# キャッチコピーにおける対句構造の解析

丹羽彩奈[1]　脇本宏平[2]　西口佳佑[2]　毛利真崇[2]　岡崎直観[1]
[1]東京工業大学　[2]株式会社サイバーエージェント
ayana.niwa at nlp.c.titech.ac.jp　okazaki at c.titech.ac.jp

## 1 はじめに

キャッチコピーでは，商品や作品を端的に表現するため，様々な修辞技法が用いられる．丹羽ら [10] は，キャッチコピーのコーパスを分析し，比喩，対句，誇張，呼びかけ，断定，反復，擬態・擬音などの修辞技法がよく用いられることを報告した．また，キャッチコピーの約 9 割は単文で構成され，その文の平均長は 16.2 文字であった．キャッチコピーの自動生成は，商品や作品のマーケティングに役立つだけでなく，文生成研究の対象としても興味深い方向性である．

本研究では，キャッチコピーで用いられる修辞技法の中でも，比較的多く（全体の約 10% ほど）用いられ，かつ挑戦的な研究対象である**対句**に着目する．対句とは，類似した文構造と対照的な意味を持つ文や句を並列させる修辞技法である．本稿では，対句を用いたキャッチコピー生成に向けた第一歩として，キャッチコピー中の対句構造を認識する手法を述べる．対句構造の認識により，対句構造の生成テンプレートを導出したり，対照的な意味を持つ単語ペアを抽出できると期待される．

管見の限り，対句の認識や生成に取り組んだ既存研究はない．しかしながら対句は並列構造の一種と捉えることができるため，並列構造解析 [3, 4, 6] と対句構造解析には共通点がある．並列構造解析では，並列を表す手がかり表現（並列キー）[*1] を検出し，その周辺の表現から並列句スパンを同定する手法が主流である．ところが，対句構造では，対句を示す手がかり表現が常にあるとは限らない．そのため並べられている句の意味的な対照性を捉える必要があり，並列構造解析よりも難易度が高い．

本研究では，キャッチコピー中の対句構造をアノテートしたデータセットを構築する．次に，キャッチコピーから対句構造を抽出するタスクを，関係抽出の一種として捉え，対句構造のスパンを認識する手法を構築する．関係抽出モデルとして，エンティティ抽出と関係抽出を同時に行う SpERT [2] をベースに，対句構造の言語的な特徴を踏まえた特徴量を導入する．SpERT は BERT [1] に基づく関係抽出モデルであり，対句構造のような長距離の依存関係を捉えやすいと期待される．

対句構造を付与したデータセットを用いた実験では，平均 F 値 71.8 の精度で対句構造解析が行えることがわかった．また，SpERT のベースラインモデルと比較すると，本研究で対句構造に特化した特徴量を導入したことで，解析性能が向上することを報告する．

[*1] 英語では「and」，日本語では「と」などの等位接続詞が並列構造を示す手がかり表現として用いられる．

## 2 関連研究

並列構造解析の研究では，並列句の類似性に着目することが多い．黒橋ら [4] は，並列句の類似度スコアを考慮した動的計画法により，並列構造の検出と範囲の推定を行なった．寺西ら [6] は，並列構造の句の持つ文法的・意味的類似性に加えて可換性という性質に着目し，双方向 LSTM による解析モデルを提案した．Ficler ら [3] は，与えられた文が並列構造を持つか否かの 2 値分類を行った後，並列句の境界を同定する階層的な手法を提案した．これらの手法は，いずれも並列キー（手がかり表現）に依存している．しかし，対句構造では明確な手がかり表現がないこと，句の構造も名詞句から文まで様々であること，句の意味的な内容から対句構造を捉える必要があること等から，並列構造解析よりも対句構造解析の方が難しいと考えられる．

対句構造解析は，与えられたテキストの中で対となる句のスパン（範囲）を認識し，二つの句が対句の関係にあるかどうかを判定することで実現できる．よって，関係抽出のタスクの一種として見ることもできる．関係抽出の研究では，長距離の依存関係や文脈を捉えやすくなった BERT [1] に基づくモデルが性能の向上に寄与することが近年多く報告されている．Shi ら [5] は，BERT エンコーダに双方向 LSTM を積み重ねる単純な手法で

も，関係抽出で高い性能を示すことを報告した．Eberts ら [2] は，固有表現抽出と関係抽出を同時に学習する SpERT を提案し，CoNLL 2004 などの複数のデータセットで最高性能を達成した．

関係抽出では，認識する固有表現（句）には人名や地名など意味カテゴリが定義される．一方，対句構造解析では対となる句に明確な意味カテゴリはなく，二つの句がペアリングされてから対句としての解釈が可能になる．そこで，本研究では SpERT のスパン抽出（固有表現抽出）を対となる句の候補の列挙に用い，対句の関係を同定する関係抽出器と同時に学習する．これにより，BERT のエンコーダで句のスパン内外の構造にも着目しながら対句構造の抽出を行うモデルを提案する．

## 3 キャッチコピーにおける対句

### 3.1 対句の言語的特徴

対句は，キャッチコピーにおいて一般的に使用される修辞技法の一つである．以下の具体例のように，類似した構造と対照的な意味を持つ文や句から構成される．

人生は、[近く で見ると悲劇] だが、
[遠く から見れば喜劇] である。

この対句は，３つの言語的特徴を持つ．

**1. 句の構造の類似性**　対句構造を構成する二つの句は，その句の内部構造に類似性がある．上述の例では，「近くで見ると悲劇」と「遠くから見れば喜劇」に構造的な類似性がある．

**2. 句の可換性**　対句構造を構成する二つの句は，互いに入れ替えても文の流暢性が保たれる．上述の例では，「人生は、[遠くから見れば喜劇] だが、[近くで見ると悲劇] である。」のように，対句を入れ替えることができる．

**3. 句の意味的対照性**　対句構造を構成する二つの句に含まれる単語は，意味的な対照性を持つことがある．上述の例では，「近く – 遠く」「悲劇 – 喜劇」という反義語が用いられている．これは，必ずしも意味が二対である必要はなく，「家族 – 世間」のように対照的な意味で用いられる対照語の関係も含む．

### 3.2 対句データセットの構築

丹羽ら [10] は，著名なコピーライターによる作品集 [8, 12, 13] と書籍 SKAT [9] から 125,886 件のキャッチコピーを収録するコーパスを構築した．本研究では，このコーパスに対して対句構造のアノテーションを行った．

**表1:** 対句のアノテーション例（[ ] 内は対となる句）．

| |
|---|
| [運がないなら勇気] を、[勇気がないなら運] で。 |
| [彼の紹介した漫画] が、[私の好きな漫画] になっていた。 |
| [アタマが水を求める] 時、[カラダはイオンを求めてる] |

アノテーションにはクラウドソーシングを活用した．まず，１つのキャッチコピーに対して５人の作業者を割り当て，作業者はそのキャッチコピーに対句が含まれるかを判定する．４人以上の作業者が対句構造を含むと指摘した約１万件のキャッチコピーに対して，アノテーションの質が高い２人の作業者を抽出し，対句構造のスパンにアノテーションを行い，10,108 件の対句データセットを構築した．文字化けなどは随時修正し，また判断がつかないものには「迷い」のフラグを立て，後で確認を行った．表1にアノテーション例を示す．なお，これらの例では「A を B で」「A が B になっていた」「A 時 B」という構造があるが，並列構造解析における並列キーのような明確な手がかり表現がないことに注意されたい．

3.1節で説明した通り，対句には単語と句という二種類の対応関係があるが，今回は単語単位ではなく句単位でアノテーションした．アノテーションすべきスパンを明確にするため，対句として解釈できる構造の中でできるだけ広いスパンを採用することにした．また，『「調べればわかる」より、「調べないでもわかる」方がいい。』のように，対応するスパンの両方にかぎ括弧や句読点が含まれる場合は，これらも含めることとした．「作り笑いが嫌いで日本を飛び出すのに、外国で愛想笑いをしますか。」のように対照語の組が複数含まれるが，句の構造が異なる場合に「迷い」が生じることが多かった．

## 4 対句構造解析

### 4.1 SpERT

本研究では，対句構造解析のアーキテクチャとして SpERT [2] を採用する．SpERT は入力文を BERT でエンコードし，各トークンの隠れ状態ベクトルからスパン分類（固有表現抽出），スパンフィルタリング，関係分類を end-to-end で実行するモデルである．スパン分類では，まず入力トークンの可能な連結からスパンの候補を列挙する．次に各スパン候補のトークンの隠れ状態ベクトルの最大値プーリングやスパンの長さの埋め込み表

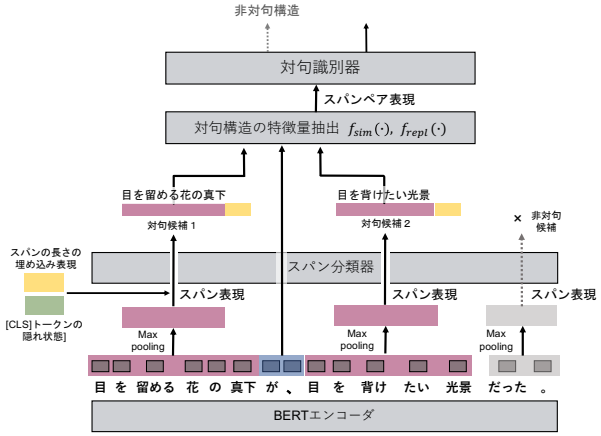


**図1:** モデル概略図（推論時）

現，文脈ベクトル（[CLS] トークンの隠れ状態）を特徴量としてスパンの認識を行う．関係抽出では，2 つのスパンの特徴量，2 つのスパンの間の表現の特徴量を連結し，2 つのスパン間の関係の識別を行う．

### 4.2 提案手法

本研究では SpERT に対句認識のための特徴量を組み込んだ手法を提案する．SpERT のスパン分類器で対句を構成しうるスパンの候補を列挙し，関係分類器で対句となるスパンのペアを選択するか，対句「無し」と判定する．提案モデルの概略図を図1に示す．対句のスパンの候補を認識するスパン分類器は，SpERT のモデルから変更を加えていない．ただし，スパン分類器は少なくとも 2 個のスパンを対句を構成するスパンの候補として出力することとした[*2]．

関係分類器（以降，対句識別器と呼ぶ）は，与えられた 2 つのスパンの組が対句か否かを識別する．対句識別器への入力として，それぞれのスパンの隠れ状態ベクトルと長さの埋め込み表現に加えて，対句の言語的性質に基づく特徴量として，句の類似性 $f_{\mathrm{sim}}(\cdot)$ と句の可換性 $f_{\mathrm{repl}}(\cdot)$ を導入する．

入力されたトークン列の中で，$[i,j]$ と $[k,l]$ の範囲の 2 つのスパンが対句識別器に与えられたとする．このとき，この 2 スパンの類似性特徴量 $f_{\mathrm{sim}}(i,j,k,l)$ は，

$$f_{\mathrm{sim}}(v_{i,j}, v_{k,l}) = [|v_{i,j} - v_{k,l}|; v_{i,j} \circ v_{k,l}], \quad (1)$$

$$v_{i,j} = \mathrm{maxpool}(h_i, h_{i+1}, ..., h_j), \quad (2)$$

$$v_{k,l} = \mathrm{maxpool}(h_k, h_{k+1}, ..., h_l). \quad (3)$$

ただし，$h_t$ は位置 $t$ のトークンに対する BERT エンコーダの隠れ状態ベクトル，$|u|$ はベクトル $u$ の要素毎の絶対値，$u \circ v$ はベクトル $u$ と $v$ の要素積，maxpool(.) は，引数のベクトル群に対する最大値プーリングを表す．

可換性特徴量 $f_{\mathrm{repl}}(i,j,k,l)$ は，$[i,j]$ と $[k,l]$ の範囲の 2 つのスパンに対して，各スパンの開始・終了トークンの入れ替えやすさ，前後のトークンとの接続の良さを定量化する．なお，前半のスパンが先頭トークンから始まる場合，または後半のスパンが最終トークンで終わる場合は，その接続の良さは考慮しない．

$$\begin{aligned} & f_{\mathrm{repl}}(i,j,k,l) \\ =& [|h_i - h_k|; \mathrm{avg}(h_{i-1} \circ h_k, h_i \circ h_{k-1}); \\ & |h_j - h_l|; \mathrm{avg}(h_j \circ h_{l+1}, h_l \circ h_{j+1})] \end{aligned} \quad (4)$$

ただし，avg(·) は，引数のベクトルの平均値を表す．

対句識別器には，SpERT でも使われているスパン長の埋め込み表現と文脈ベクトルに加え，$f_{\mathrm{sim}}(\cdot)$ と $f_{\mathrm{repl}}(\cdot)$ の特徴ベクトルを連結したものを与える．全体のモデルを学習するときの損失関数としては，SpERT と同様にスパン分類器と対句識別器の損失の和を用いる．

## 5 実験

本研究で構築した対句データセットのうち，1,265 件を学習データ，270 件ずつを開発データと評価データに用いた．1 スパンあたりの平均トークン数は 4.76，最大トークン数は 16 である．スパン分類器と対句識別器を学習する際の負例として，ランダムにスパンとスパンペアをサンプリングしたものを用いた．負例サンプル数は，スパン，スパンペアともに 100 とした．最大スパン長は 10 トークンとした．また，対句ペアはトークン数に大きな差がないことを踏まえ，スパンペアのトークン数の差が 3 以下のものだけを採用した．これらのハイパーパラメータは，対句データセットの統計に基づいて決定した．BERT の事前学習モデルには日本語 Wikipedia で学習された公開モデル [11]，日本語のトークナイザとして Juman++ (v2.0.0-rc2) [7] を用いた．

対句識別器の性能は，対句と判定されたスパンペアに対して．固有表現抽出と同様にスパンの一致度合いをトークン単位で評価した．スパン分類器の性能は，分類器が出力した全スパンに対して，正解スパンと完全一致したスパンかどうかで評価した．

表 2に，スパン分類と対句識別の性能を示した．対句識別器に入力するスパンペアの特徴量として，(1) Baseline (SpERT の手法から変更せず，スパン表現をそのまま連

[*2] スパン分類器が 1 個のスパンしか認識しなかった場合は，スパンのスコアが高い順に 2 個のスパンを強制的に候補として加えた．

**表2:** 対句識別器の予測精度

| | 特徴量 | | | スパン分類 | | | 対句識別 | | |
|---|---|---|---|---|---|---|---|---|---|
| | スパンベクトル | $f_{sim}$ | $f_{repl}$ | P | R | F | P | R | F |
| 系列ラベリング | | | | 0.608 | 0.607 | 0.608 | | | |
| (1)Baseline | ✓ | | | 0.722 | 0.692 | **0.707** | 0.774 | 0.601 | 0.676 |
| (2)SIM | | ✓ | | 0.714 | 0.687 | 0.701 | 0.738 | 0.681 | 0.708 |
| (3)REPL | | | ✓ | 0.717 | 0.694 | 0.706 | 0.737 | 0.703 | **0.718** |
| (4)BOTH | | ✓ | ✓ | 0.702 | 0.688 | 0.696 | 0.736 | 0.683 | 0.709 |

結した特徴量を用いた場合），(2) SIM（類似性），(3) REPL（可換性），(4) BOTH（可換性＋類似性）の４種類を比較した．なお，これらの性能の数値は 5 回の実験の平均値である．参考として，BERT の出力ベクトルを直接ラベル分類器に接続し，系列ラベリングタスクとして学習した場合の結果も示した．

表 2に示したとおり，対句構造の言語的性質である可換性や類似性を考慮することで，スパン表現を直接入力する Baseline と比較して F 値が平均で約 3.2〜4.2 ポイントほど向上することを確認した．また，可換性のみを考慮した REPL が最も性能が高かった．これは，対句は文構造の多様性により，句の類似性よりも接続点のみに着目した方が，効果的に対句構造を認識できたことを示唆している．Baseline は，スパンベクトルを一貫してモデル全体で学習するため，スパン分類器としては良い性能を示す．しかし，対句構造の特徴を埋め込んでいないため，正解スパンペアを対句ペアと認識する際に性能が低下する．以上より，対句構造解析を関係抽出タスクとして解く場合，対句構造の性質，特に可換性を埋め込むことが有効であることが分かった．

表3に，今回の提案手法における不正解事例の傾向を示した．これらの事例から分かる通り，正解スパンとの関係性が強いトークンは一緒にスパンとして出力されることがある．このような場合は，類似性・可換性だけでは対処できないため，例 2 であれば「神話 – 事実」「は– が」「いらない– いる」といったように，トークン単位での対応関係をモデルに組み込み，対応先がない「安全に、」をスパンから取り除く機構が必要である．

## 6 おわりに

本稿では，対句構造の解析に向け，対句構造データセットの作成を行った．さらに，対句構造解析を関係抽出タスクとして見なし，対句構造の特徴量を組み込んだ解析手法の提案を行った．提案手法は，対句の言語的性質の埋め込みとスパン分類器のスパン候補の絞り込みにより，効率的に性能を向上させた．特に，可換性が対句解析において重要な手がかりであることが分かった．今後は，句内部のアラインメントの考慮，反義語などの事前知識の導入，並列構造解析における有効性の検証を続けていきたい．また対句構造を含むキャッチコピーの自動生成にも取り組む予定である．

**表3:** 提案手法による解析の失敗傾向

| | |
|---|---|
| 正解 1 | [「がんばれ」も嬉しい] ですが、[「おいしい」はもっと嬉しい] |
| 予測 1 | [「がんばれ」も嬉しいです] が、[「おいしい」はもっと嬉しい] |
| 正解 2 | 安全に、[神話はいらない]。[事実がいる]。 |
| 予測 2 | [安全に、神話はいらない]。[事実がいる]。 |

## 参考文献


[1] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proc. of NAACL*. 2019, pp. 4171–4186.
[2] Markus Eberts and Adrian Ulges. “Span-based Joint Entity and Relation Extraction with Transformer Pre-training”. In: *ArXiv* abs/1909.07755 (2019).
[3] Jessica Ficler and Yoav Goldberg. “A Neural Network for Coordination Boundary Prediction”. In: *Proc. of the EMNLP*. 2016, pp. 23–32.
[4] Sadao Kurohashi and Makoto Nagao. “A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures”. In: *Computational Linguistics* 20.4 (1994), pp. 507–534.
[5] Peng Shi and Jimmy Lin. “Simple BERT Models for Relation Extraction and Semantic Role Labeling”. In: *CoRR* abs/1904.05255 (2019).
[6] Hiroki Teranishi, Hiroyuki Shindo, and Yuji Matsumoto. “Coordination boundary identification with similarity and replaceability”. In: *Proc. of IJCNLP*. 2017, pp. 264–272.
[7] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. “Juman++: A Morphological Analysis Toolkit for Scriptio Continua”. In: *Proc. of the EMNLP*. Brussels, Belgium, Nov. 2018, pp. 54–59.
[8] 仲畑貴志. ホントのことを言うと、よく、しかられる。勝つコピーのぜんぶ. 株式会社宣伝会議, 2018.
[9] 宣伝会議賞実行委員会. *SKAT.2-SKAT.17*. 株式会社宣伝会議, 2003-2018.
[10] 丹羽 彩奈 et al. “キャッチコピーの自動生成に向けた分析”. In: 言語処理学会第 *25* 回年次大会 *(NLP2019)*. Mar. 2019, P3–12 (4 pages).
[11] 柴田 知秀, 河原 大輔, and 黒橋 禎夫. “BERT による日本語構文解析の精度向上”. In: 言語処理学会第 *25* 回年次大会 *(NLP2019)*. Mar. 2019, P205–208 (4 pages).
[12] 谷山雅計. 広告コピーってこう書くんだ!読本. 株式会社宣伝会議, 2007.
[13] 青田 光章, 秋山 晶, 東 秀紀, ほか. 最新約コピーバイブル. 株式会社宣伝会議, 2007.