

テキストの局所一貫性に基づく半教師あり暗黙的談話関係認識

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

談話関係認識では、隣接する文間の関係性 (e.g., Comparison, Cause) の同定を行う。連結詞 (e.g., *however*, *because*) が文中に顕在している場合、これらの連結詞が強力なヒントとなるため談話関係の同定は比較的容易である [9]。一方、連結詞が顕れない文間の暗黙的談話関係の認識は、依然として困難な課題である。

低精度の原因として、そのタスクの難しさに対し教師データの数が不足していることが挙げられる。そのアノテーション作業には高度な言語知識が要求されるため、暗黙的談話関係認識のためのラベル付きデータを収集する負担は大きい。¹ 本研究の目標は、暗黙的談話関係の認識に有効な言語知識 (言語表現) を生コーパスから教師なし学習によって獲得することである。

本稿では、テキスト一貫性の教師なし学習が談話関係に関する言語知識の獲得に有効であるという仮説に基づいた半教師あり学習手法を提案する。テキストが首尾一貫しているならば、そのテキスト中の文同士には論理的、統語的、または意味的な関係性があると考えられる。すなわち、文間の関係性に関する情報は、テキストの一貫性を識別するのに有効な特徴であり、本研究ではテキスト一貫性 (教師なし) 学習を通して談話関係に関する特徴表現を学習することを目指す。学習された言語表現は、半教師あり学習の形式に沿って暗黙的談話関係認識に知識転移する。また、テキスト一貫性に関する二つの設計案を提案し、比較を行う。

実験結果から、テキスト一貫性を学習することによって暗黙的談話関係認識のマクロ F 値が第一レベル関係クラスで約 3%、第二レベル関係タイプで 5% 以上改善することが確認された。特に、Comparison や Temporal 等の教師データが他よりも少ない談話関係の認識改善に有効であることがわかった。また、トピック指向の局所一貫性に関する学習が順序指向の場合を

上回る傾向にあることがわかった。

2 関連研究

暗黙的談話関係認識のための教師データを拡張する方法として様々なアプローチがこれまで提案された。Marcu と Echihabi [8] は、連結詞を含む隣接文のペアを検出し、連結詞を除去することで暗黙的談話関係認識のための擬似データを生成する手法を提案した。しかし、これらの生成データで学習されたモデルは期待されるような精度の改善を達成できないことが知られている [12]。これは、生成されたデータとオリジナルのデータとのドメイン不一致によって引き起こされると考えられている。このドメイン不一致の問題を軽減する方法として、マルチタスク学習によって生成データとオリジナルデータを別々に扱うアプローチ [6, 7, 5] や、取り除いて良い連結詞とそうでない連結詞を区別するアプローチ [11]、ドメイン適合によってドメイン間距離を小さくするアプローチ [4] 等が提案された。また、連結詞ベースの単語埋め込みを学習し、半教師あり学習によって知識転移を行う手法も提案された [2, 13]。

これらの先行研究は、いずれも生コーパスにおける連結詞の検出処理を必要としており、コーパス中の連結詞が顕れない多くの文を活用できていない。² 一方、本稿で提案するテキスト一貫性に基づくアプローチは明示的な連結詞の検出処理を必要とせず、生コーパス全体を活用して言語知識の獲得を可能とする。

連結詞の検出に依存しない半教師あり手法もいくつか提案されているが [3, 1]、これらの研究は線形学習器を使用しているか、あるいは語彙レベルでの言語知識の転移を行っている。一方、本稿で提案する手法は 3.2 節で記述する非線形モデルによって文または談話レベルの知識転移を行っている。

¹ 談話関係認識のデータセットとして現在最大規模とされる Penn Discourse Treebank (PDTB) 2.0 コーパス [10] においてさえ、ラベル付きデータは 16,224 個のみである。

² 例えば、British National コーパスの約半数の文は連結詞を含んでおらず [12]、本研究の手法によってこれらの文も活用ができる。

3 提案手法

3.1 テキストの局所一貫性識別

本稿では、テキスト一貫性を局所文脈における確率の積としてモデル化する。すなわち、テキスト X が一貫している確率を

$$P(\text{coherent} | X) = \prod_{x \in X} P(\text{coherent} | x) \quad (1)$$

とする。ここで、 $P(\text{coherent} | x)$ は局所テキスト x が一貫している確率であり、3.2 節で記述する方法によってモデル化する。局所テキスト x は中心文とその前後の文脈文とを合わせたタプル $x = (s_-, s, s_+)$ と定義する。文脈窓幅を広げることでより広い文脈に基づいた言語知識の獲得が可能になると予想されるが、それは学習時の計算コストの増大およびデータスパースネス問題とトレードオフの関係にある。

テキスト一貫性識別のためのデータセット $\mathcal{D} = \mathcal{P} \cup \mathcal{N}$ は生コーパスから自動的に構築した。ここで、 \mathcal{P} と \mathcal{N} はそれぞれ正例集合と負例集合を表す。 N 文からなるコーパス $\mathcal{C} = s_1, s_2, \dots, s_N$ が与えられたとき、正例集合は次の方法で構築する：

$$\mathcal{P} = \{(s_{i-1}, s_i, s_{i+1}) \mid i \in [2, N-1]\}. \quad (2)$$

テキストの局所一貫性は二つの観点から破綻されうる。これは負例集合 \mathcal{N} をどのように構築するかに対応する。一つ目は文順序に関する局所一貫性の破綻であり、次の順序指向負例サンプリングによって表される：

$$\mathcal{N} = \{x' \mid x' \neq x \wedge x' \in \phi(x) \wedge x \in \mathcal{P}\}. \quad (3)$$

ここで $\phi(x)$ は x の順列集合を表す。二つ目はトピックに関する局所一貫性の破綻であり、次のトピック指向負例サンプリングによって表される：

$$\mathcal{N} = \{(s_-, s', s_+) \mid s' \in \mathcal{C} \wedge (s_-, s, s_+) \in \mathcal{P}\}. \quad (4)$$

ここで s' はコーパス \mathcal{C} 全体からランダムにサンプリングされた文とする。局所テキスト (s_-, s, s_+) で共有されているトピック連続性は、ランダムサンプリングされた s' によって満たされないと期待できる。

3.2 システム全体と半教師あり学習

本稿ではニューラルネットワークを用いてシステム全体を構築する。全体の概要図を図 1 に示す。システムは文符号化器 E 、局所一貫性分類器 F_c 、暗黙的談話

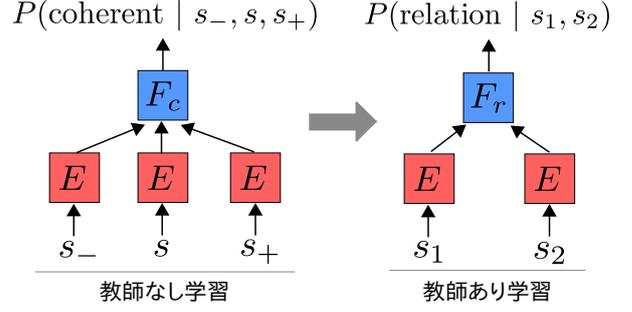


図 1: 本稿で構築したシステム全体の概要図。文符号化器 E 、局所一貫性分類器 F_c 、談話関係分類器 F_r の三つのモジュールから構成される。まずテキストの局所一貫性識別の教師なし学習を行い、その後に暗黙的談話関係認識の教師あり学習を行う (i.e., 半教師あり学習)。

関係分類器 F_r の三つのモジュールから構成される。まずテキストの局所一貫性識別の教師なし学習を行うことでモジュール E と F_c のパラメータを最適化し、次に暗黙的談話関係認識の教師あり学習によって E と F_r のパラメータを最適化する。ここで E は両タスクで共有され、暗黙的談話関係認識に有効な文情報の抽出をテキスト一貫性識別を通して学習することが目的である。

文符号化器 E は入力文をベクトル表現に変換する。まず双方向 LSTM によって L トークンからなる入力文 $s = w_1, \dots, w_L$ を処理する：

$$\vec{h}_t = \text{FwdLSTM}(\vec{h}_{t-1}, w_t) \in \mathbb{R}^D, \quad (5)$$

$$\overleftarrow{h}_t = \text{BwdLSTM}(\overleftarrow{h}_{t+1}, w_t) \in \mathbb{R}^D. \quad (6)$$

ここで FwdLSTM と BwdLSTM はそれぞれ前方向 LSTM と後方向 LSTM を表す。初期ステップにおける隠れ状態はゼロベクトルで初期化した： $\vec{h}_0 = \overleftarrow{h}_{L+1} = \mathbf{0}$ 。予備実験において従来のプーリング関数 (e.g., 総和プーリング, 平均プーリング, 最大化プーリング) を試みたが、前方向 LSTM と後方向 LSTM の最終ステップにおける隠れ状態を連結する方法が最も有効であった：

$$h = \left(\vec{h}_L^\top, \overleftarrow{h}_1^\top \right)^\top \in \mathbb{R}^{2D}. \quad (7)$$

したがって本稿では式 (7) を集約関数として用いた。

テキスト局所一貫性分類器 F_c と暗黙的談話関係分類器 F_r は二つの多層パーセプトロンによって構成する。活性化関数としては ReLU を用い、出力層にはソフトマックス関数を適用した。入力層の次元数はそれぞれ $2D \times 3$, $2D \times 2$ であり、層の数は 2 層とした。

	第一レベル関係クラス		第二レベル関係タイプ		テキスト一貫性
	正解率 (%)	マクロ F 値 (%)	正解率 (%)	マクロ F 値 (%)	正解率 (%)
<i>IRel</i> のみ	51.49	42.29	37.49	24.81	-
<i>IRel</i> + <i>O-Coh</i> (小)	52.16	41.39	37.77	25.46	57.96
<i>IRel</i> + <i>O-Coh</i> (大)	52.29	42.48	41.29	30.70	64.24
<i>IRel</i> + <i>T-Coh</i> (小)	51.70	40.84	37.91	25.35	83.04
<i>IRel</i> + <i>T-Coh</i> (大)	53.54	45.03	41.39	29.67	91.53

表 1: テキストの局所一貫性識別と暗黙的談話関係認識の結果. *IRel* は暗黙的談話関係認識, *O-Coh* および *T-Coh* はそれぞれテキストの順序指向またはトピック指向一貫性識別で学習したことを表す. 小, 大はそれぞれ利用された生コーパスによる教師データの相対的サイズを表す: WSJ (37K データ), BLLIP (22M データ).

4 実験

4.1 準備

暗黙的談話関係認識のデータセットとして Penn Discourse Treebank (PDTB) 2.0 コーパス [10] を用いた. 従来のセクション分割に従い, セクション 2-20 を訓練用, セクション 0-1 を調整用, セクション 21-22 をテスト用とした. 対象とする談話関係に関しても従来研究に従い, 第一レベル関係クラス (4 ラベル) または第二レベル関係タイプ (11 ラベル) の多値分類によって実験を行った.

テキスト一貫性識別のためのデータセットは生コーパスから構築した. 生コーパスとしては, Wall Street Journal (WSJ) コーパス³ または BLLIP North American News Text コーパス (complete)⁴ を用いた. これらのコーパスをもとに, 3.1 節で述べた方法によってデータセットを構築し, その結果約 45,000 個 (WSJ) または 23,000,000 個 (BLLIP) からなる正例集合が得られた. WSJ の方では, 暗黙的談話関係認識と同様の方法でセクション分割を行い, それに従って調整用とテスト用の正例集合を用意した. BLLIP の方では, ランダムにそれぞれ 10,000 個の正例を選択することで調整用, テスト用の正例集合とした. 調整用, テスト用データの負例に関しては 3.1 節の方法によって正例データと同数だけ作成し, 訓練用の負例に関しては学習時の各イテレーションにおいて動的に構成した.

単語埋め込みと双方向 LSTM の隠れ状態, 多層パーセプトロン分類器の隠れ層の次元数を調整し, それぞれ 100, 200, 100 次元とした. 単語埋め込みはテキスト一貫性識別および暗黙的談話関係認識の学習より以前に BLLIP と GloVe によって学習し, それ以後は固定とした. パラメータの最適化は AdaDelta によって

行った. ミニバッチサイズは 44 とし, ミニバッチ内でのクラス数の分布が均等になるようにミニバッチを動的に構成した.

4.2 結果

テキスト一貫性の教師なし学習が暗黙的談話関係認識を改善するのか検証するために, 提案手法とベースラインモデル (i.e., 暗黙的談話関係認識のみ) とのスコアの比較を行った. 評価尺度としては従来研究に従い正解率とマクロ F 値を用いた. 表 1 に示されるように, テキスト一貫性識別を学習することで暗黙的談話関係認識のマクロ F 値が第一レベル関係クラスで約 3 %, 第二レベル関係タイプで 5 % 以上改善することが確認された. 正解率においても提案手法はベースラインモデルを上回ることがわかった. また, テキスト一貫性識別におけるスコアが高いほど (小 vs. 大), 暗黙的談話関係認識のスコアも高くなることが確認された. これらの結果から, テキスト一貫性を学習することによって談話関係認識に有効な言語知識が獲得されたと考えられる.

トピック指向および順序指向の一貫性識別を比較すると, 特に大規模な生コーパス (i.e., BLLIP) を用いる場合, トピック指向が順序指向を暗黙的談話関係認識において上回る傾向にあることが確認できる. これは, 順序指向の局所一貫性識別がトピック指向の問題に比べて難しく, その結果トピック指向に比べて汎用性のある知識の獲得に失敗しているためと予想される. 実際, 順序指向の局所一貫性識別の学習はトピック指向の学習よりも早く低い精度で収束してしまうことが確認された.

次に, 参考のために, 本研究と同様に外部コーパスを活用している既存研究との比較を行った. 表 2 に結果を載せる. これらの手法が明示的に検出された連結

³<https://catalog.ldc.upenn.edu/LDC99T42>

⁴<https://catalog.ldc.upenn.edu/LDC2008T13>

	正解率 (%)	マクロ F 値 (%)
Rutherford と Xue (2015) [11]	57.10	40.50
Liu ら (2016) [7]	57.27	44.98
Braud と Denis (2016) [2]	52.81	42.27
Wu ら (2017) [13]	58.85	44.84
<i>IRel</i> only	51.49	42.29
<i>IRel</i> + <i>T-Coh</i>	53.54	45.03

表 2: 本稿のシステムと同様に外部コーパスを用いている先行研究との比較結果. 第一レベル関係クラスの多値分類における評価.

	Exp.	Cont.	Comp.	Temp.
学習データ数	6,673	3,235	1,855	582
<i>IRel</i> のみ	67.11	51.94	33.79	28.30
<i>IRel</i> + <i>T-Coh</i>	66.95	51.82	37.12	29.11

表 3: Expansion, Contingency, Comparison, Temporal の談話関係に関する one-vs-others 二値分類の結果と, 各談話関係に該当する教師データの個数. 評価尺度はマクロ F 値とした.

詞に基づいているのに対し (2 章参照), 本稿の提案手法ではそのような連結詞の検出処理が不要であり, よりシンプルなパイプラインであるにも関わらず, 既存研究のマクロ F 値を僅かに上回ることが確認された.

さらに, テキスト一貫性モデルの学習の各談話関係ごとへの有効性を調べるために, one-vs-others 二値分類設定においても実験を行った. すなわち, 第一レベル関係クラス (i.e., Expansion, Contingency, Comparison, Temporal) それぞれを正例とし, それ以外を負例にしたときのモデルの学習と評価を行った. 表 3 に示されるように, テキスト一貫性の学習は, 教師データが他よりも少ない Comparison および Temporal の暗黙的談話関係の認識において有効であり, 教師データ量の不足を補う働きがあると考えられる.

5 おわりに

本稿では, テキストの局所一貫性モデルの教師なし学習による, 暗黙的談話関係認識のための半教師ありアプローチを提案した. 実験結果から, テキスト一貫性の学習によって暗黙的談話関係の認識が向上し, 特に教師データの数が少ない談話関係において有効であることが確認された. また, トピック指向の局所一貫性識別が順序指向を上回る傾向にあることがわかった.

テキスト一貫性の学習によって獲得される言語知識

が, 従来の連結詞ベースの手法によって得られる知識とどのような関係にあるのか, またトピック指向および順序指向で獲得される言語知識にはどのような特徴と差異があるのかについては依然としてわかっておらず, 今後検証, 分析していく必要があり, これらを今後の研究課題とする.

参考文献

- [1] Chloé Braud and Pascal Denis. Comparing word representations for implicit discourse relation classification. In *EMNLP 2015*, 2015.
- [2] Chloé Braud and Pascal Denis. Learning connective-based word representations for implicit discourse relation identification. In *EMNLP 2016*, 2016.
- [3] Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In *EMNLP 2010*, 2010.
- [4] Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. Closing the gap: domain adaptation from explicit to implicit discourse relations. In *ACL 2015*, 2015.
- [5] Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-yu Niu, and Haifeng Wang. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *EMNLP 2017*, 2017.
- [6] Man Lan, Yu Xu, and Zhengyu Niu. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *ACL 2013*, 2013.
- [7] Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. Implicit discourse relation classification via multi-task neural networks. In *AAAI 2016*, 2016.
- [8] Daniel Marcu and Abdessamad Echiabi. An unsupervised approach to recognizing discourse relations. In *ACL 2002*, 2002.
- [9] Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. Easily identifiable discourse relations. In *COLING 2008*, 2008.
- [10] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Milt-sakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The penn discourse treebank 2.0. In *LREC 2008*, 2008.
- [11] Attapol T. Rutherford and Nianwen Xue. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *NAACL-HLT 2015*, 2015.
- [12] Caroline Sporleder and Alex Lascarides. Using automatically labelled examples to classify rhetorical relations: an assessment. *Natural Language Engineering*, Vol. 14, No. 03, 2008.
- [13] Changxing Wu, Xiaodong Shi, Yidong Chen, Jinsong Su, and Boli Wang. Improving implicit discourse relation recognition with discourse-specific word embeddings. In *ACL 2017*, 2017.