

自然言語生成における テンプレートの導出と活用

丹羽彩奈（東京工業大学）

人工知能学会 第118回人工知能基本問題研究会 11/26

自己紹介

● 丹羽彩奈

- 東京工業大学 情報理工学院 博士課程2年
- 岡崎研究室所属
- **自然言語処理**に関する研究をしています



 @ayaniwa

● 自然言語処理の和を広げるぞ活動

- YANS若手の会 運営委員
 - 萌芽的な研究を促進するシンポジウム等を運営
- NLP Dの会 幹事
 - 年に3~4回、博士課程に在籍している全国の学生が集い研究に関する議論を行う会を開催
- NLPコロキウム オーガナイザ
 - 月に1~2回、最新のNLP研究に関するトークイベントを開催

 @yans_official

 @nlp_colloquium

興味があること

目的

書き手・話し手の「意図」を効果的に伝えることばの生成

- ことばはものを伝える道具である
 - どういう単語選択でどういう言い回しをしたら効果的に伝わるのか？

そのための方針

意図と発話内容がペアになっている言語資源は少ない

→言語資源の量に囚われない言語生成システムを構築する

実現させるための手段

テンプレートベースの生成手法を発展させること
少資源のドメインにも適用可能（後述）

コンテンツ

1. 自然言語生成入門

- 自然言語生成とは
- 近年の潮流
- 実用化されている自然言語生成技術とその裏側

2. 自然言語生成とテンプレート

- テンプレートとは
- なぜテンプレート？
- テンプレートベースの文生成手法に求められること

3. テンプレートの導出と活用

- どう獲得し、使うのか？
- 現状と今後の展望

コンテンツ

1. 自然言語生成入門

- 自然言語生成とは
- 近年の潮流
- 実用化されている自然言語生成技術とその裏側

2. 自然言語生成とテンプレート

- テンプレートとは
- なぜテンプレート？
- テンプレートベースの文生成手法に求められること

3. テンプレートの導出と活用

- どう獲得し、使うのか？
- 現状と今後の展望

自然言語処理とは？

言葉（=自然言語）で伝達される情報を理解・検索・抽出・翻訳・整理・分析し、地球規模のコミュニケーションを支援するソフトウェア技術



機械翻訳

ある言語の文章を
別の言語に翻訳する



質問応答

自然言語で与えられた
質問に答える



対話エージェント

コンピュータと
人間の間で会話をする



自動要約

情報を集約して
文章を生成する



情報検索

大量の文書の中から
必要なものを探す



情報抽出

文章から事実や
知識を抽出する



意見・感情分析

文章から人間の主觀
的評価を抽出する

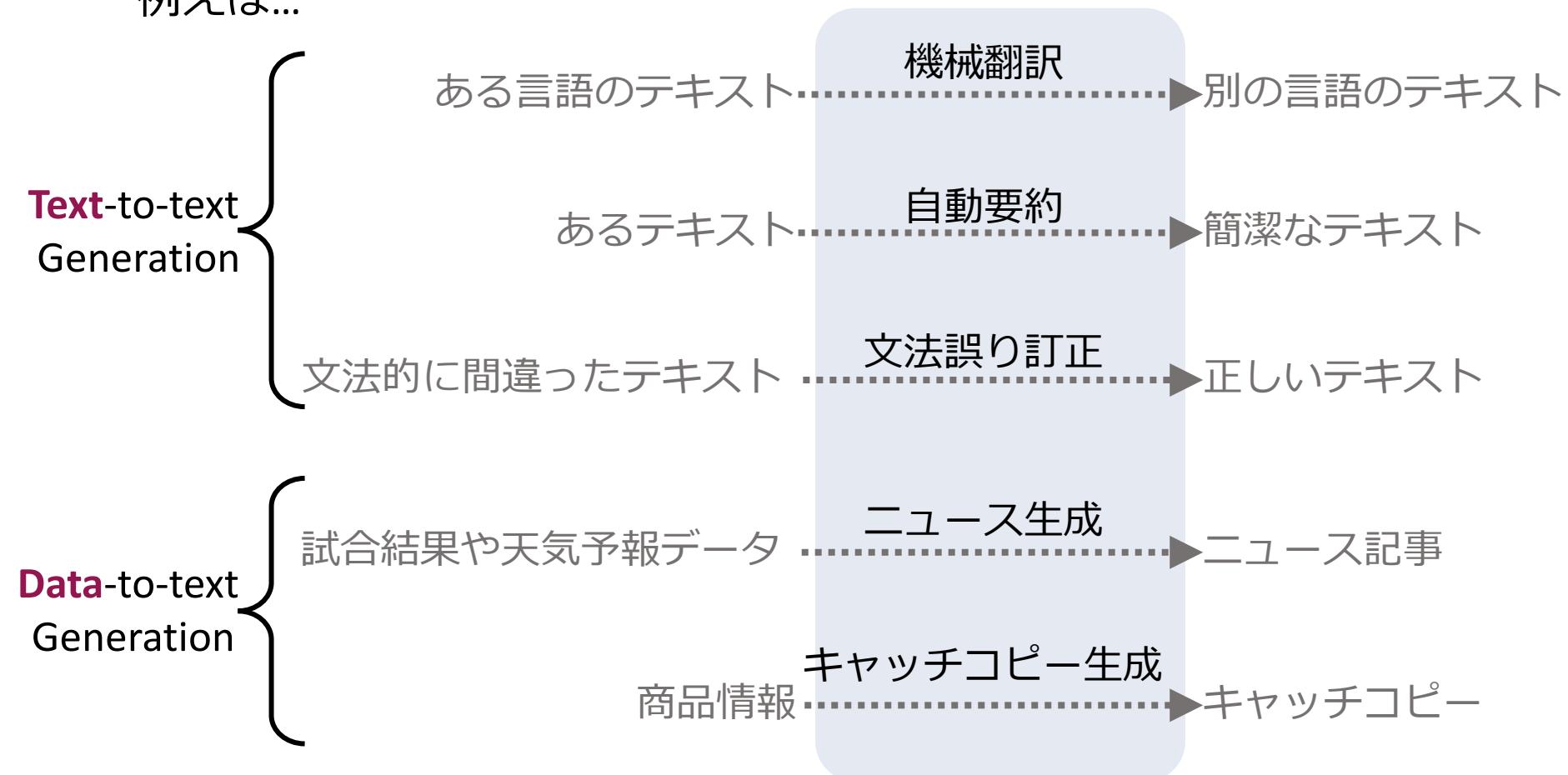


教育支援

人が良い文章を
書くように支援する

自然言語処理の一分野 自然言語生成

一般的には出力が自然言語文であるタスク全般を指す [McDonald, 1993]
例えば...



自然言語生成技術の歴史

ルールベース

機械学習

DNN

テンプレートベース

I am a .

student, teacher, doctor

大規模コーパスの誕生

Penn Treebanks

[Marcus et al., 1993]

京都大学テキストコーパス

[Kurohashi and Nagao, 1998]

Shared Taskの開始

参照表現生成

[Gatt and Belz, 2008]

画像からの案内文生成

[Koller et al., 2009]

...

エンコーダ・デコーダモデル

アテンションメカニズム

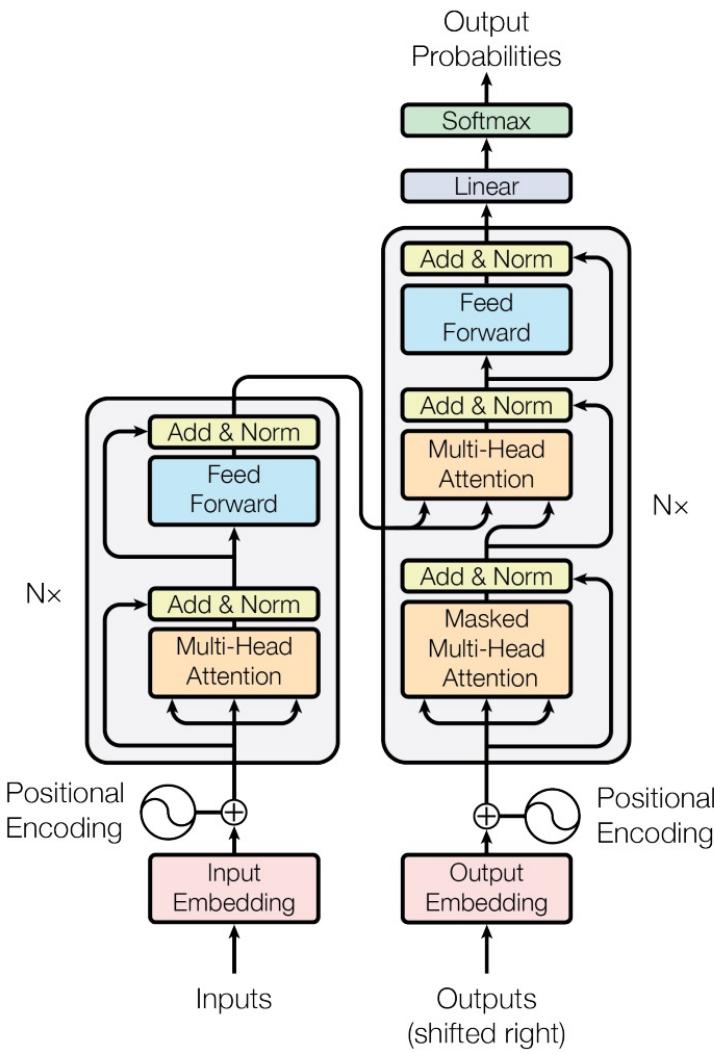
Transformerアーキテクチャ

事前学習

計算パワーの向上も一因に

→自然言語生成の性能が
飛躍的に向上

Transformerアーキテクチャ



[Vaswani et al., 2017]

それまでの標準アーキテクチャであった
RNNやCNNを置き換えた手法

詳細についてはネット上にたくさん記事があるので割愛

強み① RNNより高速な計算

強み② CNN同様並列化しやすい

強み③ 長距離の依存関係を学習しやすい

性能・汎用性ともに優れたモデルアーキテクチャで注目を浴び、その後生成系のタスクのスタンダードになってきている

- NLPのみならず、CVやARにまで適用され高性能を達成。理論分析も進む

事前学習

大量のラベルなしデータから**汎用的な言語知識**を獲得するための学習

Step1 事前学習



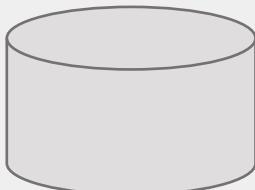
ウィキペディア
フリー百科事典

大量のラベル無し
データで学習

事前学習 モデル

Step2 転移学習

特定タスクに適用
文書分類
系列ラベリング
質問応答...



少量の
ラベルあり
データで学習

アプローチ① 素性抽出器

事前学習モデルの
パラメータを固定
自然言語を**特徴量**に変換

アプローチ② ファイン チューニング

事前学習モデルの
パラメータを初期値と
して**再学習**

近年、自然言語処理における有効性が多く報告されている
ほとんどのモデルでTransformerアーキテクチャが採用されている

自然言語生成と実応用

Transformerアーキテクチャや事前学習などの手法開発とともに
自然言語生成技術の実応用も進んできた

地震ニュースの自動生成

Los Angeles Times

Magnitude 6.2 quake hits Taiwan



Los Angeles Times

BY QUAKEBOT
OCT. 23, 2021 11:12 PM PT

A magnitude 6.2 earthquake was reported Saturday evening at 10:11 p.m. Pacific time 13 miles from Yilan, Taiwan, according to the U.S. Geological Survey.

According to the USGS, the epicenter was further than 100 miles from a city.

In the past 10 days, there have been 3.0 or greater centered nearby.

SUBSCRIBERS .

POLITICS

FOR SUBSCRIBERS:

Is your compa

at home? Sinc

LIFESTYLE

FOR SUBSCRIBERS:

Pick up the pe

stores you'll fi

BUSINESS

FOR SUBSCRIBERS:

April 29, 2021



Click to copy

RELATED TOPICS

Earnings

Business

決算報告の自動生成

AP通信×Automated Insights

McDonald's: Q1 Earnings Snapshot

CHICAGO (AP) — McDonald's Corp. (MCD) on Thursday reported first-quarter net income of \$1.54 billion.

On a per-share basis, the Chicago-based company said it had profit of \$2.05. Earnings, adjusted for non-recurring gains, came to \$1.92 per share.

The results topped Wall Street expectations. The average estimate of 14 analysts surveyed by Zacks Investment Research was for earnings of \$1.81 per share.

The world's biggest hamburger chain posted revenue of \$5.12 billion in the period, also exceeding Street forecasts. Thirteen analysts surveyed by Zacks expected \$5.05 billion.

McDonald's shares have increased slightly more than 8% since the beginning of the year, while the S&P's 500 index has increased 11%. The stock has risen 25% in the last 12 months.

This story was generated by Automated Insights (<http://automatedinsights.com/ap>) using data from Zacks Investment Research. Access a Zacks stock report on MCD at <https://www.zacks.com/ap/MCD>

This story was generated by Automated Insights (<http://automatedinsights.com/ap>) using data

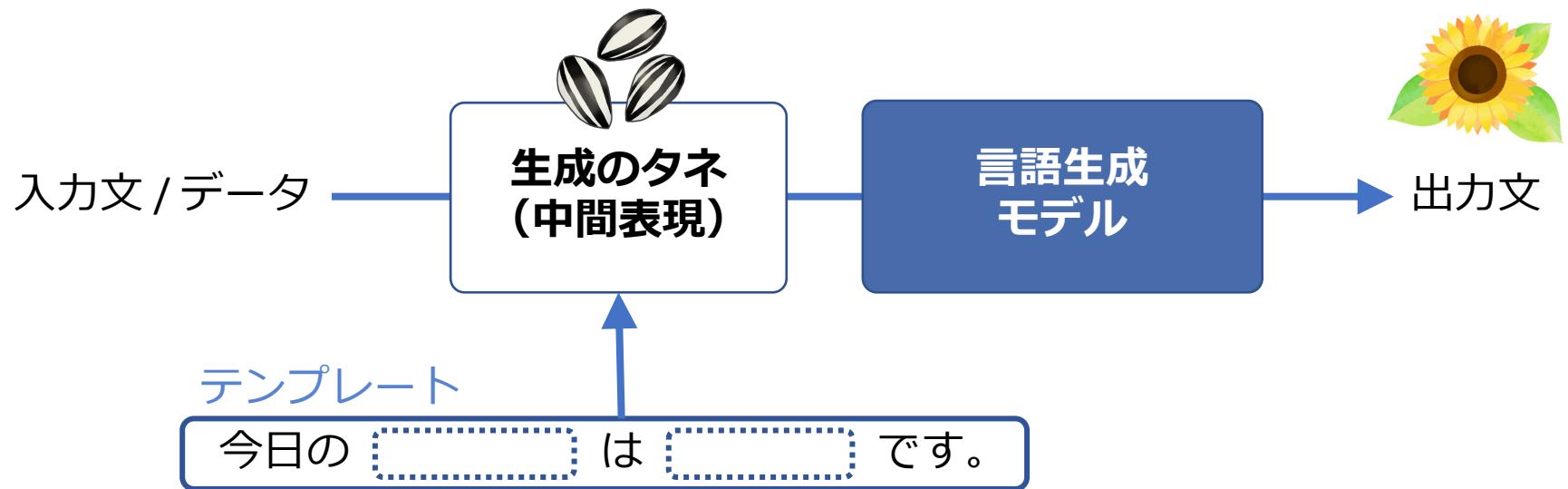
導入後出稿量が12倍になったケースも※

人間が取材などより重要なタスクに時間を割けるように

※<https://insights.ap.org/industry-trends/study-news-automation-by-ap-increases-trading-in-financial-markets>

実用化成功の背景

実用化が成功しているひとつの背景は「生成のタネ（種）としてテンプレートを採用していること」



本日はこのテンプレートを用いた自然言語生成に関する取り組みについてお話しします

コンテンツ

1. 自然言語生成入門

- 自然言語生成とは
- 近年の潮流
- 実用化されている自然言語生成技術とその裏側

2. 自然言語生成とテンプレート

- テンプレートとは
- なぜテンプレート？
- テンプレートベースの文生成手法に求められること

3. テンプレートの導出と活用

- どう獲得し、使うのか？
- 現状と今後の展望

自然言語生成とテンプレート

テンプレートとは、テキストセグメントの系列のこと

- 必ずしもスロットつきの不完全文を指すわけではない（後述）

(例) 試合レポート生成のためのテンプレート

試合の名前 が 開催日時 に 開催場所 で開催され、

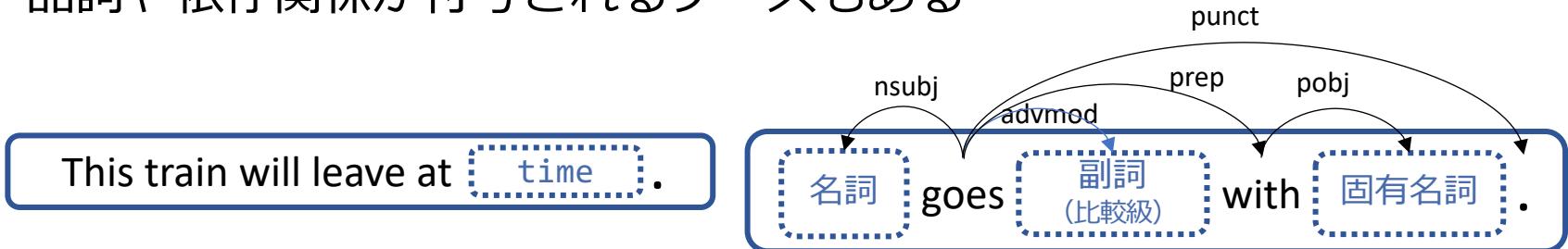
負けたチーム名 を 勝ったチームのスコア - 負けたチームのスコア

でくだし、 勝ったチーム名 が首位に立った。

テンプレートの種類

① ハード（ルールベース）テンプレート

テンプレートの使い方（例：スロットの穴埋め）が明示されている品詞や依存関係が付与されるケースもある



② ソフトテンプレート

使い方について明示的なルールがない。あくまで出力を制御する役割

記事

european stock markets
advanced strongly thursday ...

テン
プレ

european shares sharply lower
on us interest rate fears

要約

european shares rise strongly
on bargain-hunting

参考: [Cao et al., 2018]の例

参考になる事例（=近傍の事例）をソフトテンプレートとして活用
※近傍の事例ベースの手法との違いは明確ではない

テンプレートの嬉しさ

① 生成される文の質が高い

文法的・意味的に正しい文を生成しやすい

② モデルの出力を制御しやすい

近年論文でよく見るキーワード

Controllable, Constrained

1. 言及内容

- 真実性と事実性
 - 事実と矛盾していることを言わないようにしたい
- キーワード
 - 文に含める単語を指定したい

2. 文字数

- 適用先に合わせて変更する需要がある
 - 例) Yahooニュースの見出しは最大14.5文字

3. 表現・言い回し

- 文の丁寧さや能受動などの態

③ 解釈性が向上する

テンプレートに基づく文生成は、なぜその出力をしたのかがわかりやすい

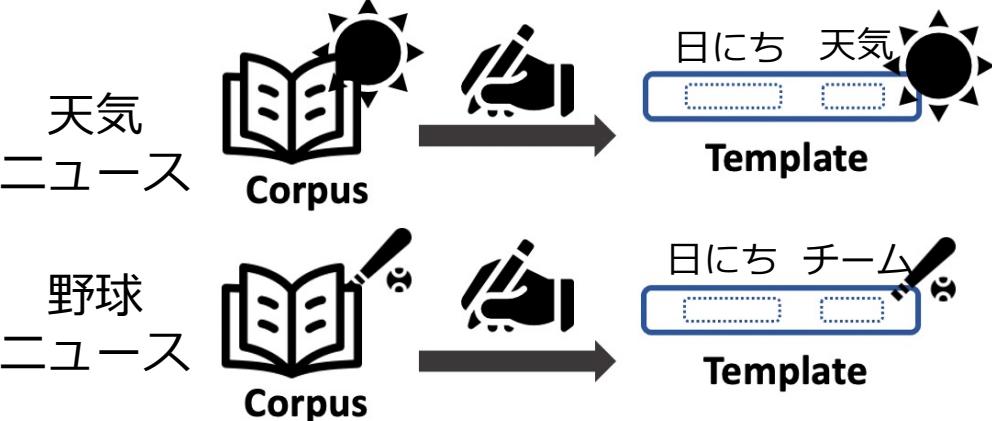
④ データ数が少ないドメインにも適用可能

ハードテンプレートの場合、文をゼロから生成する必要がない

これは機械学習を導入する前からテンプレートが使われていたことからもわかる

作成方法から見る テンプレートに求めること

人手作成（従来）



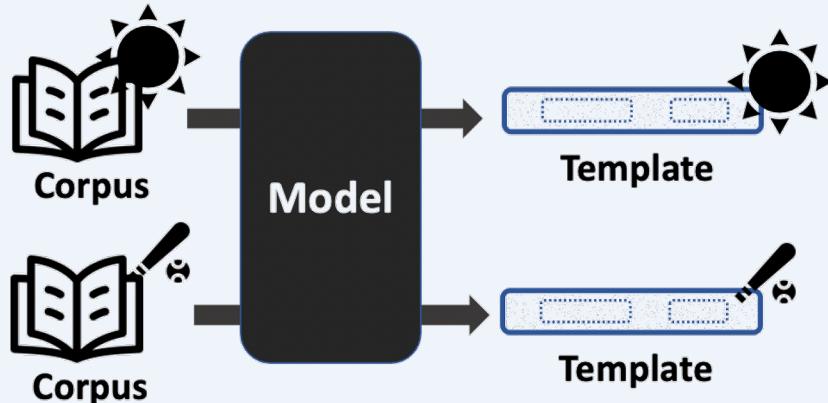
人手でテンプレートを作成し
スロット情報を付与。
スロットに自動で単語を当てはめる

強みは正確さと内容の一貫性

弱みはスケーラビリティのなさ

- 時間がかかる
- 手間がかかる
- ドメイン知識を要する

自動導出・活用



テンプレートを自動で導出し
モデルに組み込む

強みはスケーラビリティ

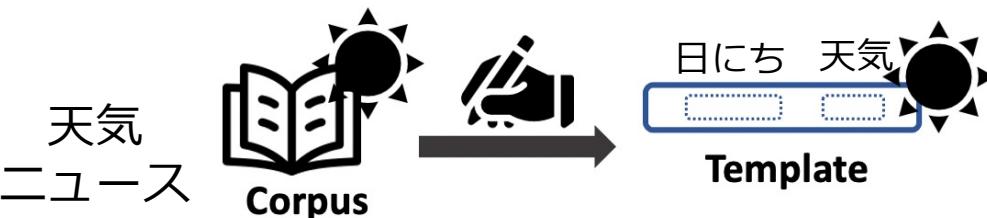
- ドメインごとに作り直し不要

弱みはテンプレート・生成される文の質

逆の関係

作成方法から見る テンプレートに求めること

人手作成（従来）

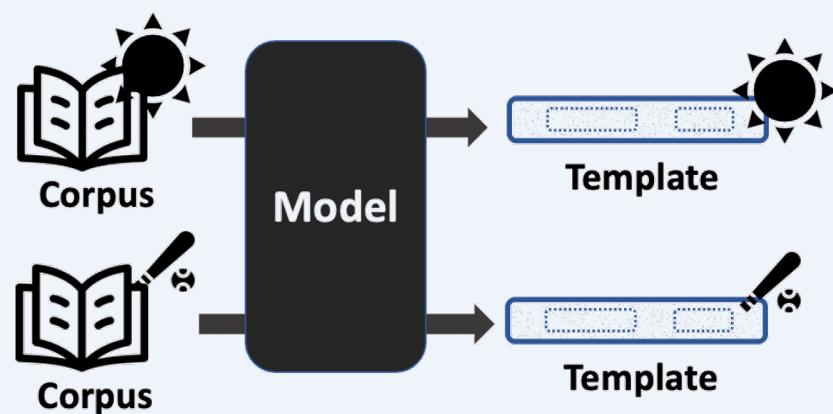


人手でテンプレートを作成し
スロット情報を付与。
スロットに自動で単語を当てはめる

強みは正確さと内容の一貫性
弱みはスケーラビリティのなさ

- 時間がかかる
- 手間がかかる
- ドメイン知識を要する

自動導出・活用



テンプレートを自動で導出し
モデルに組み込む

近年はこちらに注目
いかに生成文の質を下げずに
テンプレートを自動導出・
活用できるが重要なテーマ

コンテンツ

1. 自然言語生成入門

- 自然言語生成とは
- 近年の潮流
- 実用化されている自然言語生成技術とその裏側

2. 自然言語生成とテンプレート

- テンプレートとは
- なぜテンプレート？
- テンプレートベースの文生成手法に求められること

3. テンプレートの導出と活用

- どう獲得し、使うのか？
- 現状と今後の展望

テンプレートベースの文生成

テンプレートの導出および活用法について、簡単な既存研究の紹介とともに自分の

研究テーマ①②

をお話しします



This train will leave
at time.

モデル

This train will leave at 12:00.

第一ステップ

コーパスからテン
プレートを導出

第二ステップ

導出したテンプ
レートを活用し
文を生成

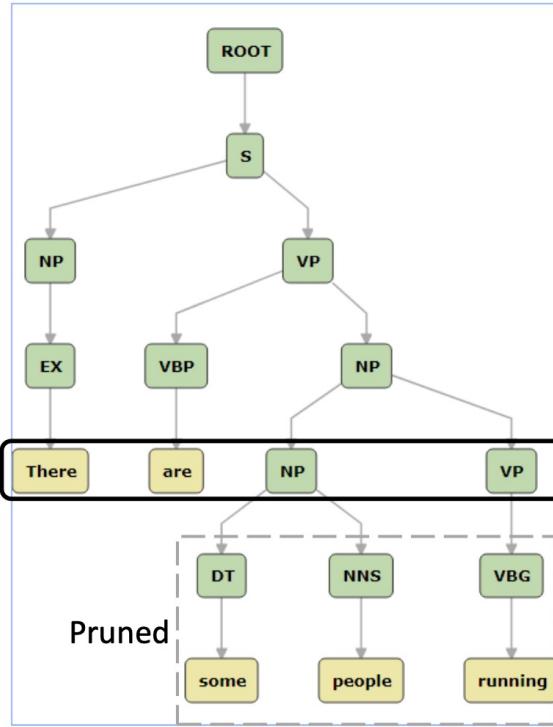
ハードテンプレートのお話

研究テーマ①
テンプレートを抽出する

研究テーマ②
テンプレートのスロット
に単語を埋める

導出編: 既存研究

抽出する



学習事例を木構造に変換し、特定の深さの系列をテンプレートとする

[Yang et al., 2020]

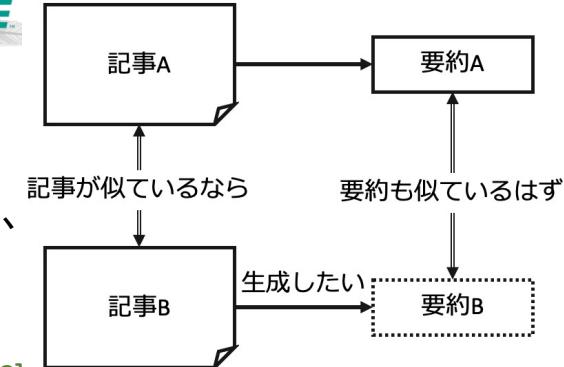
検索・探索する

情報検索システム

近傍探索ライブラリ Faiss

似ている入力事例があったら、
その出力事例が参考になる
(=テンプレートになる)

[Cao et al., 2018, Khandelwal et al., 2020]



学習する

The — is a — providing
— is an — serving
... is an expensive — offering
food cuisine price range ...
cuisine foods with a price bracket ...
foods and has a pricing ...
located in the Its customer rating is ...
located near near Their customer rating is ...
near ... Customers have rated it ...

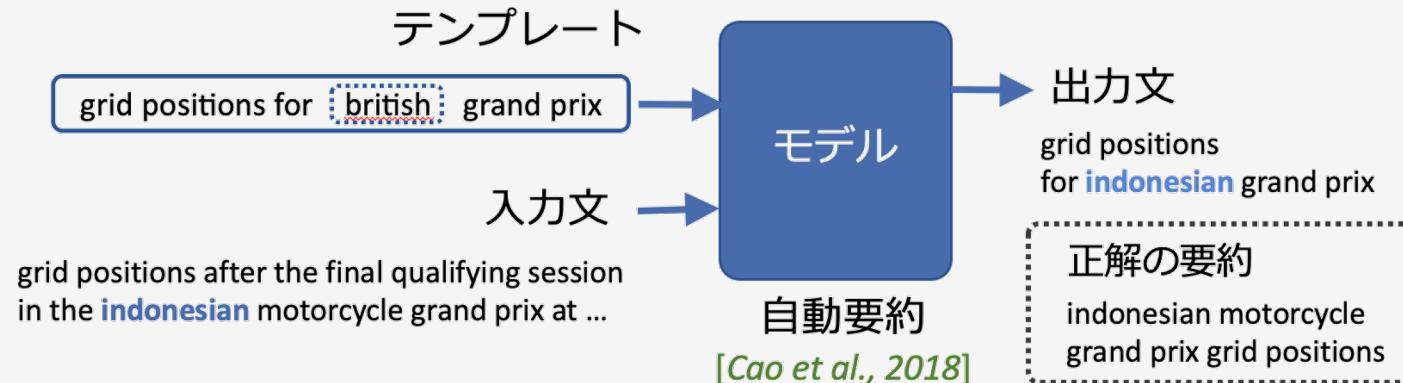
[Wiseman et al., 2019]

ニューラル隠れセミマルコフモデルにより
文中の単語の遷移パターンをテンプレート化

活用編: 既存研究

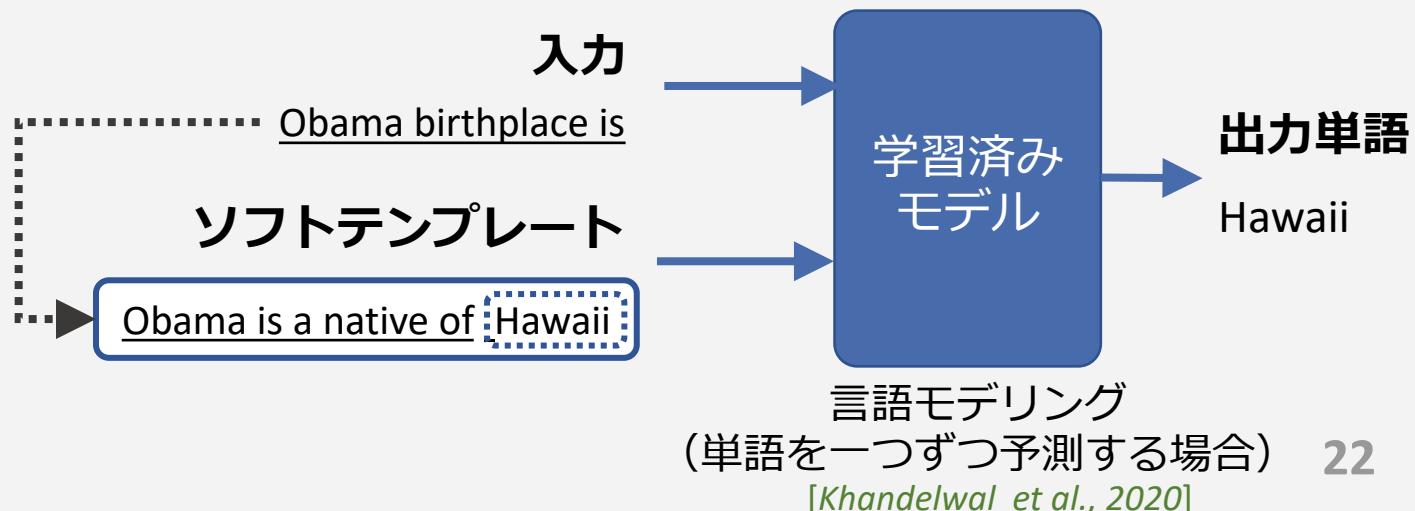
スタンダードな学習方法

入力文とテンプレートをそれぞれエンコード（ベクトル化）して文を生成



推論時のモデルの出力だけを制御する手法も近年提案されている

入力とのベクトルの類似度が高いngramの事例を探索



Construction of a Corpus of Rhetorical Devices in Slogans and Structural Analysis of Antitheses.

Ayana Niwa, Naoaki Okazaki, Kohei Wakimoto, Keisuke Nishiguchi, and Masataka Mouri.

ACM Transactions on Asian Low-Resource Language Information Processing (TALLIP),
Association for Computing Machinery, 2021

どんな研究？

対句を含む文の生成に向けて、文構造テンプレートを抽出する
テンプレートは対句の言語的性質を利用した単語の除去によって
作成

$$\text{ABCDE} - \text{BD} = \text{A } \square \text{ C } \square \text{ E}$$

やったこと

1. 対句コーパスの構築
2. 対句範囲の同定のための構造解析手法の開発

本研究の背景

● 言語の表現力を高める高度な作文技術 修辞技法

—— 体言止め ——
暮らすなら、**都心**。
名詞

—— 比喩 ——
雲に乗った**ような**気持ちだ
例える

—— 反復 ——
人民の、**人民**による、**人民**のための政治。
繰り返す

—— 押韻 ——
セ**ブン** イレ**ブン**、いい気**分**
bun bun bun

品詞や意味、文構造、音韻上の制約をもち、
ある種の文の「型」が存在することが多い

テンプレートを用いた生成が有効

本研究の背景

- 本研究では、数ある修辞技法の中でも文構造的・意味的制約を併せ持つ（=高度な制御を要する）対句の生成に着目

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表現。

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。



反義関係を持つ単語が含まれる

反義関係を持つ単語 対義語

反義関係: 反対あるいは対照的な意味を持つ単語間の関係性
(対義語)



反義関係には多様な関係性が含まれる

- 二律相反のペア (素数—非素数)
- 相互の役割などの関係性によるもの (生徒—教師)
- ある基点を挟んだ相対的・対照的な概念を表すもの (明日—昨日)
- などなど...

この対義語の語彙知識を獲得することも対句生成には必須

- 既存の辞書では不十分

本研究の背景

- 本研究では、数ある修辞技法の中でも文構造的・意味的制約を併せ持つ（=高度な制御を要する）対句の生成に着目

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表現。

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。



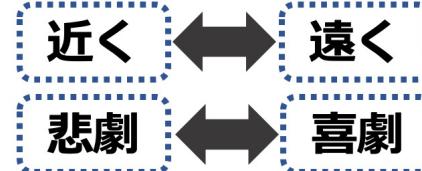
ゴール

生成に必要な以下の二つの知識を抽出すること

① テンプレート

人生は、□□□だが、□□□である。

② スロットを埋める単語候補
→ 対義語の語彙知識



本研究の背景

- 本研究では、数ある修辞技法の中でも文構造的・意味的制約を併せ持つ（=高度な制御を要する）**対句**の生成に着目

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表現。

人生は、近くて見ると悲劇だが、遠くから見れば喜劇である。

方法

対句構造解析を行う

{ 近くて見ると悲劇
遠くから見れば喜劇

本研究でやったこと

- 対句コーパスの構築
- 構造解析手法の提案

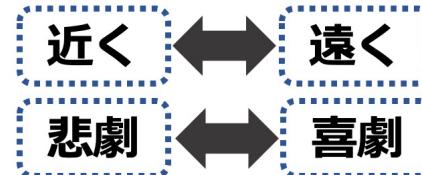
ゴール

生成に必要な以下の二つの知識を抽出すること

① テンプレート

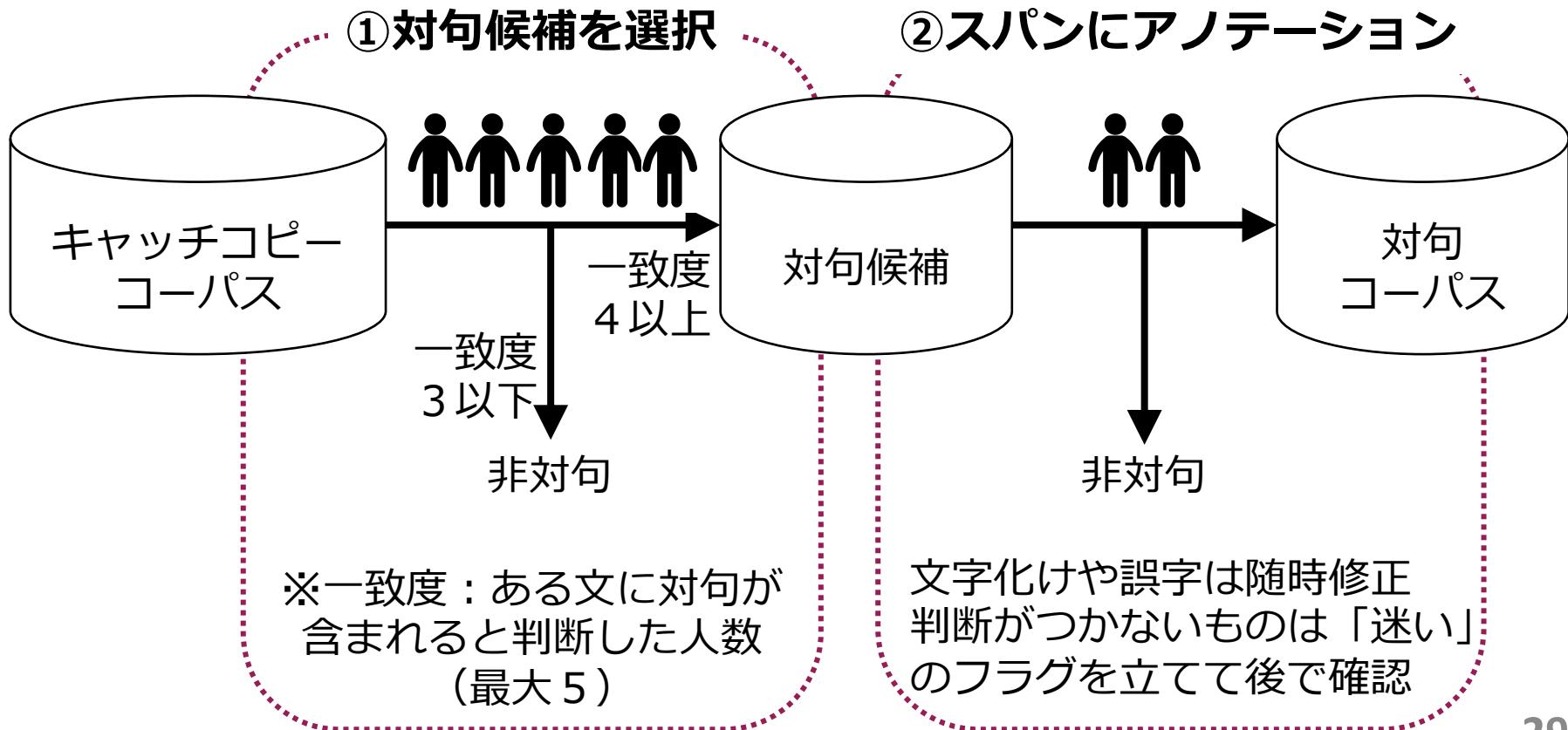
人生は、□□□だが、□□□である。

② スロットを埋める単語候補
→ 対義語の語彙知識



対句コーパスの構築

クラウドソーシングを利用し、キャッチコピーコーパスに
対句構造でアノテートした**対句コーパス**を構築



対句構造解析に有用な 3つの言語的性質

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。

● 句構造の類似性



Dependency treeに基づく句構造の比較

● 句の可換性

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。

● 句の意味的対照性

- a. <遠く - 近く>
- b. <喜劇 - 悲劇>

対義語ペアを含む

対句構造解析における困難な性質

対句は特定の接続詞を持たない任意の文構造で用いられる

[水を汲みに行く 5 時間が消えた]時、[学校に行く 5 時間が生まれた]。

[大人が幸せでない]国で、[子どもが幸せになれる]わけがない。

[初めての「おいしい」]より、[二度目の「おいしい」]。

→対句範囲の
手がかりがない

対句と共通する性質を持つ並列句の場合

うちの息子は[厳父]と[岳父]の違いも知らない、愚息だ

句を連接させる働きのある等位接続詞などの語（並列キー）を伴う

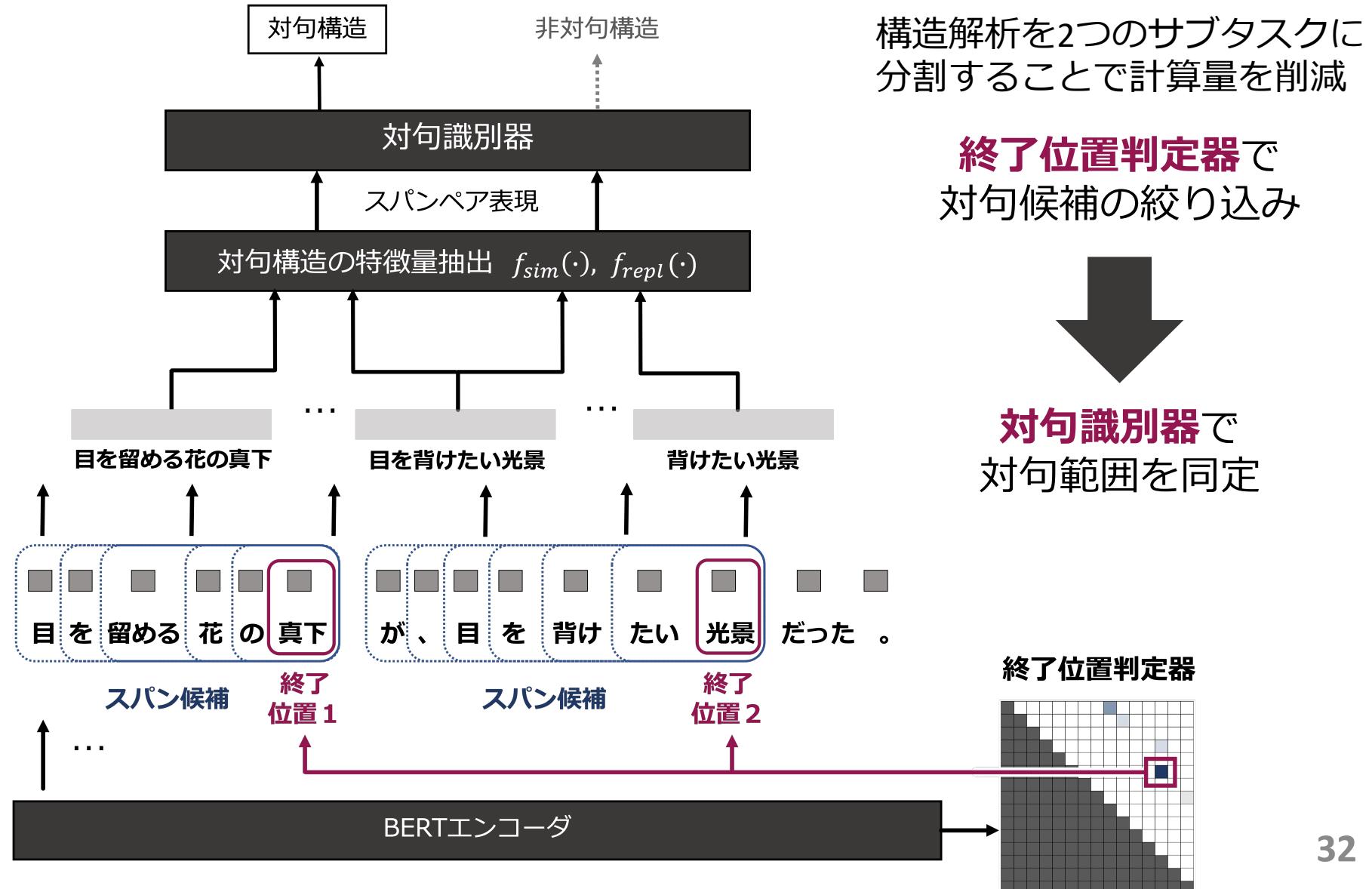


対句構造を特定の構文規則に当てはめて解析することは困難

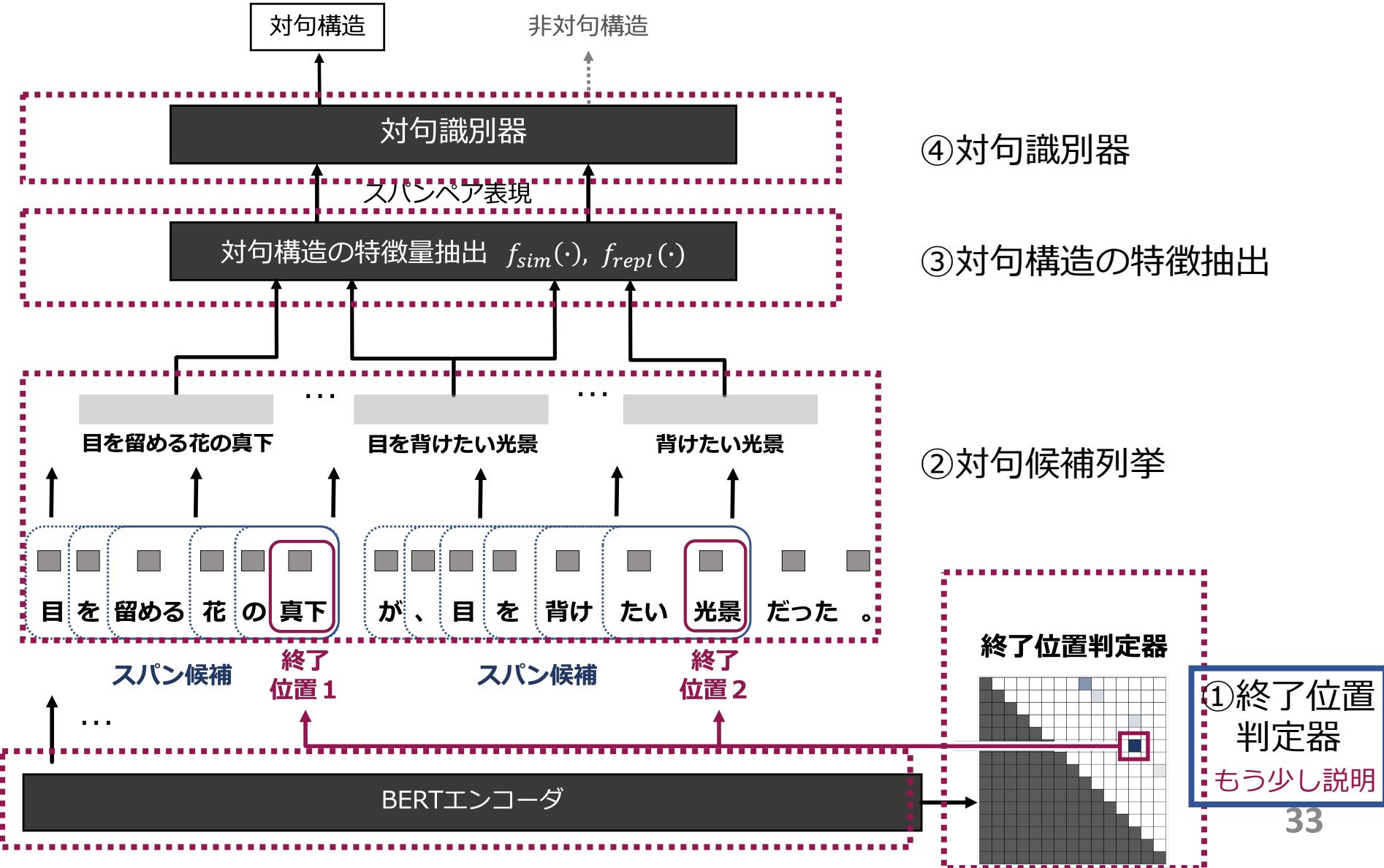
並列構造の解析と比較すると計算量が問題になる



提案手法の概要

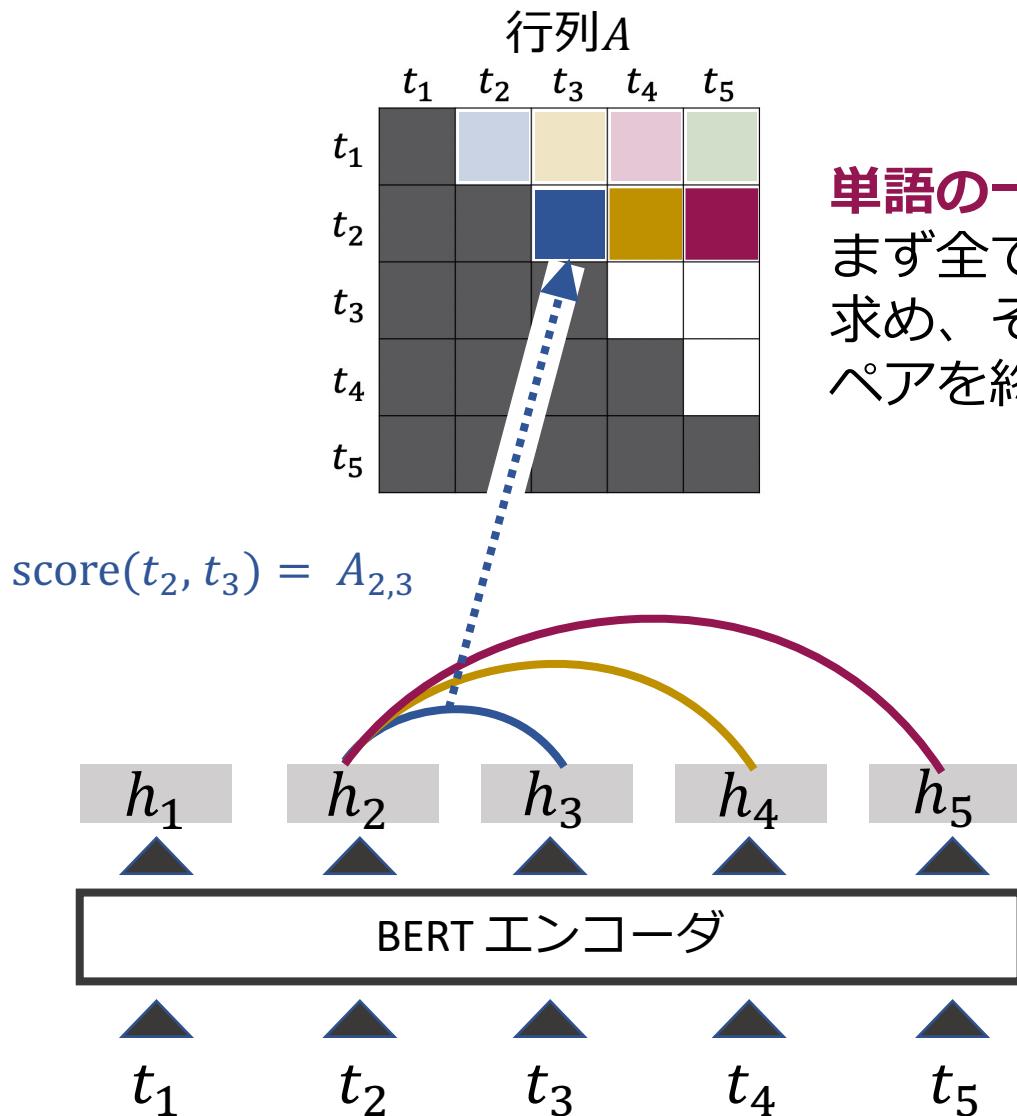


提案手法の概要



提案手法

①終了位置判定器



単語の一対一の対応関係を見るために、
まず全てのトークン間の関係性スコアを
求め、その内最大スコアを持つトークン
ペアを終了位置ペアとする

Notations

H : BERTエンコーダの出力系列
 $H = (h_c, h_1, \dots, h_N)$
 t_i : i 番目のトークン

提案手法の長所



計算量削減



単語の対応関係を考慮

目を留める花の**真下**が、目を背けたい**光景**だった。

目を留める**花**の**真下**が、目を**背け**たい**光景**だった。

目を**留める**花の**真下**が、**目**を**背け**たい**光景**だった。



類似性・可換性を両方を組み込み
対句が持つ構造の多様性に対応

対句構造解析 予測結果

対句と判定されたスパンペアに対して、
スパンの一致度合いをトークン単位で評価

類似性特徴量 SIM 可換性特徴量 REPL 両方	特徴量	対句識別器		
		P	R	F
比較手法 直接スパンで 絞り込み	SIM	0.554	0.806	0.655
	REPL	0.577	0.807	0.672
	BOTH	0.559	0.807	0.659
提案手法 終了位置で 絞り込み	SIM	0.773	0.780	0.776
	REPL	0.781	0.757	0.769
	BOTH	0.788	0.757	0.772

単語の対応関係を利用した
対句候補の絞り込み

提案手法は単語の対応関係を考慮することで比較手法よりも

F値が9ポイント以上向上

提案手法は比較手法に比べて平均学習時間を87%以上削減できる

対句の言語知識の抽出

実験に用いたデータ以外のキャッチコピー約10万件に学習済みモデルを適用
ランダムにサンプリングされた100件のキャッチコピーに対して人手評価

- 比較・逆説表現など物事を対比させるのに有効な特定の接続詞を持つテンプレートが獲得できた

〔スパン1〕ではなく 〔スパン2〕でした。

必要なのは、〔スパン1〕より 〔スパン2〕。

〔スパン1〕けど、〔スパン2〕。

〔スパン1〕より、〔スパン2〕。

〔スパン1〕のに、〔スパン2〕。

〔スパン1〕より 〔スパン2〕を考える。

〔スパン1〕けど、〔スパン2〕なあ。

〔スパン1〕より 〔スパン2〕を。

- 100対もの既存の辞書にはない対義語ペアを獲得できた

未来 ←→ 過去

社員 ←→ アルバイト

競合 ←→ 協力

お金 ←→ カード

年齢 ←→ 肌年齢

愛人 ←→ 恋人

研究① まとめ

本研究のまとめ

対句範囲の同定により、対句を含む文のテンプレートと語彙知識を抽出するための対句構造解析を行った

- 手がかりのない対句範囲を同定するため 2 つのサブタスクに取り組んだ
- 提案手法は、効率的なスパンの絞り込みと単語の対応関係を踏まえた識別により、比較手法に比べ短い学習時間で予測精度を上回った
- 対句を含む文を生成するためのテンプレートと語彙知識が抽出できた

本研究におけるテンプレートまとめ

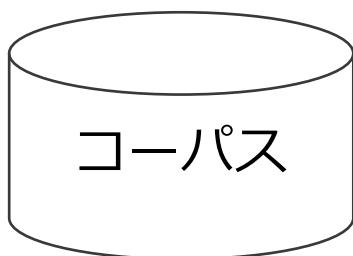
構造解析の結果をもとに、元の文からテンプレートとスロットに当てはめる単語候補（対義語）を抽出
対句の言語的性質を活用するアプローチをとった

テンプレートベースの文生成

テンプレートの導出および活用法について、簡単な既存研究の紹介とともに自分の

研究テーマ①②

をお話しします



This train will leave
at time.

モデル

This train will leave at 12:00.



第一ステップ

コーパスからテン
プレートを導出

第二ステップ

導出したテンプ
レートを活用し
文を生成

研究テーマ①

テンプレートを抽出する

研究テーマ②

テンプレートのスロット
に単語を埋める

Predicting Antonyms in Context using BERT

Ayana Niwa, Keisuke Nishiguchi, and Naoaki Okazaki.

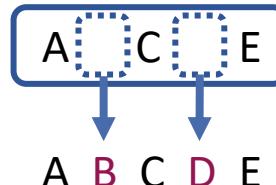
Proceedings of the 14th International Conference on Natural Language Generation
(INLG2021), Association for Computational Linguistics, 2021

どんな研究？

先ほどの研究では、対句生成のためのテンプレートとスロットに埋めるべき対義語の候補を獲得できた。

しかし、実際の生成時には**適切な対義語ペアとテンプレートの組み合わせ (=文脈)**を考慮する必要がある

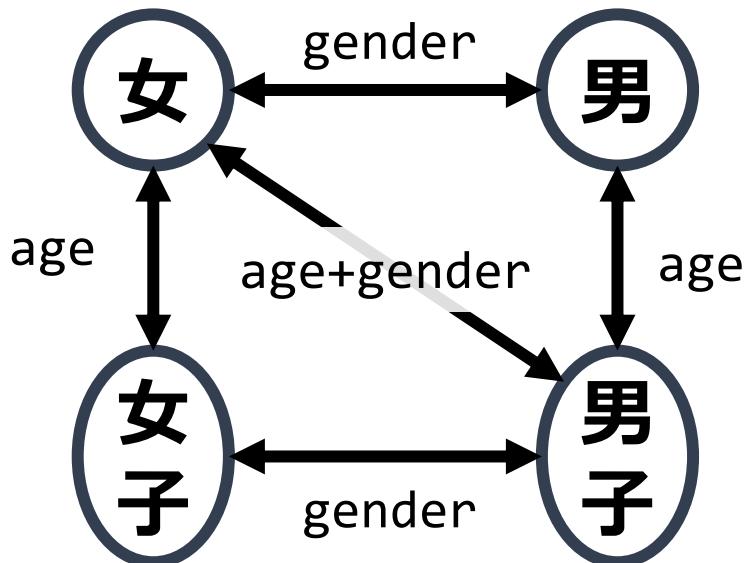
そこで、文脈に依存する対義語を予測するため、テンプレートに基づく対義語予測タスクに取り組んだ



反義関係がもつ一対多対応

単語が意味的対照性を持つには、少なくとも一つの対照的な性質を持ってば良い

[Leech, 1976]



上記を踏まえると…

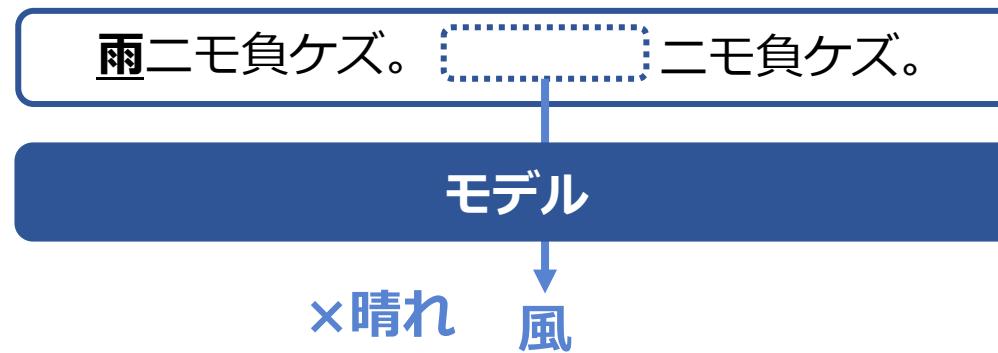
- 一つの単語に対する対義語は、対照的な性質を持つ**複数の単語が候補になりうる**
- さらに、複数の対義語候補の中でも対義語ペアが表現されている**文脈**によって適切なものは異なる

したがって、**対義語はその文脈とともに考えられるべき**
しかし、今までの対義語予測の研究では文脈は考慮されてこなかった

本研究で提案するタスク

そこで本研究では、テンプレートを活用し、
新タスクとして**文脈を考慮した対義語の穴埋め問題**を提案

スロット部分に当てはまる対義語を予測



●このタスクは、以下の二つの能力が求められる

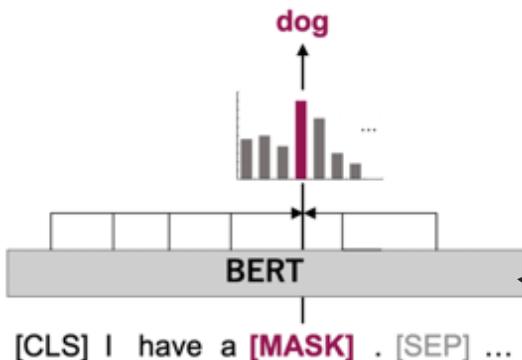
- (1) 単語間の**対照的な性質**を捉え、**文脈に合った対義語**を予測すること
 - スロット間の意味的対照性
 - テンプレートとスロットの意味的一貫性
- (2) 穴埋め後の文としての**自然さ**を考慮すること
 - テンプレートとスロットの文法的一貫性

アーキテクチャ

- 事前学習およびファインチューニングのアプローチが多くのタスクで有効性が示されていることから、モデルのアーキテクチャにBERTを採用

BERT: Transformerエンコーダを12/24層重ねた事前学習済み巨大モデルのこと

事前学習タスクのひとつ
マスク言語モデル



この形式のタスクが可能なのは、BERTが**双方向の文脈を考慮**できるから



→テンプレートのスロットの穴埋めに適している
[MASK]をスロットと考えれば良い

しかしながら、**学習データ** (i.e., 対義語ペアを対照的な意味合いで表現している文章) の収集は容易でない

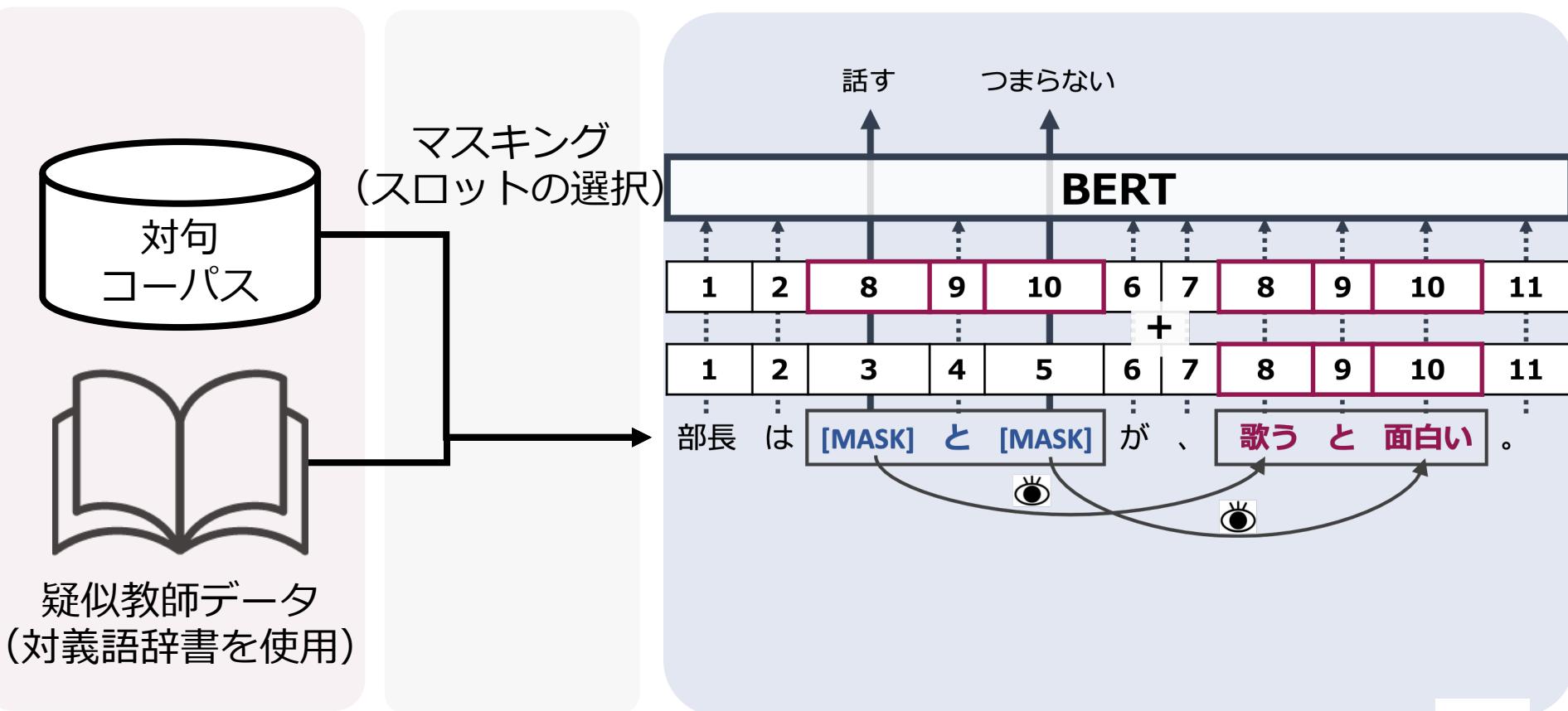
→どう学習データを収集し、その少ないデータをもとに効率的に学習するかが鍵

提案手法 概要

何を入力＝テンプレートとするか？

どのように学習事例を作成するか？

どのようにBERTを対義語予測に適応させるか？

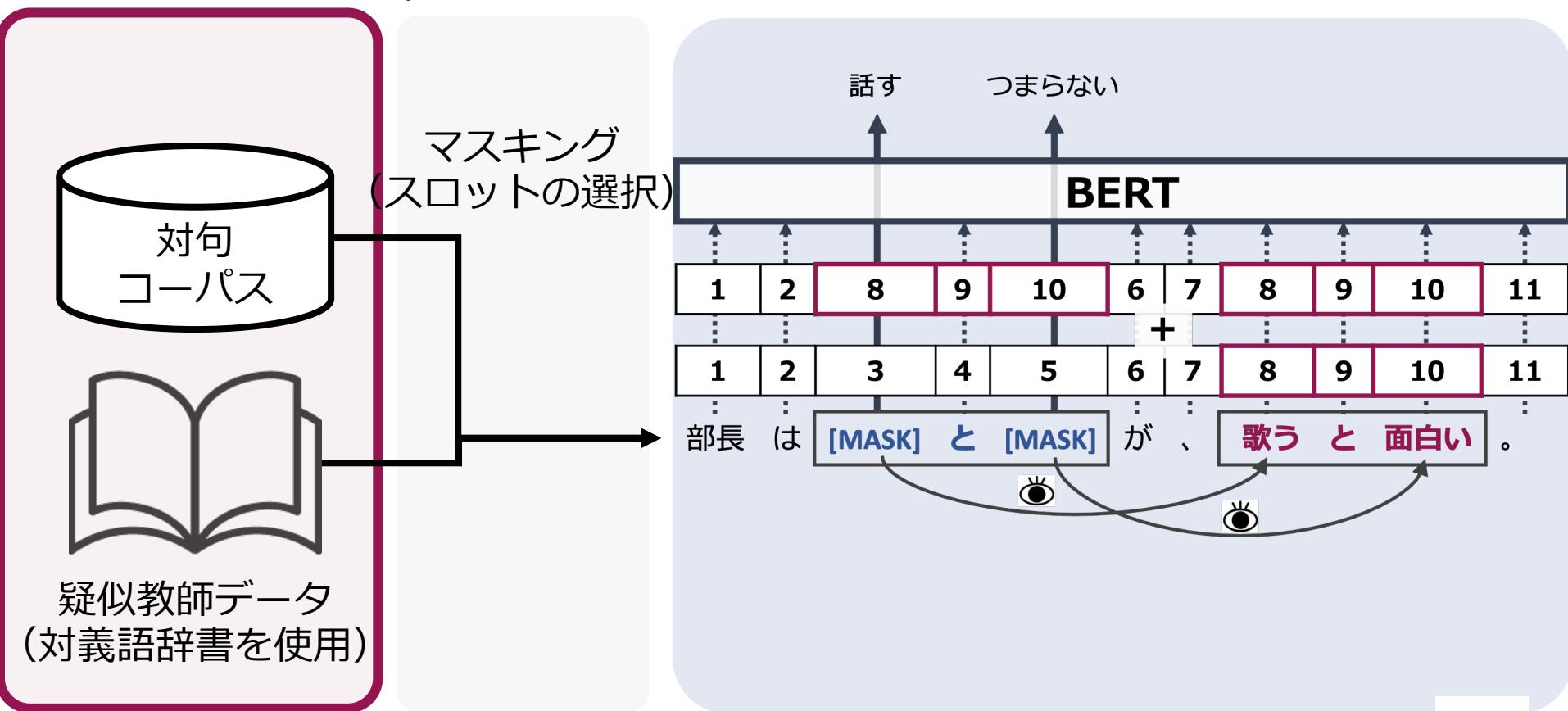


提案手法 概要

何を入力＝テンプレートとするか？

どのように学習事例を作成するか？

どのようにBERTを対義語予測に適応させるか？



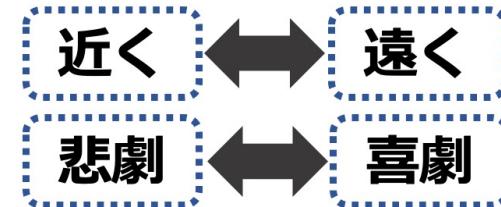
提案手法 対句コーパスの活用

本研究では、対句コーパスを使って
BERTを対義語予測タスクに適応させる

対句コーパスを対義語予測に用いるメリット

- ①対義語ペアが対比的な文脈で表現されている
- ②一つの対句データから複数の対義語ペアが獲得可能

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。



提案手法 疑似教師データの収集

対句を利用した教師データだけでは不足する可能性がある

→**対義語辞書内に含まれる対義語ペアを自動アノテーションすることで、疑似教師データを自動収集する。**

- ・ 辞書内にある対義語ペアを含めば、文脈的にも対照的に表現されていると仮定



対義語辞書

反対語・対立語辞典（三省堂）

愛情と**書いて**、ドライとは**読み**ません。
書ける、**読める**、伝えられる
「文章を**読む**のが苦手.....」「文章を**書く**のに
自信がない.....」というあなたへ

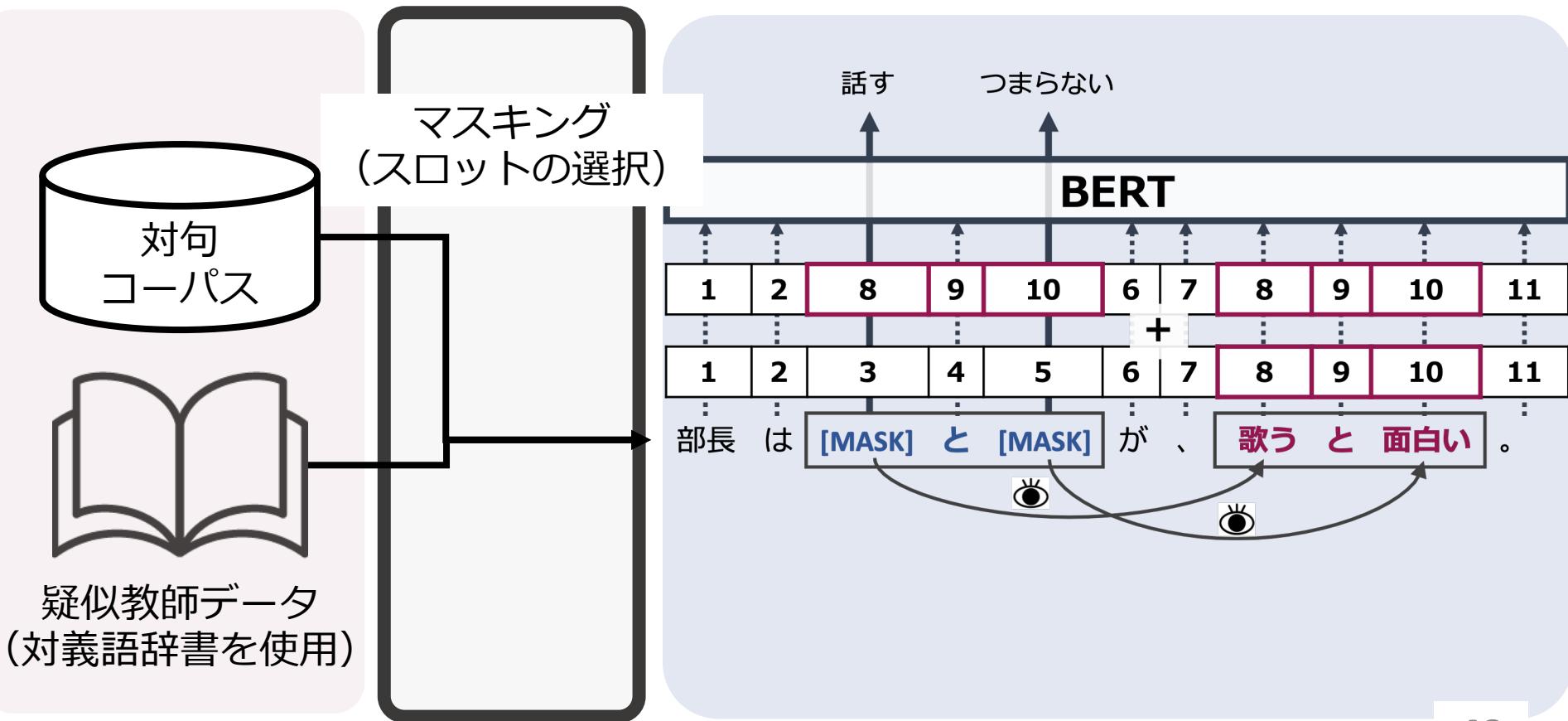
これにより、対義語の語彙知識をBERTに与えることができる

提案手法 概要

何を入力＝テンプレートとするか？

どのように学習事例を作成するか？

どのようにBERTを対義語予測に適応させるか？



提案手法 対照的マスキング

対句ペアの差分となる単語を片方の句ずつマスキングを行うことで対義語予測に特化したテンプレートを作成する

父は保険で楽をした。母は保険で苦労した。

- 学習事例 1 [MASK]は保険で[MASK] [MASK]した。母は保険で苦労した。
- 学習事例 2 父は保険で楽をした。 [MASK]は保険で [MASK]した。

※元々のBERTはランダムにトークンをマスクする（ランダムマスキング）

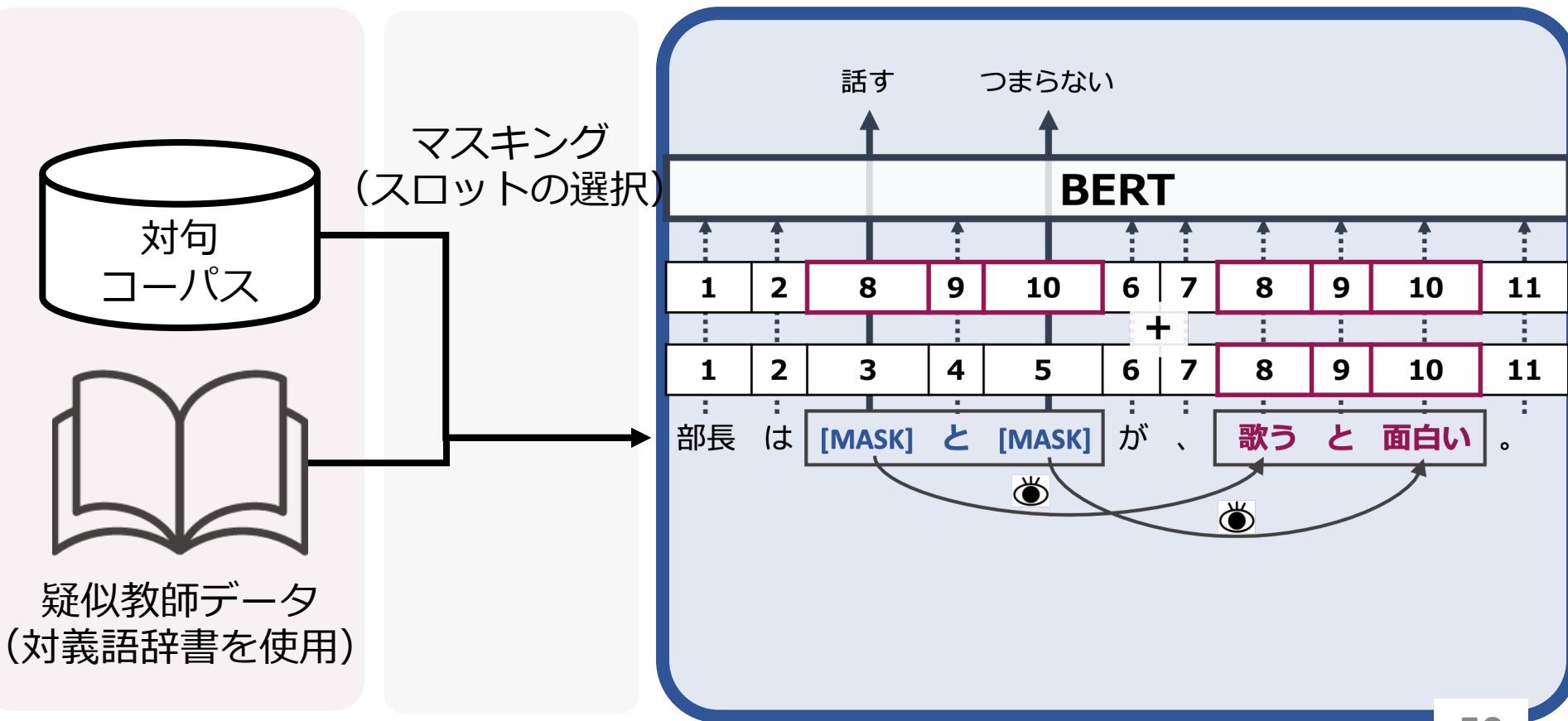
これにより、対義語穴埋めのためのスロット位置を選択的に決定

提案手法 概要

何を入力 = テンプレートとするか？

どのように 学習事例を 作成するか？

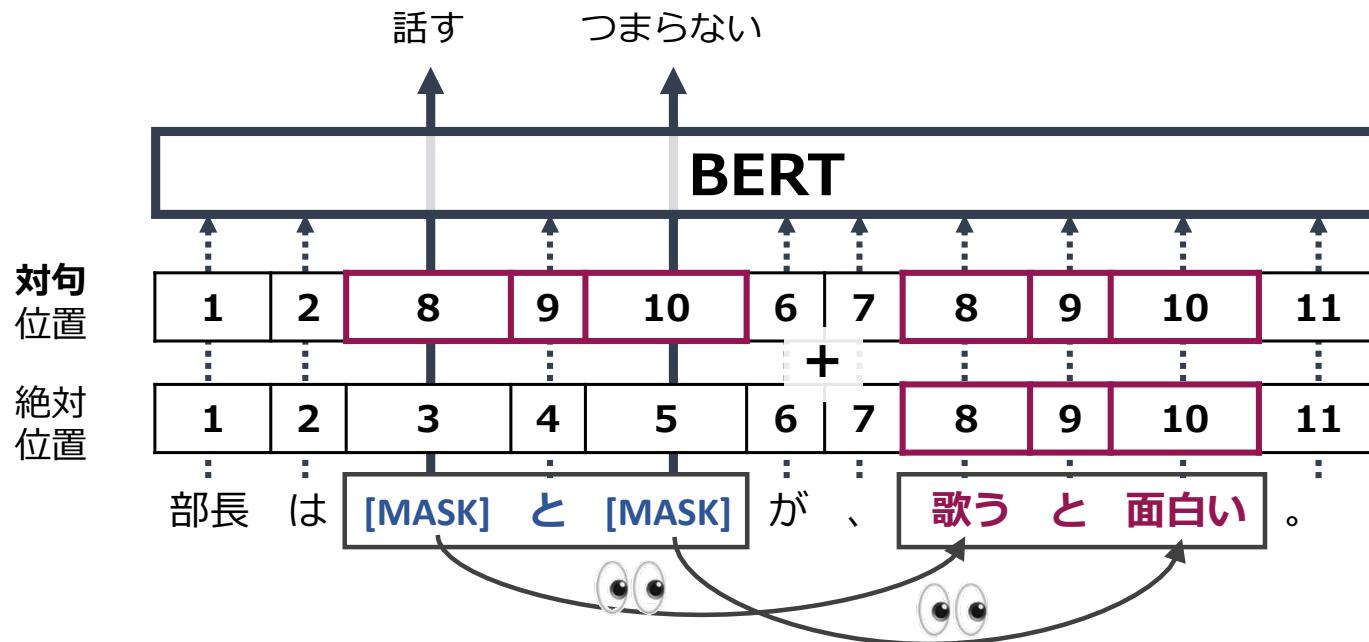
どのようにBERTを対義語予測に適応させるか？



提案手法

対句位置エンコーディング

対義語を予測するには、対となる単語の情報を踏まえることが必要
→ 位置エンコーディングを拡張し、対句の位置をBERTに教える



対句構造を構成するスパン $[i, j]$ 、 $[k, l]$
スパン $[i, j]$ が[MASK]トークンを含む
スパン $[i, j]$ に $[k, l]$ が対応していることを示す

$$対句位置 a_t = \begin{cases} k + \left\lfloor \frac{(l-k)(t-i)}{j-i} \right\rfloor & (a_t \in [i, j]) \\ t & (otherwise) \end{cases}$$

提案手法の長所



データ収集の難しさに対応するため、対義語を対照的な文脈で表現している保証のある対句データを有効活用



BERTを用いることで文脈を考慮してテンプレートのスロットを埋められる

[] は保険で [] [] した。母は保険で苦労した。



対句位置エンコーディングで対となる単語情報を、疑似教師データで対義語の語彙知識を学習可能

実験結果

マスクされたトークン（スロット）に対して正解が予測されたか否かを正解率で評価

	All		Word level	
	Acc@1	Acc@10	Acc@1	Acc@n
辞書を用いた手法	-	-	9.6	-
事前学習済みBERT (ファインチューニングなし)	15.0	40.9	15.7	39.1
事前学習済みBERT (対句データでファインチューニング)	24.4	51.4	25.0	44.4
- ランダムマスキング + 対照的マスキング	28.8	52.6	27.4	47.4
+ 対句位置エンコーディング	28.7	53.5	27.4	48.0
+ 擬似教師データ	29.3	53.8	30.4	49.1
人手作成 (最もスコアが低い場合)	-	-	31.5	52.3
人手作成 (最もスコアが高い場合)	-	-	34.5	59.1
人手作成 (3人の回答を統合した場合)	-	-	51.8	66.6

提案手法により性能の向上が見られる

- 対句データでファインチューニングすることや対となる単語情報の埋め込みなど、**対義語予測に特化した学習の重要性**

実験結果

人手で作成したデータとも比較した

- 事例一つにつき 3人のアノテータが最大5つの解答候補を考えた

	全データ		単語レベル	
	Acc@1	Acc@10	Acc@1	Acc@n
辞書を用いた手法	-	-	9.6	-
事前学習済みBERT (ファインチューニングなし)	15.0	40.9	15.7	39.1
事前学習済みBERT (対句データでファインチューニング)	24.4	51.4	25.0	44.4
- ランダムマスキング + 対照的マスキング	28.8	52.6	27.4	47.4
+ 対句位置エンコーディング	28.7	53.5	27.4	48.0
+ 擬似教師データ	29.3	53.8	30.4	49.1
人手作成 (最もスコアが低い場合)	-	-	31.5	52.3
人手作成 (最もスコアが高い場合)	-	-	34.5	59.1
人手作成 (3人の回答を統合した場合)	-	-	51.8	66.6

人手作成の正解率は、アノテータ間ではらつきがあるものの自動予測に比べては高いものの全体的に低い数値。タスクの難しさを表す
三人の回答を統合して評価すると、特にtop-1正解率が大きく向上
• 複数通りの回答候補が存在するため (数さえ出せれば正解は当てられる)

主観評価

人手で予測した単語と提案手法が予測した単語をそれぞれランダムに100件抽出し、意味的対照性と文の自然さを人手で評価

[再掲] このタスクに求められる能力

- (1) 単語間の**対照的な性質**を捉え、文脈に合った対義語を予測すること
- (2) 穴埋め後の**文としての自然さ**を考慮すること

	意味的対照性	文の自然さ
人手作成	94	90
提案手法	88	85

正解と一致せずとも、対義語として判定できる事例が多く存在

提案手法でも**文脈を踏まえた対義語予測ができる**

- 正解単語1つのみを用いた正解率では正しく評価できない

穴埋め問題として対義語を予測することは現状かなりのレベルでできていることが分かった

出力例

正解が予測できた例

別れの曲だったのに、[MASK]の曲になった。

正解: 出会い

ベースライン 別れ, 最後, 今, 人生

提案手法 出会い, 憧れ, 最高, 始まり

人手作成 1 出会い, 再会, 初恋, 永遠

人手作成 2 出会い, 始まり, 邂逅

人手作成 3 出会い

別れの曲 – 出会いの曲 という離れた位置にある二単語の意味的対照性を捉えられている

正解ではないが、適切な単語を予測した例

地球の環境より、まず[MASK]の環境。 正解：心

ベースライン 宇宙, 水, 地球, 太陽, 植物

提案手法 家族, 私, 周り, トイレ, 家

人手作成 1 自宅, 自分, 部屋, 職場

人手作成 2 自分, 私, 周辺, 室内, 家内

人手作成 3 国, 家庭, 町, 周り

地球 – 自宅、自分 のように「身近さ」の観点で対照的な単語を予測できている

研究② まとめ

本研究のまとめ

与えられた文脈（テンプレート）に対して
対義語を予測する穴埋めタスクに取り組んだ

- BERTを対義語穴埋めに適応させるためのテンプレート作成方法と
して対照的マスキングを提案
- モデルのアーキテクチャに対しては対句位置エンコーディングや
辞書を用いた擬似教師データの作成を行った。
- 実験により提案手法の有効性を示した

本研究におけるテンプレートまとめ

- 文脈を考慮する対義語予測という新しいタスクに、テンプレート
の穴埋め形式で対応
- スロットの位置を選択的に決定することで、モデルが反義関係を
捉えやすくした

テンプレートを用いた 自然言語生成 まとめ

- テンプレートの自動導出および活用法において、対句や対義語といった言語的性質に着目した手法の開発に取り組んできた
- テンプレートは、出力制御の目的だけではなく、モデルを特定の言語現象（反義関係など）に適応させるためにも使える
- 今までの研究の限界は、各言語現象（対句など）以外の文構造に対する一般性がないことが挙げられる

テンプレートを用いた 言語生成の今後の展望

データからテンプレートを学習する研究を進めたい

今まで：対句などの各言語現象特有の性質を利用したテンプレート化

これから：一般化された文構造をテンプレートとして自動導出することで、
任意のドメインにおける文生成に適用できるようにしたい

この時、日本語で起きやすい語順の入れ替え（スクランブリング）も考慮する必要がある

- [太郎が [花子がソウルに住んでいると] 思っている]
- [ソウルに [太郎が [花子が 住んでいると] 思っている]]

文構造の一般化がしづらいドメインにどうアプローチするかも鍵

- 天気ニュースなどは一般化しやすい（例）明日は～から～になるでしょう
- 新聞のようにトピックが多岐にわたるドメインは一般化しづらい

テンプレートを用いた 言語生成の今後の展望

テンプレートの形式を再検討することも重要ではないか

現状

系列テンプレート

- 語順が制限される
- 言葉の階層性が捉えられない

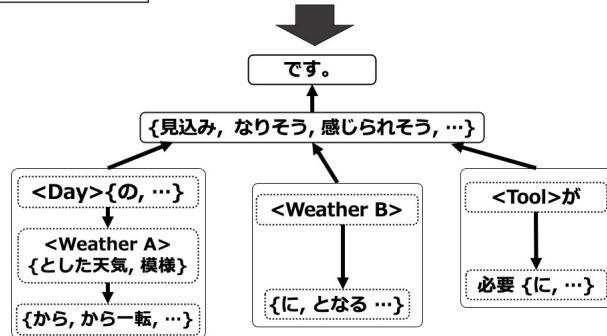
1事例につき1テンプレート

- その事例に最適なテンプレートが見つかるとは限らない
- 出力結果の多様性に欠ける

今後の展望

非系列テンプレート？

対象ドメインの
コーパス
例：気象ニュース
...
今日の～とした天気から～となる見込みです。
昨日の～模様から一転～が必要になりそうです。



1事例につき複数テンプレート
(部品としてのテンプレート?)

□は□に負けた

□を□点差で□

□は□で□



□は□点差で□に負けた

References

[McDonald, 1993] David D. McDonald. "Issues in the Choice of a Source for Natural Language Generation." 1993.

[Marcus et al., 1993] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank." 1993.

[Kurohashi and Nagao, 1998] Sadao Kurohashi and Makoto Nagao. "Building a Japanese parsed corpus while improving the parsing system." 1998.

[Gatt and Belz, 2008] Albert Gatt and Anja Belz. "Attribute selection for referring expression generation: new algorithms and evaluation methods." 2008.

[Koller et al., 2009] Alexander Koller, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, Jon Oberlander, and Kristina Striegnitz. "The software architecture for the first challenge on generating instructions in virtual environments." 2009.

[Vaswani et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." 2017.

References

- [Cao et al., 2018] Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. "Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization." 2018.
- [Yang et al., 2020] Jian Yang, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Ming Zhou. "Improving Neural Machine Translation with Soft Template Prediction." 2020.
- [Wiseman et al., 2019] Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. "Learning Neural Templates for Text Generation." 2018.
- [Khandelwal et al., 2020] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. "Generalization through Memorization: Nearest Neighbor Language Models." 2020.
- [Leech, 1976] Geoffrey Leech. "Semantics." 1976.