

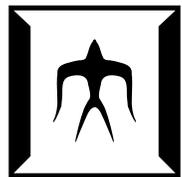
# 自然言語生成における 内容の制御

岡崎 直観

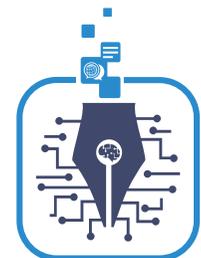
東京工業大学  
情報理工学院

[okazaki@c.titech.ac.jp](mailto:okazaki@c.titech.ac.jp)

<https://www.nlp.c.titech.ac.jp/>

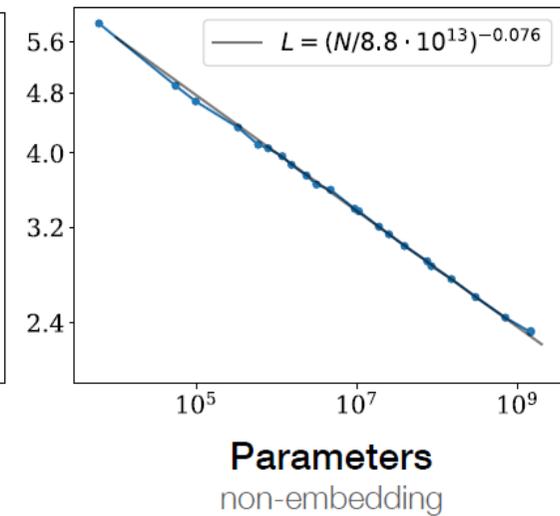
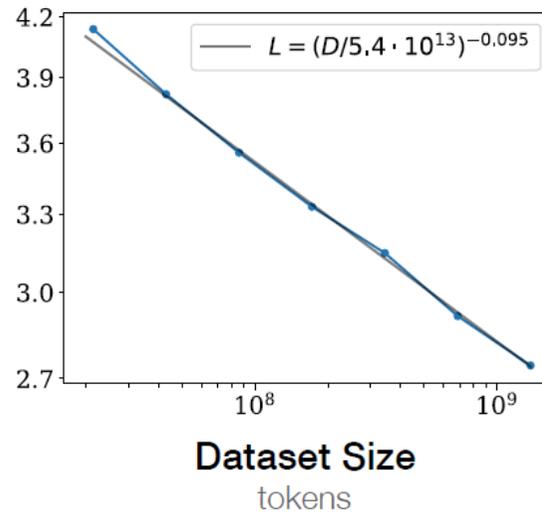
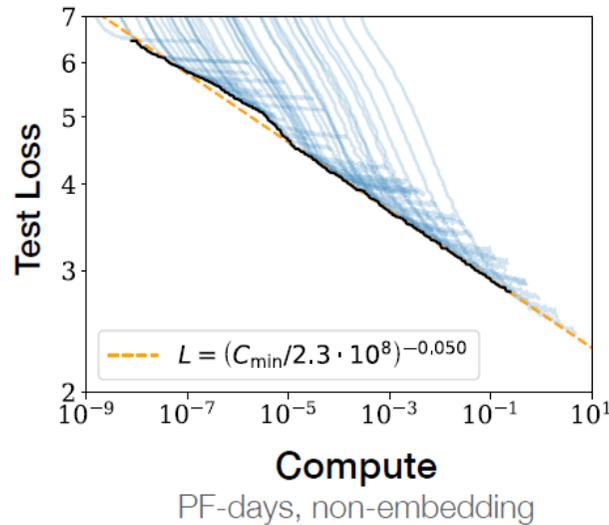
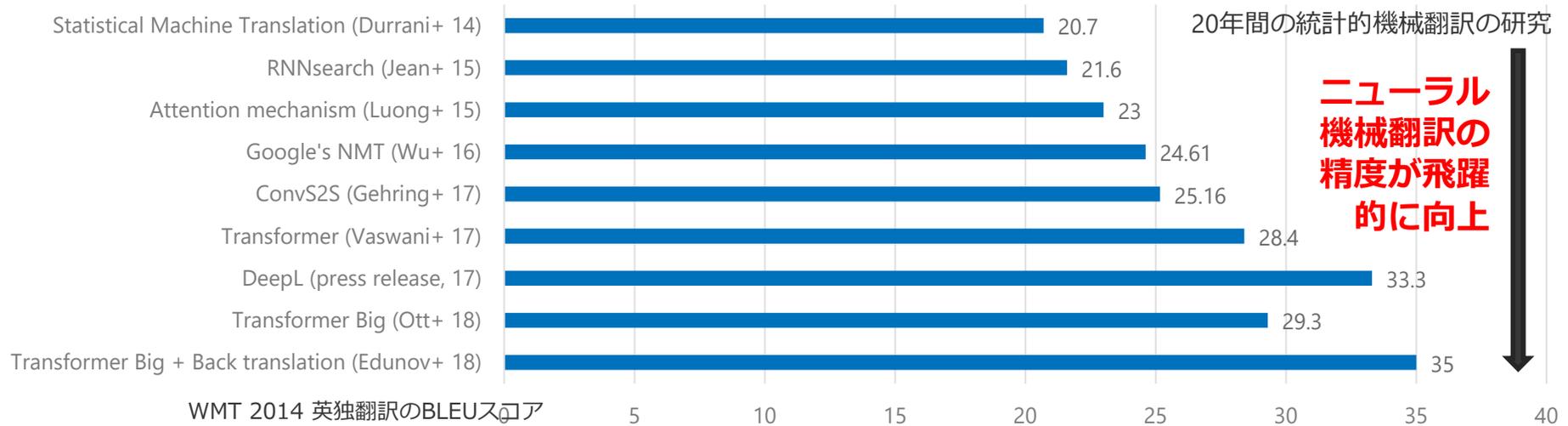


東京工業大学  
Tokyo Institute of Technology



OKAZAKILAB

# 自然言語生成（機械翻訳・言語モデル）の精度向上



言語モデルの精度と**計算能力**、**訓練データ量**、**パラメータ数**の間にべき乗則 (Kaplan+ 2020)

J Kaplan et al. (2020) Scaling Laws for Neural Language Models. *arXiv:2001.08361*.

## 自然言語生成を「制御」したい

- ニューラル機械翻訳は訓練データに依存
  - 大量の訓練データがあれば、高精度な翻訳モデルを構築できる
  - どのような生成結果になるかは、動かしてみるまで分からない
  - 訓練データから少し逸脱したタスクの生成を行うには工夫が要る
- 教師あり学習に基づく生成モデルを活かしながら、所望する出力が得られるようにその振る舞いを制御したい
- テキスト生成の制御
  - 長さの制御
  - 見出しの忠実性の制御
  - 言及すべきキーワードの制御
  - 生成文の構造の制御

# 見出しの自動生成

朝日新聞は一つの記事に対して5種類の見出しを作成している<sup>[1]</sup>

- **Print**: 新聞紙面用
- **Sum** (50): 新幹線向けの要約
- **Large** (26): asahi.comのニュース用
- **Middle** (13): SNS向け
- **Short** (10): デジタルサイネージ向け

※カッコ内の数字は最大文字数



小4算数・理科 過去最高点

国際学力調査

脱ゆとりの成果

自由席 Non-Reserved 新元号は「令和」

Sum

ソーシャルランキング → もっと見る

フェイスブック はてなブックマーク

1 東京都で新たに47人が感染 2798

2 米で黒人男性を警官が射殺 2557

3 京大総長、留学生給付金批判 2256

Middle

[1] 人見 雄太, 田口 雄哉, 田森 秀明, 岡崎 直観, 乾 健太郎. 小規模リソースにおける生成型要約のためのスタイル転移. 言語処理学会第26回年次大会 (NLP2020), pp. 929–932, 2020年3月.

[2] <https://www.asahi.com/articles/ASN6G5D0JN6GUTIL00K.html>

朝日新聞  
DIGITAL

Large

東京都で新たに47人感染 「夜の街」の同じ店で18人<sup>[2]</sup>

新型コロナウイルス

2020年6月14日 16時08分



【動画】東京都の休業要請解除の3段階のステップとは？



「東京アラート」の発動を受けて警戒を呼びかける赤色にライトアップされた東京都庁=2020年8月2日午後11時3分、東京都新宿区、長島一浩撮影

東京都の小池百合子知事は14日、新型コロナウイルスの感染者を新たに47人確認したと発表した。そのうち18人が新宿区内の「夜の街」にある同じ店での感染だという。また、5人は集団感染が起きている武蔵野中央病院（東京都小金井市）関連だという。

都は11日、新型コロナ感染拡大への警戒を呼びかける「東京アラート」を解除した。ただ11日以降、1日あたりの都内での感染者は4日連続で20人を上回っており、再び感染者数が増加している。

# デモ：指定された長さの見出しを生成

The screenshot shows a web browser window displaying a news article on the Jiji Web website. The browser's address bar shows the URL: `jjiiweb.jiji.com/apps/contents/view/20210305/1039/viewtemplate1/jpnpol?spn=1`. The website header features the Jiji Web logo, a search bar with the placeholder text "検索ワードを入れてください", and navigation links for "検索期間", "詳細検索", "ユーザー設定", and "ログアウト". A news ticker at the top displays the headline "巨大IT解体論者を補佐官に＝競争政策担当－バイデン米大統領 (03/06 09:31)".

The main content area is titled "注目キーワード" (注目キーワード) and includes tags for "新型コロナ", "コロナ・日本関係", "ミャンマー情勢", "東京五輪", and "震災10年". The article is categorized under "政治" (政治) and has the title "コロナ終息まで2～3年＝対策分科会の尾身氏 (2021/03/05-17:19)". The article text reads:

政府の新型コロナウイルス感染症対策分科会の尾身茂会長は5日の参院予算委員会で、新型コロナウイルス終息の時期を問われ、季節性インフルエンザと同等の病気と認識されるまで「2～3年」かかるとの見通しを示した。日本維新の会の浅田均氏への答弁。

尾身氏は、今後ワクチン接種が進めば発症や重症化が予防できると分析。今年12月ごろまでに全国民の6～7割の接種が一巡したとしても、依然としてクラスター（感染者集団）や重症化は起こり得ると説明した。

その上で、尾身氏は「さらにもう1年、あるいはさらにもう1年たつと、季節性インフルエンザのように不安感、恐怖心がないということが来る。その時が終息みたいな感じになる」との見方を示した。（了）

The article includes a photo of the government's COVID-19 response subcommittee chair, Shigenori Moriyama, speaking at a press conference. A caption below the photo reads: "参院予算委員会で答弁する政府の新型コロナウイルス感染症対策分科会の尾身茂会長＝5日、国会内".

At the bottom of the article, there are navigation buttons for "< 前の記事" and "次の記事 >". Below the article, there is a section for "関連記事" (関連記事) with a list of related articles:

- > 議員やじ、感染リスクは？＝スポーツ、ライブは大声自粛－「議会の華」も専門家苦言 (02/02-05:42)
- > 自民・安藤高夫氏がコロナ感染 (01/19-12:12)
- > 入院拒否の罰則に反対＝「感染抑止に逆効果」－医学会連合 (01/14-18:52)
- > 予防接種法改正案が衆院で審議入り＝コロナワクチン無料化 (11/10-16:49)
- > 枝野氏、「責任転嫁」と加藤厚労相批判＝コロナ相談目安「誤解」発言 (05/11-18:28)

Sho Takase and Naoaki Okazaki. Positional Encoding to Control Output Sequence Length. 2019. *NAACL*, pp. 3999-4004.

## 問題点：見出し生成が「誤報」を生む

衆院選は14日に投開票される。前回2012年は19人だった県内五つの小選挙区の候補者は、今回14人に減少。少数激戦になった。



**生成された見出し:** 14 候補、最後の訴え **あす**投開票 衆院選

**実際の見出し:** 14候補、最後の訴え **きょう**投開票 **深夜に** 大勢判明

- 赤字で示した箇所は記事（入力）中で言及されていない（**逸脱した見出し**）
- 学習データにも逸脱した見出しが含まれている



- **見出し生成モデルも記事に書かれていないことを含めようとしてしまう**

# デモ：忠実性を改善した見出し生成

The screenshot shows a web browser window displaying a news article on the Jiji Web website. The browser's address bar shows the URL: `jijiweb.jiji.com/apps/contents/view/20210305/861/viewtemplate1/jpnt?spn=1`. The website header includes the Jiji Web logo and a search bar. A navigation menu on the left lists categories like 'Myクリッピング', '時事速報', 'フラッシュ・速報', 'e-World', and '最新ニュース'. The main content area features a '注目キーワード' (注目キーワード) section with tags for '新型コロナ', 'コロナ・日本関係', 'ミャンマー情勢', '東京五輪', and '震災10年'. The article title is '米、マスク論争再び=バイデン氏「原始人」発言に批判 (2021/03/05-15:01)'. The article text discusses the controversy over mask-wearing in the US, mentioning Biden's criticism of 'primitive people' and the reactions of Mississippi Governor Lee and other politicians. An image of Biden wearing a mask is included, with a caption: 'マスクをするバイデン米大統領 = 2月11日、東部メリーランド州ベセスダ (AFP時事)'. Another image shows a man speaking at a podium, with a caption: '米南部ミシシッピ州のリープス知事 = 2020年9月、ワシントン (AFP時事)'. The article concludes with a note that the text is abbreviated (了).

Kazuki Matsumaru, Sho Takase, Naoaki Okazaki. 2020. Improving Truthfulness of Headline Generation. *ACL*, pp. 1335-1346.

## 問題点：見出しの内容を制御したい

台風19号による豪雨災害。氾濫などによる浸水範囲は去年の「西日本豪雨」を超えたほか、土砂災害も1つの台風によるものとしては最も多くなるなど、国が対策の見直しを迫られる記録的な豪雨災害になりました。台風19号で亡くなった人は全国で93人で、3人が行方不明となっています。国土交通省によりますと、台風19号による豪雨で川の堤防が壊れる「決壊」が発生したのは12日時点で7つの県の合わせて71河川、140か所となっています。台風19号による豪雨で発生した土砂災害は、.....



### 見出し例1：（人的被害に焦点）

台風19号による死者は全国で93人、3人が行方不明

### 見出し例2：（物的被害に焦点）

台風19号による土砂災害は821件、住宅被害は8万棟超

### 見出し例3：（避難生活に焦点）

台風の影響により2367人が避難所で生活

# デモ：キーワード指定による見出しの制御

jjji-web | 時事通信社 x 見出し生成デモ x +

jjjiweb.jjji.com/apps/contents/view/20210306/174/viewtemplate1/jpntop?spn=1

検索ワードを入れてください 検索期間 詳細検索 ユーザー設定 ログアウト

Myクリッピング 注目キーワード 新型コロナ コロナ・日本関係 ミャンマー情勢 東京五輪 震災10年

時事速報 フラッシュ・速報 e-World 最新ニュース

トップニュース 政治 経済 企業(業種別) 社会 国際 人事 訃報 スポーツ 全ジャンル

新商品情報 解説・レポート

最新ニュース

トップニュース

政治

経済

企業(業種別)

社会

国際

人事

訃報

スポーツ

全ジャンル

新商品情報

解説・レポート

見出し生成デモ

検索ワードを入れてください

検索期間

詳細検索

ユーザー設定

ログアウト

一米政権 (03/06 10:04)

巨大IT解体論者を補佐官に＝競争政策担当一パ

注目キーワード

新型コロナ

コロナ・日本関係

ミャンマー情勢

東京五輪

震災10年

最新ニュース > トップニュース

## トップニュース

### ごみ収集に自動運転車＝負担軽減、24年度にも導入一環境省 (2021/03/06-05:23)

環境省は、2021年度から自動運転技術を活用したごみ収集の実証を始める。作業員の後ろを自動追尾する機能を持った収集車を使い、住宅地などで作業する際、車両への乗降回数を減らせる仕組みを開発。作業員の負担を軽減し、深刻な人手不足に対応する。3年間かけて技術的な検討や、関係省庁と法令上の手続きを進め、24年度にも市区町村の現場での導入を目指す。

全国のごみ収集作業員は、市区町村職員が約2万人で、民間企業の従事者が約24万人。少子高齢化に加え、身体的負担が大きい収集業務は敬遠されがちで、成り手が不足している。財政的に厳しい自治体では、1人で運転とごみの積み込みをこなす場合がある。

また、近年は高齢者らがごみを出しやすいように、集積所での回収ではなく、戸別収集する自治体が増加。小刻みに車両を進めるため、作業負担が一層大きくなっている。

こうした状況を踏まえ、環境省は自動運転技術を生かした収集の効率化を探る。4月以降、実証に参加する自動車メーカーを公募し、夏ごろ選定。既に一部メーカーが自動追尾機能を備えた車両を開発している。21年度中は公道ではなく専用のテストコースなどで、実際の収集作業員らに協力してもらいながら試行する

自動運転技術を活用したごみ収集のイメージ

現在

乗って運転してごみ運んで大変!

将来

乗り降りなくて楽!

自動追尾

自動運転技術を活用したごみ収集のイメージ

# 文構造の制御（対句の生成）

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。

（対句構造を持つキャッチコピーの生成）

## 1. 対句構造のアノテーション・解析

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。

**反義語・対称語：**（近く，遠く），（悲劇，喜劇）

## 2. 対句構造になるように単語列を生成（単語穴埋め）

必要な時の終わりが、無駄な時の始まりです。

# 対句構造のアノテーション・解析

- 10,108件の対句からなるデータセットを構築
- このデータセットで対句の識別器を学習（F1スコアは0.865）

対句の識別器

対句構造の特徴量抽出（類似性・可換性）

目を留める花の真下

留める花の真下

真下

目を背けたい光景

背けたい光景

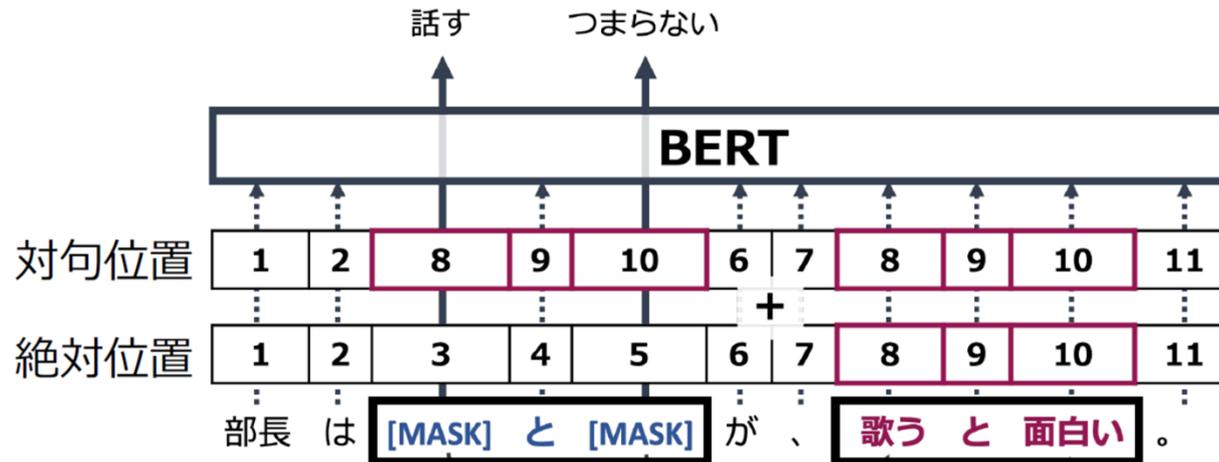
光景

終了位置判定

目を留める花の真下が、目を背けたい光景だった。

# 対句構造になるように単語列を生成

- 疑似教師データなどを用いてBERTをファインチューニングする手法を提案
- 人手評価によると、人間が単語を穴埋めするのと同程度の精度を達成



手法	Hit@1	Hit@10
辞書を用いた手法	9.6	
BERT (ファインチューニング無し)	15.7	39.1
BERT (ファインチューニング有り)	25.0	44.4
提案手法	<b>30.4</b>	<b>49.1</b>
作業員	51.8	66.6

Ayana Niwa, Keisuke Nishiguchi, and Naoaki Okazaki. Predicting Antonyms in Context using BERT. In Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 48–54, Aberdeen, Scotland, UK, August 2021.