

# 対をなす二文書間における文対応推定および応答文生成への応用

角田 孝昭      乾 孝司      山本 幹雄  
筑波大学大学院 システム情報工学研究科  
tsunoda@mibel.cs.tsukuba.ac.jp

## 1 はじめに

文や文書の構造を明らかにすることは、古くから言語学における興味の対象であった。また、計算機が広く使われるようになってからは、文・発話などの種類・相互関係を自動で推定したり、その結果を用いて文書全体の構造を推定する研究も進められている。文及び文書構造を明らかにする動機は言語学的興味のみにとどまらず、文書要約など工学的応用をはじめとする実用的な興味にも基づいている。

本研究ではその中でも文書を横断した文間の関係、特にある文書とそれに対する応答文書の間における文の対応関係を自動で推定することを目的とする。具体的な対象として、明確に対応関係が存在し、一定以上の量を確保できるデータである楽天トラベル<sup>1</sup>のレビュー文書とそれに対する宿泊施設提供者からの応答文書を用いる。この文書対における文間関係の例を図1に示す。例えば「ご朝食につきましては……」と言う応答文は、「朝食メニューが……良かったです。」と言うレビュー文を受けて書かれた文である。このような二文書において文間関係の自動推定が実現すれば文書要約や検索などのタスクにおいて有用である。

さらに、本研究では文対応推定結果を用いた新たな応用の一つとして、文生成の可能性を示す。このために、統計的機械翻訳の枠組みを援用する。すなわち、文書対から推定した文対応関係を「対訳文」とみなすことで、文から応答文への同言語間翻訳によって自動応答文生成を実現する。

本研究の文対応推定のための中心となるアイディアは、文の種類に注目する点である。例えば、「ポジティブな感想を述べる文」と「ポジティブな感想の記述に対する感謝を述べる文」とは対応している可能性が高いなどの傾向がある。そこで、既存手法として提案されている文種類と文対応を推定するそれぞれのモデルを統合し、相互に考慮しながら同時推定が可能なモデルを提案する。提案モデルにより、従来モデルよりも文対応推定の性能が向上したことを示す。

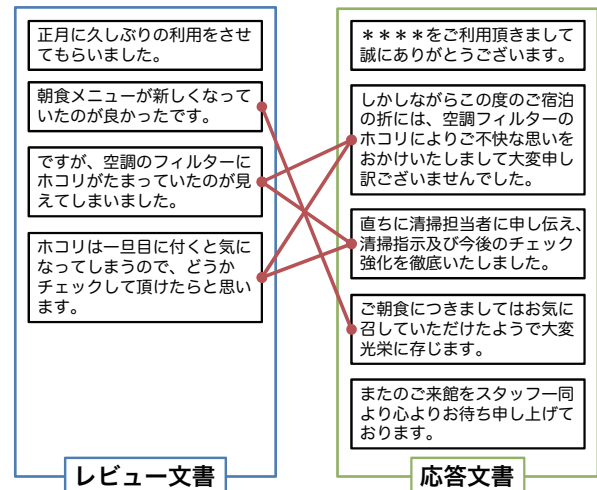


図1: 本研究で扱う文対応関係の例

## 2 先行研究

従来の文書構造の自動推定を目的とした研究は、一文書内の文間・対話の発話間の関係を扱ったものが中心であった。代表的な理論として、一文書の構造に対しては Rhetorical Structure Theory (RST) [Mann and Thompson, 1986]、対話構造に対しては隣接対 (adjacency pair) [Schegloff and Sacks, 1973] が提案されており、これらの理論に則った自動推定の研究も存在する。

一方で、二文書間において文書を横断した文同士の関係を扱ったものは少ない。文関係について分析して理論化した研究には、Radev による Cross-document Structure Theory (CST) がある [Radev, 2000]。この理論では、新聞記事間における文同士の関係を 24 種類に分類している。また、この CST に従った文対応を自動推定する研究には [宮部他, 2006] [難波他, 2005] が存在する。宮部らの研究では、同一のトピックを扱った複数の新聞記事の間における文関係の中でも、特に同等関係及び推移関係の推定に特化した手法を提案している [宮部他, 2006]。また、難波らも、宮部らと同様に新聞記事間を対象とし、特に推移関係の推定に特化した手法を提案している [難波他, 2005]。

以上の研究が対象とする新聞記事と文書-応答文書の違いを考えると、新聞記事間の文関係は主に同一

<sup>1</sup>楽天トラベル: <http://travel.rakuten.co.jp/> の「クチコミ・お客さまの声」。引用にあたり、宿泊施設名は伏せた。また、紙面の都合上、図1では一部の文を省略している。

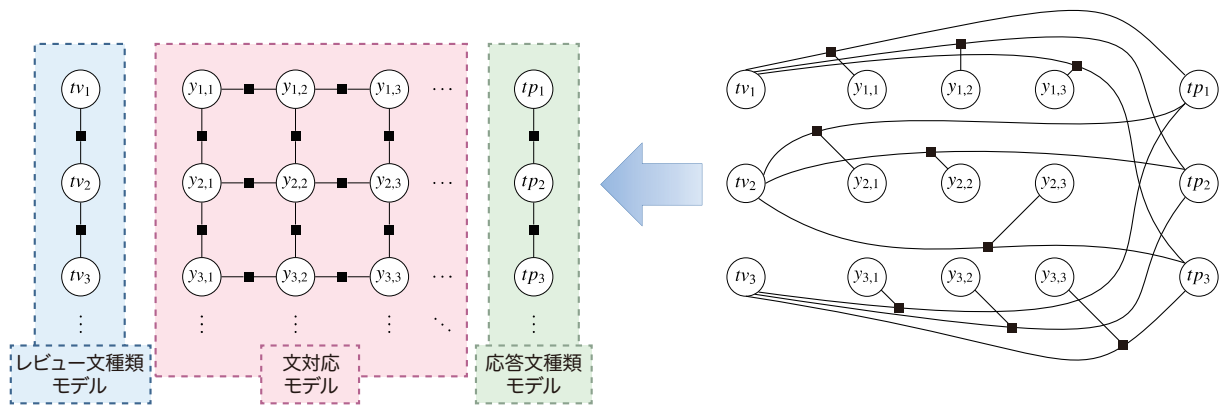


図 2: 比較手法の各モデル (左) と提案手法で追加する因子 (右)

の事象に対して書き手や文脈が変わることで発生するが、文書-応答文書間の文関係はコミュニケーションを達成するために発生すると言う違いがある。そのため、CST の分類や、宮部ら・難波らの手法を本研究に用いることは不適である。

### 3 文対応関係と文種類

#### 3.1 文対応関係

本研究で扱う文関係を定義する。本研究では、「文書とそれに対する応答文書が与えられた時、ある応答文が応答先文書のある文を原因として生起している」関係を文対応関係と呼び、これの有無を推定することを目的とする。なお、文対応関係は多対多関係を許す。

#### 3.2 文種類

本研究では、文対応推定のために文種類を利用した手法を導入する。文種類とは「ポジティブな感想を述べる文」や「お詫びを述べる文」など、文がどのような目的で書かれているかにより分類したものである。

楽天トラベルのレビュー・応答文書における文種類の分類を行なっている先行研究には [大沢他, 2010] がある。本研究ではこれを参考に、文種類が文対応関係の有無を判断するにあたってより有用な情報となるよう再分類し、レビュー文を 12 種類、応答文を 20 種類に分けた。例えばレビュー文種類としては「ポジティブ感想」「ネガティブ感想」など、応答文種類としては「ほめへの感謝」「お詫び」「対応明示」などがある。詳しい分類は紙面の都合上 [角田, 2013] に譲る。

### 4 提案手法

#### 4.1 動機

例えばレビュー文書において、文種類が「ネガティブな感想」「要求・要望」などの文に対しては、いずれかの応答文と対応する可能性が高いという傾向が存在する。同様のことは応答文書にも言える。このた

め、文対応関係の推定において文種類が果たす役割は大きい。

ただし、文種類が分かっている状況は少ない。例えば、対話構造の自動推定を目的とした [Qu and Liu, 2012] では、文種類推定器を別途作成し、対話構造推定時に文種類推定器の出力結果を素性として追加することで対処している。

しかし、文種類推定の精度には限界がある。Qu & Liu の手法では、文種類推定時に複数の文種類の確率値が拮抗している場合でも一番高い確率値である文種類のみを出力しているため、拮抗していたという情報は構造推定には用いられない。ここで我々は、文種類を固定せずに様々な文種類の可能性を文対応の推定時に考慮できれば、一番高い確率値の文種類が誤っていた場合でも推定結果を活用できると考えた。

そこで、本研究では文種類と文対応を推定するモデルを一つに統合した、より文対応推定に特化したモデルを提案する。

#### 4.2 文対応・文種類推定モデルの統合

まず、統合にあたって、本研究では元となるモデルとして文種類推定には Linear-chain CRF [Lafferty et al., 2001] を (レビュー・応答文種類別に用意するので 2 つ)、文対応推定には 2D CRF [Zhu et al., 2005] を用いた場合を考える。これを因子グラフで表した図を図 2 左に示す。

次に、以上の 3 モデルを統合した提案手法のモデルについて説明する。このモデルでは、文種類・文対応推定を同一のモデルで扱うために、新たに二文の文種類変数と、それらに対応する文対応変数を結ぶ因子を加え、各因子関数に対数線形モデルを用いて学習・推定を行う。この因子とそれに関わる枝のみを表した図を図 2 右に示す。

なお、提案手法の因子グラフは閉路を含むため、各隠れ変数の周辺確率  $P(y|x)$  は TRP [Wainwright et al., 2001] などにより近似的に求める。

## 5 実験

### 5.1 実験条件

実験には、「楽天データ公開<sup>2</sup>」に収録されている投稿 348,564 件のうち、レビュー文書・応答文書の双方が存在する 276,562 件からランダムサンプリングした 1,000 文書対を用いた。この各文書を簡易的なヒューリスティックによって文単位に分割し、レビュー文 4,813 文・応答文 6,160 文を得た。また、各文書対に含まれる文対応を目視で確認した結果、計 4,492 通り存在した。この 1,000 文書対に対して 5 分割交差検定を適用して評価を行う。

比較手法として、レビュー・応答文種類モデルには Linear-chain CRF を用いる。また、文対応モデルには図 2 左で示す行ごと・列ごとに Linear-chain CRF を適用したモデル（それぞれ L-CRF<sub>review</sub>、L-CRF<sub>reply</sub> と呼ぶ）、2D CRF を適用したモデルを用いる。

各モデルのパラメータ学習・利用には MALLET 2.0.7 [McCallum, 2002] を用いた。また、用いる素性として、文種類の推定には「文を構成する uni-gram」を、文対応の推定には「レビュー文・応答文を構成する uni-gram」「レビュー文・応答文のコサイン類似度」「予め文種類モデルで推定したレビュー文・応答文の種類（従来手法の場合のみ<sup>3</sup>）」を用いた。

文対応推定性能の評価は、情報検索などの分野で良く用いる再現率 (Recall)・適合率 (Precision)、及びそれらの調和平均である F 値から行う。

### 5.2 実験結果

実験結果を表 1 に示す。表の  $P_{avg}$ 、 $R_{avg}$ 、 $F_{avg}$  はそれぞれ適合率・再現率・F 値を表しており、各数値は 5 分割交差検定での平均値である。

従来手法と性能を比較すると、提案手法の方が文対応推定性能が高くなった。特に再現率が改善したことが F 値の改善に寄与している。

### 5.3 考察

実験結果より、従来手法に比べて提案手法の方が優位であることが分かった。特に提案手法と 2D CRF の差異は、文種類をどのように扱うかの違いのみである。すなわち、2D CRF では、与える文種類は文種類モデルによる推定で固定された素性であるのに対し、提案手法では文種類を同時に推定するため様々な文種類の可能性を併せて考慮する点が異なる。この扱い方の違いによって、文対応推定の性能差が生まれたと考えている。

<sup>2</sup>楽天データ公開: <http://rit.rakuten.co.jp/rdr/>。データは 2010 年当時の公開データを用いた。

<sup>3</sup>提案手法の場合は、レビュー文種類・応答文種類は文対応と同時に推定されるため、これらを陽に素性として追加する必要はない。提案手法においては文種類と文対応を同時に考慮する因子が存在するため、文種類を考慮した文対応推定が実現できる。

表 1: 文対応推定性能

	$P_{avg}$	$R_{avg}$	$F_{avg}$
<b>proposal</b>	<b>51.49</b>	<b>50.71</b>	<b>51.10</b>
L-CRF <sub>review</sub>	49.73	43.68	46.51
L-CRF <sub>reply</sub>	47.85	44.77	46.26
2D-CRF	49.70	47.77	48.72

## 6 応答文の自動生成

### 6.1 概要

文対応関係を推定した結果を用いれば、これを対訳文とみなして統計的機械翻訳の枠組みを用いることで、応答文自動生成への応用が可能である。同言語間での応答を統計的機械翻訳によって生成する従来研究として、英語では [Ritter et al., 2011]、日本語では [長谷川他, 2012] が存在する。

これら従来研究は文対応（応答先の発話）情報がある Twitter をデータに用いるが、本研究では対応する文は推定を要す条件の元で応答文の生成を試みる。

### 6.2 実験条件

まず、対訳文の作成方法と作成結果について説明する。前節までの手法で推定した文対応関係は多対多関係を許すため、そのままでは対訳文として用いることはできない。そこで、同一の文から複数の文への対応が存在する場合は結合して一つの文とする。但し、120 文字を超える文や結合文数が 3 文以上となる場合はデータから省いた。また、文中の固有表現を CaboCha [工藤・松本, 2001] を用いて特別な単語に置換した。

以上の条件に基づき、文対応推定用いた 5 分割交差検定のある特定の 1 つのケースにおける学習データ 800 文書対・テストデータ 200 文書対を除いた 275,562 文書対を対象に提案手法を用いて文対応推定を行い、対訳文へ変換した。結果、27,916 文対が対訳コーパスとして得られた。

言語モデル作成ツールには SRILM ToolKit 1.7.0 [Stolcke, 2002] を使用し、オーダーは 5-gram とした。翻訳モデル作成及びデコーダには Moses 0.91 [Koehn et al., 2007] を使用した。

### 6.3 実験結果・考察

テストデータにおけるレビュー文を入力した場合の応答文生成例を図 3 に示す。応答として適切な生成文は背景色で強調した。なお、BLEU 値は 12.78 であった<sup>4</sup>。

生成結果の中でも、入力例 1-4 は妥当な出力であった。例 1-2 は他のレビュー文書においても非常に類似した表現が存在するレビュー文であるが、特に例 3-4

<sup>4</sup>文生成の出力として妥当な文は異言語翻訳と比較するとより多様であるため、BLEU 値が応答生成器の指標に相応しいかは不明である。そのため、BLEU 値は参考値として掲載するにとどめる。

番号	レビュー文	実際の応答文	生成された応答文
1	いつも利用してます。	いつも****にご宿泊頂きまして誠にありがとうございます。	いつもご利用有難うございます。
2	****からすぐでとても便利だった。	当館は****西口から徒歩約30秒と大変近くにあります。	****からは、とても便利でございます。
3	楽天のポイントがまだ付いてません。	楽天トラベルに確認致しました所、楽天トラベルサイト内にてご案内させて頂いておりますように、楽天スーパーポイントの付与日は以下のとおりです。	楽天ポイントの件、申し訳ございませんでした。
4	残念なのはお風呂の蛇口。	<対応する応答文なし>	お風呂の蛇口の件大変申し訳ございませんでした。
5	無線LANもあり、食事もおいしかった。	****円プラスで朝食付きのプランは大変お得でございます。	無線LANにつきましては、大変申し訳ございませんでした。
6	天然温泉がついていていつもゆったりできます。	「いつもゆったりできます」とのお言葉を頂戴致しまして大変安心しております。	天然温泉がついていたようで大変嬉しく思っております。
7	あと枕が低く型くずれして、あまり眠れなかったもので、できれば枕をもう少しいい物に変えていただけたらうれしいです。	ご宿泊中に不快な思いをさせてしまい、誠に申し訳ございませんでした。枕につきましては慎重に検討させて頂き、皆様に喜んでいただけるように努めて参りたいと思います。	枕でございます、こちらは今後の検討材料にさせて頂きたいと思っております。で型くずれ

図 3: 自動生成された応答文の例

はより対象を限定した文であり、他の文書にはあまり類似する表現は登場しない。これは、統計的機械翻訳の汎化能力によるところが大きいと考える。

しかし、今回取り上げなかった例も含めて全体的な出力を見ると、多くは例 5-7 のような文脈に不適切な文や、未知語が含まれるために一部が不自然な文であった。これらの主な原因には、対訳文が不十分であると言う問題がある。一般に統計的機械翻訳では数十万〜数百万文と大規模な対訳文を確保してはじめて性能を発揮するため、本研究で用いた数万文では不十分であった可能性がある。しかし、データ量確保のために文対応推定の性能を犠牲にしてもノイズとなってしまうため、データ量の確保及び文対応推定性能の向上が必要である。

## 7 おわりに

本研究では、文対応推定のために文種類・文対応推定モデルを統合したモデルを提案し、従来手法よりも高い性能で文対応推定が可能であることを示した。また、文対応推定の結果を応答文自動生成へと利用し、応用の一つの可能性を示した。

今後の課題として、文対応推定に投入する素性をより充実することが挙げられる。また、文生成においても、推定した文種類を考慮した生成や、文書から応答文書全体を生成したいと考えている。

## 謝辞

本研究には楽天データ公開において公開された楽天トラベル「お客さまの声・クチコミ」データを使用させて頂きました。データを公開して頂きました楽天株式会社に感謝致します。

## 参考文献

- Andreas Stolcke, 2002. “SRILM - An extensible language modeling toolkit”, in *Proceedings of the 7th International Conference on Spoken Language Processing*, pp. 901–904.
- Andrew Kachites McCallum, 2002. “MALLET: A Machine Learning for Language Toolkit”.

- Alan Ritter, Colin Cherry, and William B. Dolan, 2011. “Data-driven response generation in social media”, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 583–593.
- Dragomir R. Radev, 2000. “A common theory of information fusion from multiple text sources step one”, in *Proceedings of the 1st ACL SIGdial workshop on Discourse and dialogue*, Vol. 10, pp. 74–83.
- E. A. Schegloff and H Sacks, 1973. “Opening up closings”, *Semiotica*, Vol. 8, No. 4, pp. 289–327.
- John Lafferty, Andrew McCallum, and Fernando Pereira, 2001. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, in *Proceedings of the 18th International Conference on Machine Learning*, pp. 282–289.
- Jun Zhu, et al., 2005. “2D Conditional Random Fields for Web information extraction”, in *Proceedings of the 22nd international conference on Machine learning*, pp. 1044–1051.
- Martin Wainwright, Tommi Jaakkola, and Alan Willsky, 2001. “Tree-based reparameterization for approximate inference on loopy graphs”, in *Advances in Neural Information Processing System (2001)*, pp. 1001–1008.
- Philipp Koehn, et al., 2007. “Moses: open source toolkit for statistical machine translation”, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*, pp. 177–180.
- William C Mann and Sandra A Thompson, 1986. “Rhetorical Structure Theory: Description and Construction of Text Structures”, *Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*, pp. 85–96.
- Zhonghua Qu and Yang Liu, 2012. “Sentence Dependency Tagging in Online Question Answering Forums”, in *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics*, pp. 554–562.
- 大沢裕子・郷垂里沙・安田励子, 2010 年, 「Web サイトにおけるクチコミの苦情と返答ー「宿泊予約サイト」を対象にー」, 言語処理学会第 16 回年次大会発表論文集, pp.322–325.
- 工藤拓・松本裕治, 2001 年, 「チャンキングの段階適用による係り受け解析」, 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2001, No. 20, pp.97–104.
- 角田孝昭, 2013 年, 「対をなす二文書間における文対応関係の自動推定」, 筑波大学大学院 システム情報工学研究科 修士論文.
- 難波英嗣・国政美伸・福島志穂・相沢輝昭・奥村学, 2005 年, 「文書横断文間関係を考慮した動向情報の抽出と可視化」, 情報処理学会研究報告, 自然言語処理研究会, Vol. NL-169, pp.67–74.
- 長谷川貴之・鍛冶伸裕・吉永直樹・豊田正史, 2012 年, 「聞き手の感情を喚起する発話の分類と生成」, 電子情報通信学会思考と言語研究会 (2012-10).
- 宮部泰成・高村大也・奥村学, 2006 年, 「文書横断文間関係の特定」, 言語処理学会第 12 回年次大会, pp.496–499.