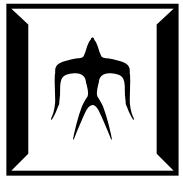
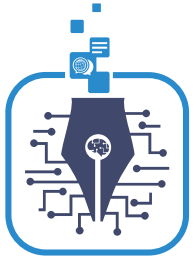
自然言語生成における内容の制御

岡崎　直観

東京工業大学
情報理工学院
okazaki@c.titech.ac.jp
https://www.nlp.c.titech.ac.jp/

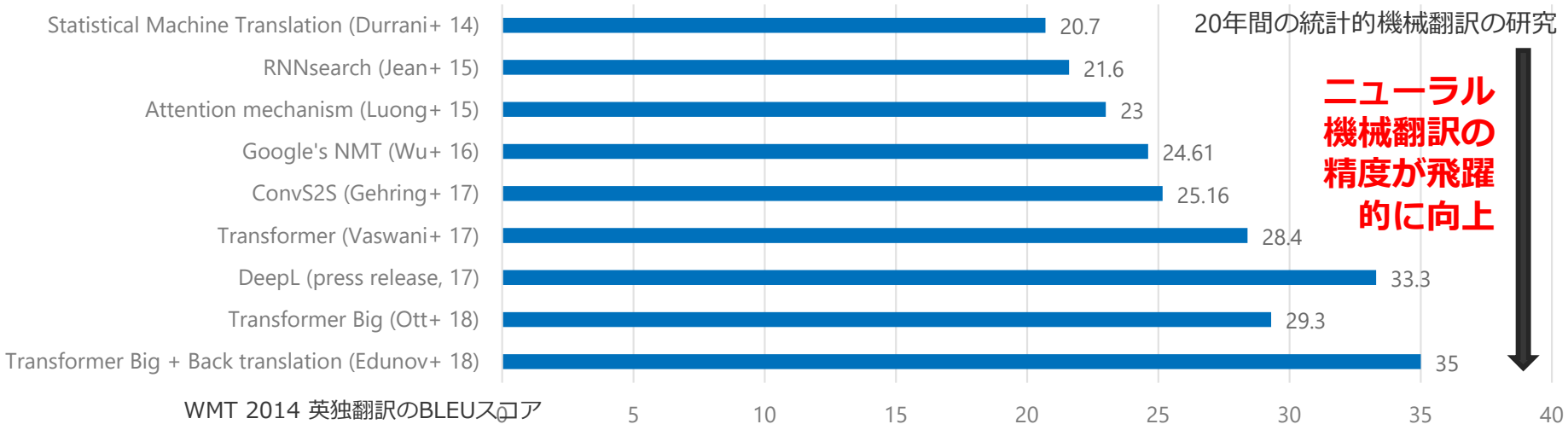


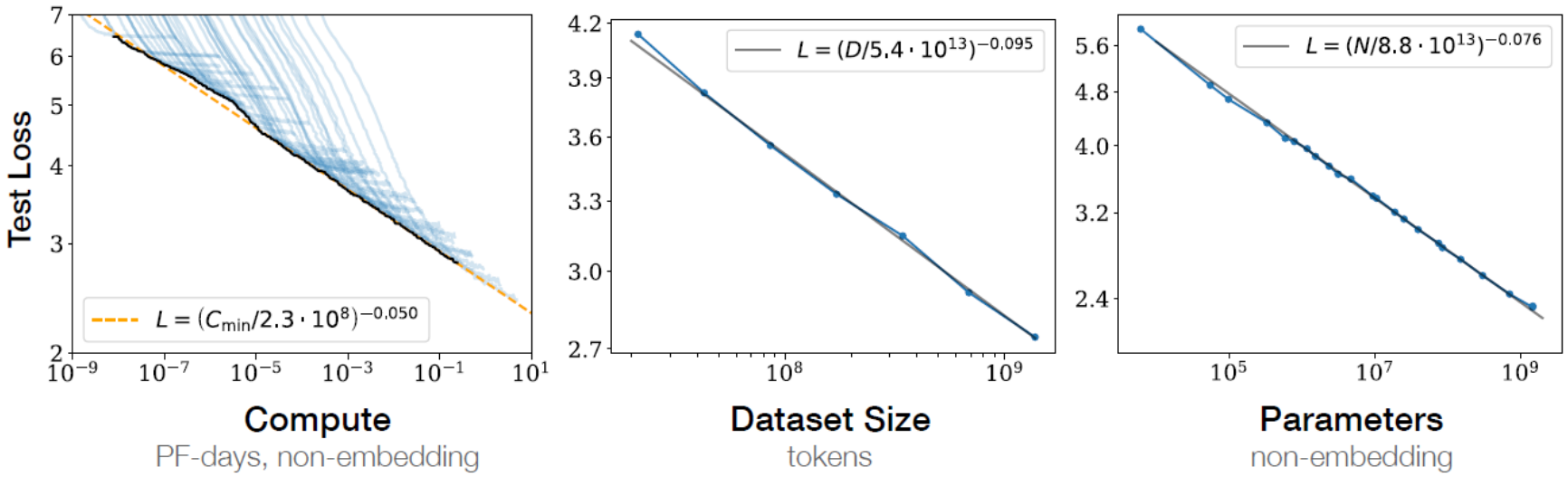
東京工業大学
Tokyo Institute of Technology



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[1]

- **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」

Print

Sum

Middle

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[2]

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6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。

[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

デモ：指定された長さの見出しを生成

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

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

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）

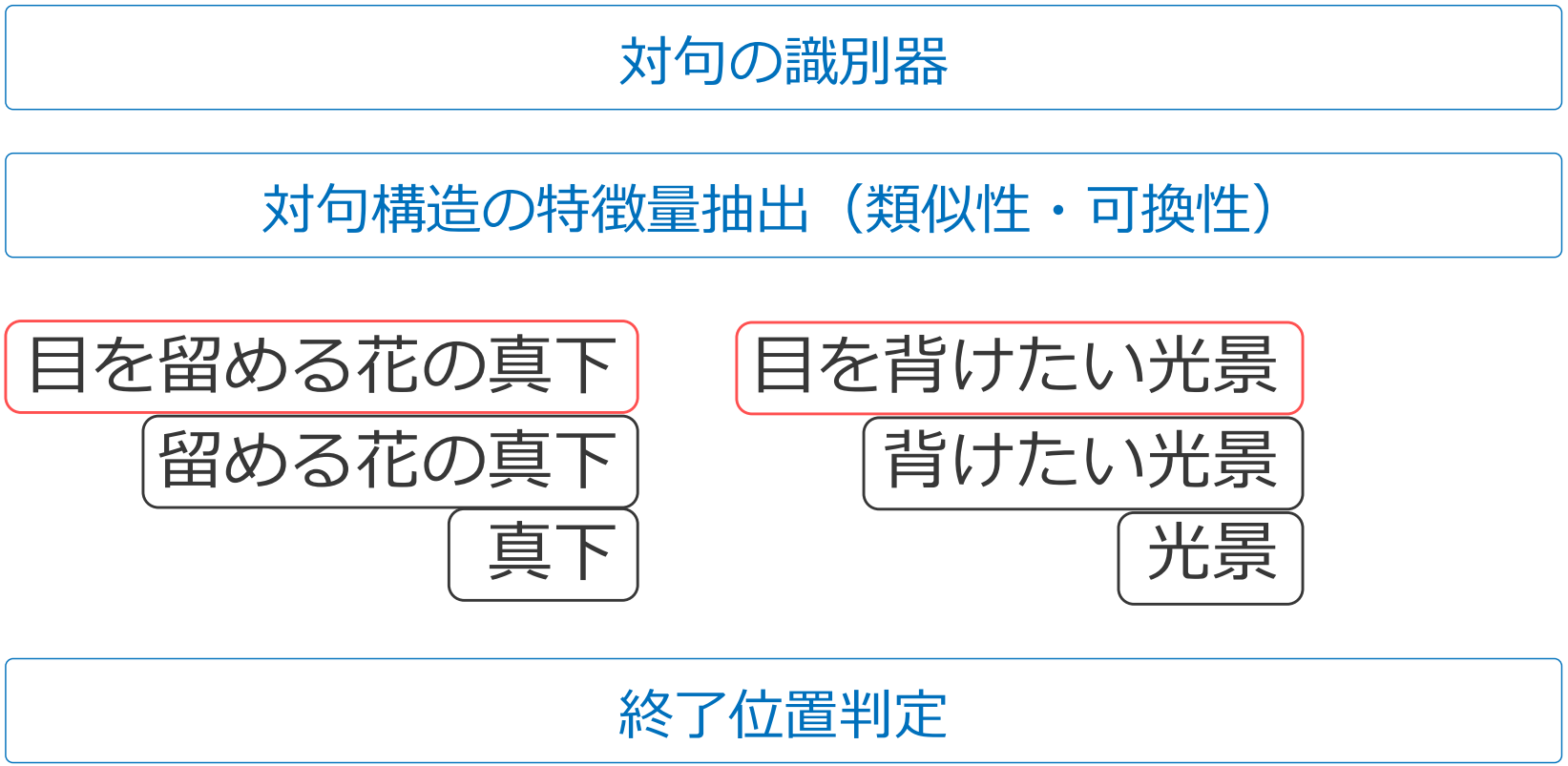
2. 対句構造になるように単語列を生成（単語穴埋め）

   必要な時の終わりが、無駄な時の始まりです。

Ayana Niwa, Naoaki Okazaki, Kohei Wakimoto, Keisuke Nishiguchi, and Masataka Mouri. Construction of a Corpus of Rhetorical Devices in Slogans and Structural Analysis of Antitheses. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(6), November 2021.

対句構造のアノテーション・解析

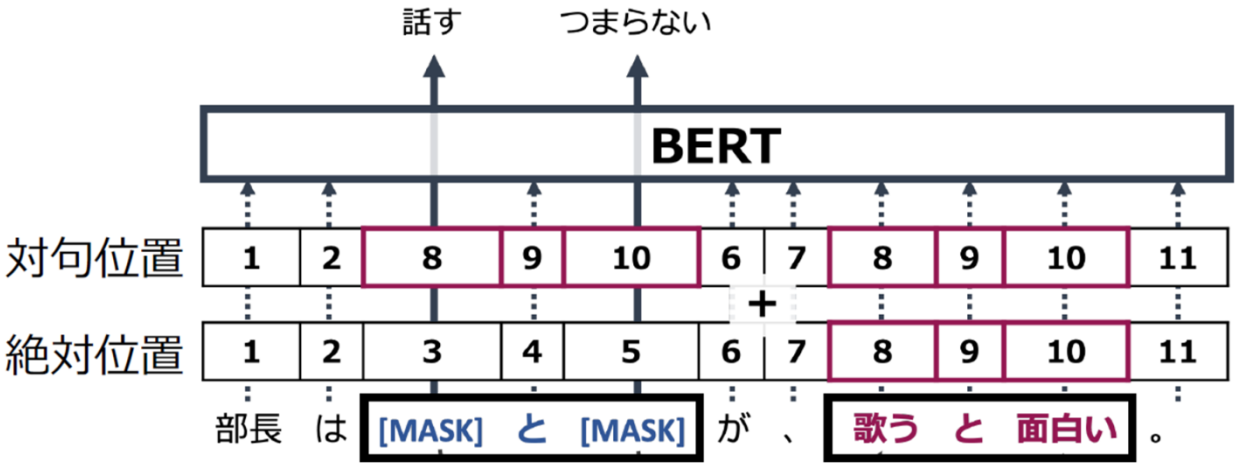
- 10,108件の対句からなるデータセットを構築
- このデータセットで対句の識別器を学習（F1スコアは0.865）



Ayana Niwa, Naoaki Okazaki, Kohei Wakimoto, Keisuke Nishiguchi, and Masataka Mouri. Construction of a Corpus of Rhetorical Devices in Slogans and Structural Analysis of Antitheses. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(6), November 2021.

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	**30.4**	**49.1**
作業者	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.