスコアを二つの独立なスコアに分解する:

$$s_\theta(i,j) = f_\theta(\boldsymbol{r}_{i,j}) + g_\theta(\boldsymbol{c}_{i,j}). \tag{6}$$

ここで，f_θ はテキスト領域 $e_{i:j}$ からその構成素らしさを計算し，一方で g_θ はその文脈 $(e_{0:i-1}, e_{j+1:n-1})$ から間接的に $e_{i:j}$ の構成素らしさを計算する．本稿では 活性化関数が ReLU の二層パーセプトロンによって f_θ および g_θ をそれぞれ実装した．

2.2 最適な木構造の探索

DCCM では，CKY と同様の動的計画法によってスコア最大となる談話構成素構造の大域探索を行う．区間 (i,j) をカバーする部分木のスコアを

$$\mathrm{C}[i,j] = s_\theta(i,j) + \max_{i \le k < j} (\mathrm{C}[i,k] + \mathrm{C}[k+1,j]) \tag{7}$$

のように再帰的に定義する．入力文書の談話構成素構造を解析するには，まずボトムアップに $\mathrm{C}[0,n-1]$ を求め，それから再帰的に選択された分割点 k をトップダウンに辿っていけばよい．

2.3 学習

学習は Viterbi EM によって行う．Viterbi EM では次の対数尤度を逐次的に最大化する:

$$\hat{\theta} = \operatorname*{argmax}_\theta \sum_d \log P_\theta(\hat{T}^\theta \mid d). \tag{8}$$

ここで，$\hat{T}^\theta$ はパラメータ θ に基づいて求めたスコア最大の木構造である．Viterbi EM では内向き外向きアルゴリズムによる従来の EM とは異なり，E ステップでは現在のパラメータに基づいてスコア最大の木構造を求め，M ステップではこの木構造に基づいて勾配降下法等によってパラメータを更新する．

実際には，2.4 で説明するように，本稿では負例情報をより明示的に導入するために式 (8) の代わりに次のヒンジ損失関数を最小化した:

$$\sum_{T':\ T \neq T'} \max\left(0, s\left(T'\right) + \Delta\left(T, T'\right) - s\left(T\right)\right). \tag{9}$$

ここで $T' \neq T$ は文書 d に対する負例木構造であり，$\Delta(T,T')$ は木構造間の距離とする．木構造の距離は

$$\Delta\left(T, T'\right) = |T| - |T \cap T'| \tag{10}$$

のように定義した．すなわち，$\Delta(T,T')$ は T と T' の間で異なる構成素の個数を表す．

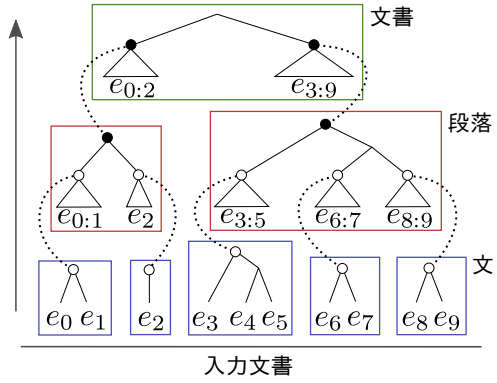


図 2: 組み合わせ逐次解析器では，談話の階層性と分岐傾向に基づいてボトムアップに逐次的に木構造を構築する．

2.4 初期化と負例木構造サンプリング

EM ベースの文法推定は (現在のパラメータに基づく) E ステップと (E ステップの結果に基づく) M ステップを交互に繰り返すため，学習できる文法の傾向が初期のパラメータに強く依存してしまうという弱点がある．そこで，文，段落，文書という談話の階層性とそれぞれの分岐傾向という事前知識に基づいて，より有効な初期化と明示的な負例情報の導入を行うために，組み合わせ逐次解析器 (Combinational Incremental Parser; CIP) という木構造サンプリング手法を提案する．

初期化に関しては，ランダムに初期化したパラメータ θ から Viterbi EM を開始する代わりに，CIP によってサンプリングされた談話構成素構造をもとに M ステップを繰り返し行い，そこで得たパラメータを Viterbi EM の初期値とする．式 (9) の負例木構造 T' も同様に，異なる設定の CIP によってサンプリングする．

CIP の概要図を図 2 に示す．CIP では，まず入力文書を自動的に抽出された文境界に基づいて分解し，各文ごとに手法 m_s によって談話構成素構造の解析を行う．そして，同様にして抽出された段落境界に基づいて，各段落ごとに手法 m_p によって解析を行う．ここで，各解析における終端ノードはその段落を構成する各文の解析結果 (木構造) とする．最後に，各段落の解析結果を受けて，手法 m_d によって文書全体の解析を行う．本稿では各 CIP を $\langle m_s, m_p, m_d \rangle$ と表記する．

ベースとなる解析手法 m としては，教師なし構文解析における既存研究 [5, 6] を参考に Random BottomUp (BU), Random TopDown (TD), Right Branching (RB), Left Branching (LB) の 4 つを定義した．これら 4 つの詳細については紙面の都合上割愛する．

負例サンプリング	UP (%)	UR (%)	F_1 (%)
$\langle \mathrm{LB}_s, \mathrm{LB}_p, \mathrm{RB}_d \rangle$	57.6	59.9	58.7
BU	59.0	61.3	60.1
BU + B_s	59.3	61.6	60.4
BU + (B_s, B_p)	57.3	59.6	58.4
Mixture	**60.7**	**63.0**	**61.9**

表 1: DCCM の学習における負例サンプリング方法の比較．B_s, B_p はそれぞれ文境界，段落境界情報を表す．

3 実験結果と考察

3.1 コーパスと評価方法

実験用のコーパスとしては，WSJ の記事に対し RST に基づいた談話構造がアノテーションされている RST Discourse Treebank (RST-DT) [2][1] を用いた．

定量評価として，文法推定における既存研究に従い [5, 6, 7]，各解析器の出力が言語学者の考える談話構成素構造 (RST-DT におけるアノテーション) に比べてどれほど近いかを評価した．評価尺度としては Unlabeled Precision (UP) と Unlabeled Recall (UR), またそれらの Micro F_1 スコアを用いた．

3.2 結果と考察

まず，CIP の全組み合わせ ($4^3 = 64$ 通り) について比較を行った．その結果，最もスコアの高い組み合わせは $\langle \mathrm{RB}_s, \mathrm{RB}_p, \mathrm{LB}_d \rangle$ ($F_1 = 60.1$) であり，逆に最もスコアの低い組み合わせは $\langle \mathrm{LB}_s, \mathrm{LB}_p, \mathrm{RB}_d \rangle$ ($F_1 = 50.4$) であり，約 10 ポイントもの差があることがわかった．尚，境界情報を用いない場合の最高スコアは 19.3 (BU)，最低スコアは 7.6 (RB, LB) であった．以上の結果から，本稿では $\langle \mathrm{RB}_s, \mathrm{RB}_p, \mathrm{LB}_d \rangle$ を DCCM の初期化に用いた．

次に，表 1 に示す 5 通りの負例サンプリング方法について比較実験を行った．ここで “Mixture” は BU,

[1] https://catalog.ldc.upenn.edu/LDC2002T07

手法	UP	UR	F_1
$\langle BU_s, BU_p, BU_d \rangle$	54.4	56.5	55.5
$\langle TD_s, TD_p, TD_d \rangle$	53.7	55.8	54.8
$\langle RB_s, RB_p, RB_d \rangle$	57.9	60.2	59.0
$\langle LB_s, LB_p, LB_d \rangle$	50.5	52.5	51.5
$\langle LB_s, LB_p, RB_d \rangle$	49.4	51.4	50.4
$\langle RB_s, RB_p, LB_d \rangle$	59.0	61.3	60.1
DCCM (RB w/o boundaries)	55.9	58.1	57.0
DCCM ($\langle RB_s, RB_p, RB_d \rangle$)	58.6	60.8	59.7
DCCM ($\langle RB_s, RB_p, LB_d \rangle$)	**60.7**	**63.0**	**61.9**

表 2: DCCM およびベースライン (CIP) による談話構成素構造のスコアの比較．“DCCM” の横の括弧内は初期化手法を表している．

BU + B_s, BU + (B_s, B_p) の 3 通りによる負例集合を混合した場合を表す．表 1 の結果から，負例サンプリングにおいては境界情報を用いない方が良いことがわかった．これは，サンプリングされる負例木構造の多様性が重要であることを示している．さらに，興味深いことに，“Mixture” のように境界情報を活用しない結果と活用した結果を混合することでスコアが向上することが確認された．これは，負例集合の多様性だけでなく，負例としてのモデルにとっての難易度 (正例っぽさ) のバランスが重要であることを示している．以上の結果を受けて，本稿では “Mixture” を DCCM の学習における負例サンプリング方法として用いた．

次に，DCCM とベースライン (CIP) の比較を行った．表 2 (および表 1) から，本稿で提案した初期化および負例サンプリングを行うことで，DCCM がベースラインを上回ることが観測できた．特に，DCCM の初期化で用いた $\langle RB_s, RB_p, LB_d \rangle$ を上回ったという結果は，Viterbi EM を通して言語的に妥当な談話構成素性を幾ばくか学習できたことを示している．

最後に，DCCM によって学習された談話構成素の傾向を定性的に分析した．RST-DT のテストセットに含まれるすべての談話区間 (i, j) $(i \neq j)$ について，その区間スコア $s_\theta(i, j)$ を求め，スコアの大きさに基づいてランク付けし，スコアが相対的に高い談話区間を確認した．ページ数制限の都合上具体的な結果の記載は割愛するが，DCCM は特に (1) 右から左の節を明示的な連結詞とともに修飾するようなテキスト表現 (e.g., “[X [because Y.]]”) と (2) Discourse GraphBank [10] 等でも談話関係として定義されている `Attribution` (e.g., “[X say that [Y]]”) に関して相対的によくその談話構成素らしさを学習できていることがわかった．一方で，より複雑で長区間にまたがる談話構成素についてはまだまだ上手く学習できておらず，今後はこの点の改善方法を検討していく必要がある．

4 おわりに

本稿では，談話構成素構造の教師なし解析手法 (DCCM) を提案し，またその学習に有効な初期化および負例情報の導入を行うための木構造サンプリング手法 (CIP) を提案した．実験の結果，提案した初期化と負例サンプリングを行うことで DCCM がベースラインを上回り，また言語的に妥当な談話構成素らしさをいくらか発見できることを確認した．

謝辞

本研究成果は，独立行政法人情報通信研究機構 (NICT) の委託研究「 多言語音声翻訳高度化のためのディープラーニング技術の研究開発 」により得られたものです．

参考文献

[1] Y. Bisk and J. Hockenmaier. Probing the linguistic strengths and limitations of unsupervised grammar induction. 2015.

[2] L. Carlson, D. Marcu, and M. E. Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *SIGDIAL'01*, 2001.

[3] A. Clark. Unsupervised induction of stochastic context-free grammars using distributional clustering. In *CoNLL'01*, 2001.

[4] D. Gaddy, M. Stern, and D. Klein. What's going on in neural constituency parsers? An analysis. In *NAACL-HLT'18*, 2018.

[5] D. Klein and C. D Manning. Natural language grammar induction using a constituent-context model. In *NIPS'01*, 2001.

[6] D. Klein and C. D. Manning. A generative constituent-context model for improved grammar induction. In *ACL'02*, 2002.

[7] D. Klein and C. D. Manning. Corpus-based induction of syntactic structure: Models of constituency and dependency. In *ACL'04*, 2004.

[8] K. Lari and S. J. Young. The estimation of stochastic context-free grammars using the Inside-Outside algorithm. *Computer Speech and Language*, Vol. 4, pp. 35–56, 1990.

[9] W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, Vol. 8, No. 3, pp. 243–281, 1988.

[10] F. Wolf and E. Gibson. Representing discourse coherence: A corpus-based study. *Computational, Linguistics*, Vol. 31, No. 2, pp. 249–287, 2005.