NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所
NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

ウェブアーカイブを用いた学術研究
ー『国語研日本語ウェブコーパス』の開発とその利用

国立国語研究所
コーパス開発センター
浅原正幸

自己紹介

- 2004年　奈良先端科学技術大学院大学　助教

（2007年7月から 2011年11月まで）
国立国会図書館関西館電子図書館課非常勤調査員
ウェブアーカイブ関連

- 2012年　国立国語研究所　特任准教授
- 2014年　国立国語研究所　准教授

（2011年4月から2016年3月まで）
ウェブコーパスの開発
『国語研日本語ウェブコーパス』(NWJC)と検索ツール『梵天』

- 2019年　国立国語研究所　教授

『NWJC2vec』単語ベクトル／『NWJC-BERT』深層学習モデル

国語研コーパス開発センター
「超大規模コーパス」プロジェクト（2011-2016）

『国語研日本語ウェブコーパス』と検索ツール『梵天』

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所
NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

国語研コーパス開発センター
「超大規模コーパス」プロジェクト：概要

言語研究に資する 100億語規模の Web コーパスを構築する

- 現在の日本語 Web テキストの総量を数十兆語と想定し、適切にサンプリングされた分布で収集されることが望ましい
- Web 上の多様なテキストについて、文体分析を行い、言語の運用実態を明らかにする
- Web 特有の表現に適応した言語解析モデルを構築する
- 計算機の利用を不得手とする方が利用できる検索環境を構築する

『国語研日本語ウェブコーパス』として
検索ツール『梵天』により 2017年3月一般公開
https://bonten.ninjal.ac.jp/

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所
NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

『国語研日本語ウェブコーパス』利用

検索系『梵天』　https://bonten.ninjal.ac.jp/

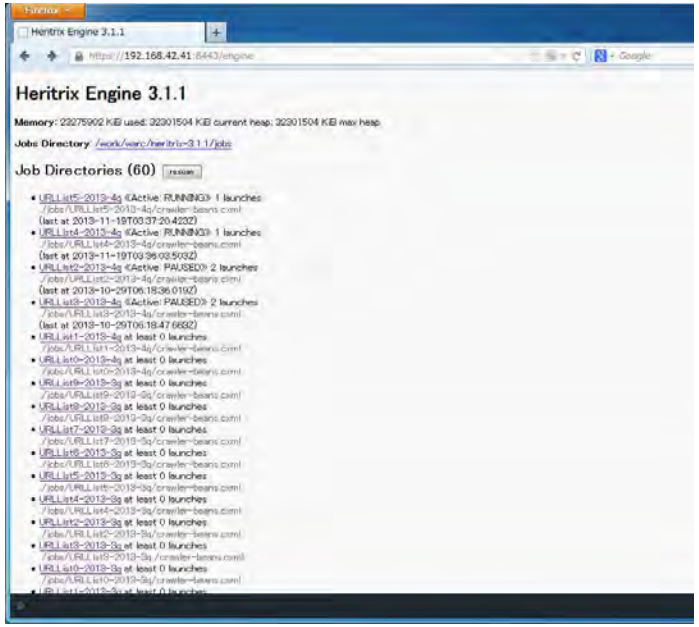
- 文字列検索
- 品詞列検索
- 係り受け部分木検索

2014-4Q データ		
収集WARCファイル数	916	
収集URL数	83,992,556	8399万URL
文数(のべ数)	3,885,889,575	38億文
文数(異なり数)	1,463,142,939	14億文
国語研短単位数	25,836,947,421	258億単位

※残念ながら 2021年9月に公開停止予定

『国語研日本語ウェブコーパス』基盤技術:収集

- HERITRIX クローラ
  - IIPC (International Internet Preservation Consortium) で各国国立図書館がWeb アーカイブを構築するために利用しているオープンソースクローラ
  - WARC 形式と呼ばれる Web アーカイブを保存するファイル形式で保存



- あらかじめ URL のリストを準備
  - 1億 URLを3か月でクロール
  - 同一URLを 1年間で 4回クロール
  - 1年ごとにクロールすべき URL を更新

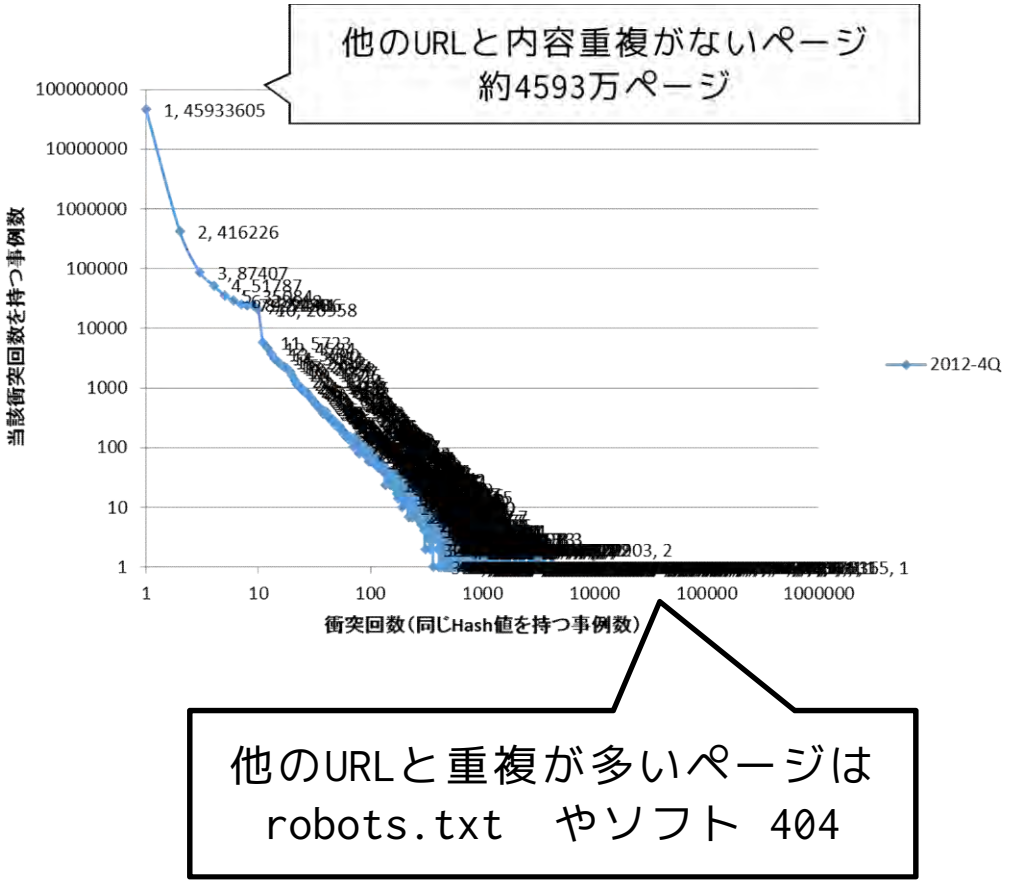
NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所 NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

『国語研日本語ウェブコーパス』
収集:収集ページの統計

	2012-4Q	2013-1Q	2013-2Q	2013-3Q
ページ数　（1期）	61,668,805	58,844,092	61,479,268	57,892,917
内容の重複なしページ数	45,933,605	42,932,982	45,111,527	42,192,931
4期通しての統計				
総異なり URL数（4期）	64,539,233			
(内) 内容の更新なしページ数	27,604,915			
(内) 内容の更新ありページ数	36,934,706			

『国語研日本語ウェブコーパス』収集:収集ページの重複

- 収集ページの2種類の重複
- 異なる URL で同じ内容のもの
  - 内容のコピー&ペースト
  - ミラーサイト
  - robots.txt
  - ソフト 404
- 同一 URL で異なる収集時期で同じ内容のもの
  - 定期的に内容が変更されないサイト
  - 1年を通して同じ内容のサイト
- 1年ごとの収集 URL の選別時に重複が検出されたサイトを排除



『国語研日本語ウェブコーパス』解析

HTMLからテキストコーパスの作成

- 正規化
  - HTML からのテキスト抽出、文字コード変換、文抽出
- 形態素解析・係り受け解析
  - 単語分かち書き
  - 構文情報付与
- n-gram 作成
  - 単語の n個組の頻度情報

『国語研日本語ウェブコーパス』解析：正規化

- 日本語 Google N-gram と同じ手法を利用
  - Unicode NFKC による文字の正規化
  - 文境界
    - ".", "!", "?", "．", "。", "！", "？"
  - 文抽出（以下のものを排除）
    - 5 文字以下もしくは 1024 文字以上（空白を除く）
    - ひらがなが文全体の 5% 未満（空白を除く）
    - 日本語の文字が文全体の 70% 未満（空白を除く）
      - 日本語の文字として扱うコードポイント
      - U+3040 - U+30FF（ひらがな，カタカナ）
      - U+31F0 - U+31FF（カタカナ拡張）
      - U+3400 - U+34BF（CJK 統合漢字拡張 A の一部）
      - U+4E00 - U+9FFF（CJK 統合漢字）
      - U+F900 - U+FAFF（CJK 互換漢字）
- nwc-toolkit https://code.google.com/archive/p/nwc-toolkit/ を利用

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所 NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

『国語研日本語ウェブコーパス』解析：形態素解析・係り受け解析

- 形態素解析・係り受け解析
  - MeCab-0.996 + IPADIC-2.7.0 + CaboCha-0.69
  - MeCab-0.996 + unidic-mecab-2.1.2 + CaboCha-0.69(UniDic主辞規則)

	2012-4Q	2013-1Q	2013-2Q	2013-3Q
収集 WARC ファイル数	814	870	910	905
URL数	61,668,805	58,844,092	61,479,268	57,892,917
形態素数(IPADIC)（文抽出なし）	64,714,650,129 647億形態素	62,077,520,745 620億形態素	63,414,252,638 634億形態素	65,736,027,334 647億形態素
形態素数(IPADIC)（文抽出あり）	33,767,409,441 337億形態素	32,651,138,004 326億形態素	33,073,991,355 330億形態素	30,923,912,566 309億形態素
文数（のべ数）	2,678,315,774	2,600,122,908	2,659,617,620	2,478,309,312
文数（異なり数）	1,097,011,506	1,048,772,913	1,063,649,324	1,007,771,383

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所 NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

『国語研日本語ウェブコーパス』
解析：n-gram データ（1-gram～4-gram）

	Rank	1-gram	2-gram	3-gram	4-gram
本言語資源	1	の	し て	て い ます	し て い ます
MeCab/IPADIC	2	に	まし た	て い た	て い まし た
	3	て	て い	し て い	さ れ て いる
2012-4Q	4	が	て いる	し て いる	し て い た
文単位での	5	は	し た	と 思い ます	さ れ て い
重複性排除有	6	を	で は	さ れ て	た の です が
	7	た	に は	に なっ て	て きまし た
	8	で	さ れ	の です が	れ て い ます
	9	と	ませ ん	し まし た	は あり ませ ん
	10	し	い ます	さ れ た	に なり まし た
本言語資源	1	の	まし た	記事 へ の	記事 へ の トラック
MeCab/IPADIC	2	に	でしょう	お願い し ます	専用 ページ を 表示
	3	を	行っ て	Q & A	利用 する こと が
2012-4Q	4	は	思っ て	続き を 読む	機能 を 利用 する
文単位での	5	て	情報 を	マーク へ 投稿	おすすめ の 知恵 ノート
重複性排除無	6	が	利用 規約	専用 ページ を	正確 性 の 保証
	7	た	おすすめ の	機能 を 利用	お客様 自身 の 責任
	8	で	記事 へ	済み の 質問	回答 を 指示 する
	9	と	追加 する	おすすめ の 知恵	便利 に 新規 取得
	10	し	場合 は	エンターテインメント と 趣味	はてな ブック マーク へ
Web日本語Nグラム	1	の	し て	て い ます	し て い ます
MeCab/IPADIC	2	に	まし た	し て い	さ れ て いる
	3	を	て い	て い た	さ れ て い
[工藤・賀沢 2007]	4	は	て いる	し て いる	は あり ませ ん
	5	て	し た	さ れ て	れ て い ます
	6	が	ませ ん	に なっ て	て い まし た
	7	た	さ れ	し まし た	に なり まし た
	8	で	に は	さ れ た	し て おり ます
	9	と	で は	れ て いる	て きまし た
	10	し	い ます	あり ま せん	し て い た

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所
NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

『国語研日本語ウェブコーパス』
解析：n-gram データ (1-gram～4-gram)

本言語資源 MeCab/IPADIC 2012-4Q 文単位での 重複性排除有	1	されています	ではないでしょうか	のではないでしょうか
	2	ではありません	ていたのですが	のタグが付けられた質問
	3	と思っています	のではないでしょう	ではないかと思います
	4	していました	のではないかと	に関するウェブ上の情報を探す
	5	ではないでしょう	に行ってきました	ああああああああああああ
	6	のではないか	ような気がします	のではないかと思い
	7	はないでしょうか	タグが付けられた質問	していたのですが
	8	になっています	のタグが付けられた	思っていたのですが
	9	ていましたが	させていただきました	えええええええ
	10	ていたのです	たいと思っています	と思っていたのです
本言語資源 MeCab/IPADIC 2012-4Q 文単位での 重複性排除無	1	記事へのトラックバック	機能を利用することが	機能を利用することができ
	2	機能を利用すること	利用することができませ	利用することができません
	3	利用することができ	正確性を保証して	正確性を保証しており
	4	正確性を保証し	お客様自身の責任と判断	お客様自身の責任と判断で
	5	お客様自身の責任と	すべての機能を利用する	すべての機能を利用すること
	6	はてなブックマークへ投稿	知恵袋のすべての機能を	知恵袋のすべての機能を利用
	7	更新情報が届きます	おすすめの解決済みの質問	ニックネームのMy知恵袋で確認でき
	8	おすすめの解決済みの	記事へのトラックバックURL	質問年月や画像の有無を
	9	すべての機能を利用	ニックネームのMy知恵袋で確認	質問や知恵ノートは選択さ
	10	質問年月や画像の	することができません	以上更新がないブログに表示
Web日本語Nグラム MeCab/IPADIC [工藤・賀沢 2007]	1	されています	無料でお届けします	料無料でお届けします
	2	ではありません	料無料でお届けし	配送料無料でお届けし
	3	でお届けします	配送料無料でお届け	国内配送料無料でお届け
	4	無料でお届けし	国内配送料無料でお	以上国内配送料無料でお
	5	1500円以上国内配送	円以上国内配送料無料	円以上国内配送料無料で
	6	料無料でお届け	以上国内配送料無料で	1500円以上国内配送料無料
	7	配送料無料でお	1500円以上国内配送料	はインラインフレームを使用して
	8	国内配送料無料で	を使用しています	フレームを使用しています
	9	以上国内配送料無料	インラインフレームを使用して	インラインフレームを使用してい
	10	円以上国内配送料	この記事へのトラックバック	部分はインラインフレームを使用し

国語研コーパス開発センター
共同研究プロジェクト(2017-2021)

自然言語処理
深層学習モデルの構築

単語埋め込み

単語をベクトル表現に変換する技術
（可変長離散記号列から固定長連続空間への写像）
→分布意味論

- Word2vec: Google [Mikolov+ 2013]
- GloVe/gensim: Stanford U. [Pennington+ 2014]
- FastText: Facebook [Bojanowski+ 2016]

“king” − “man” + “woman” = “queen”
「王」 - 「男性」 + 「女性」 = 「女王」

文脈化単語埋め込み

ELMo (Deep Contextualized Word Representations): Allen AI
[Peters+ 2018] https://arxiv.org/pdf/1802.05365.pdf

従来の単語埋め込み：単語の出現ごとに同じベクトル
文脈化単語埋め込み：単語の出現ごとに異なるベクトル
→多義語について異なるベクトルを割り当てられる

従来の単語埋め込み:	文脈化単語埋め込み:
距離が短い (0,3,0,…) 時間が短い (0,3,0,…)	距離が短い (1,3,0,…) 時間が短い (0,3,1,…)
同じベクトル	違うベクトル

事前学習モデル BERT

- 事前学習モデル
  - 穴埋め問題

```
Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon
```

  - 隣接文推定

```
Sentence A: the man went to the store .
Sentence B: he bought a gallon of milk .
Label: IsNextSentence
```

```
Sentence A: the man went to the store .
Sentence B: penguins are flightless .
Label: NotNextSentence
```

- 転移学習
  - 既存の学習済モデル（出力層以外の部分）を、
    重みデータは変更せずに特徴量抽出機として利用する。
- Fine-tuning
  - 既存の学習済モデル（出力層以外の部分）を、
    重みデータを一部再学習して特徴量抽出機として利用する。

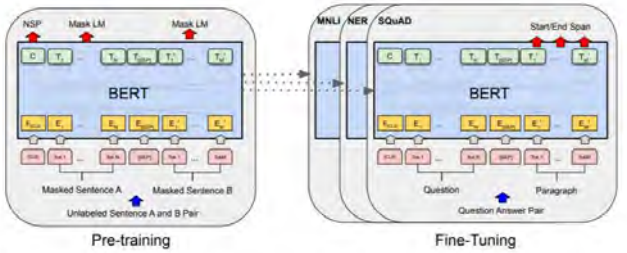


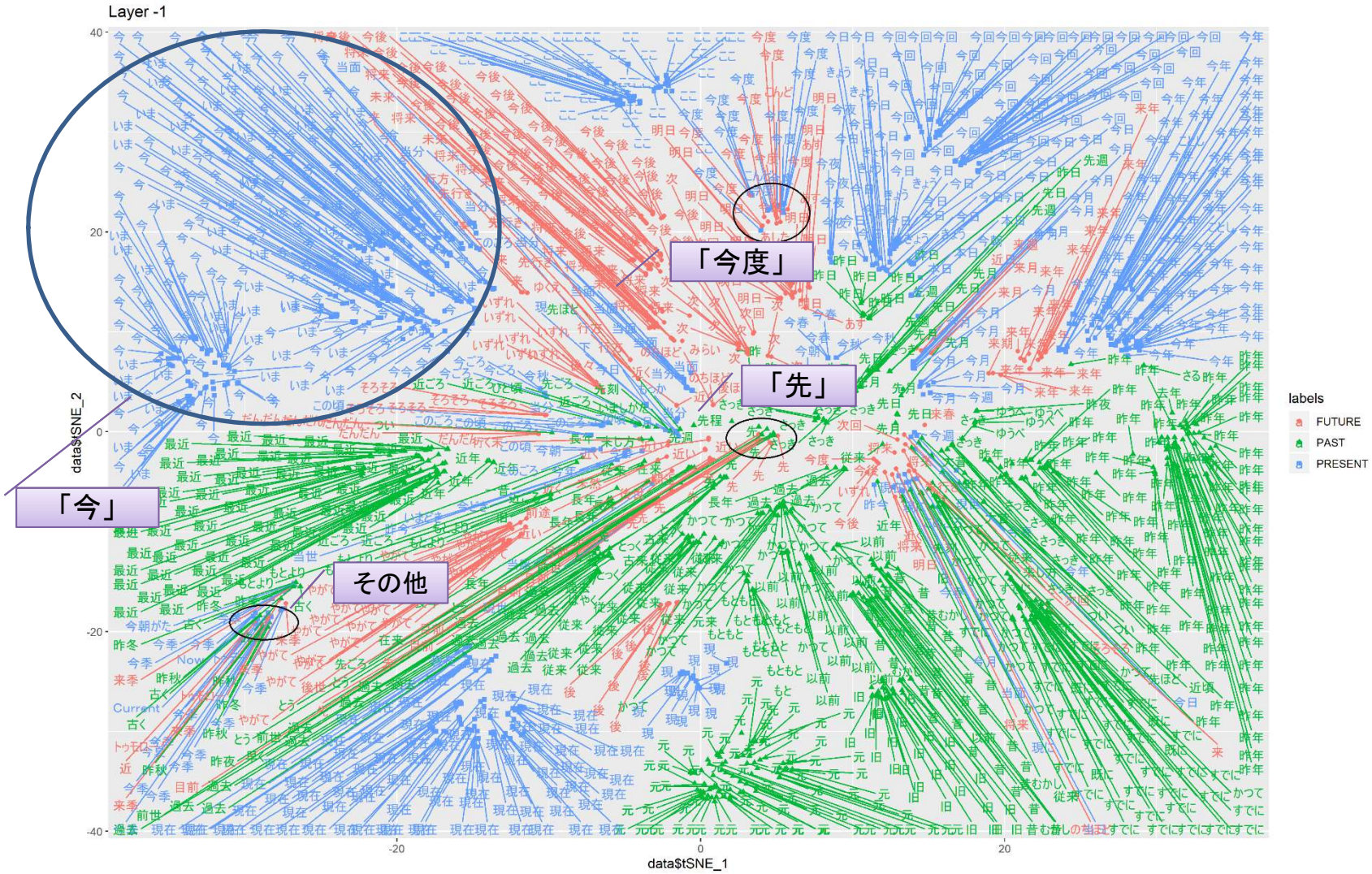
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architec-

NWJC-BERT
『国語研日本語ウェブコーパス』によるモデル

- 『国語研日本語ウェブコーパス』1,287,504,831文による
  訓練済み BERT モデル

  各種、自然言語処理の応用に利用できるほか、
  文脈化単語埋め込みが得られる

  次のスライドで時間（現在、過去、未来）を表す表現 のベクトルを可視化


Layer -1
「今」
「今度」
「先」
その他
data$tSNE_2
data$tSNE_1
labels
FUTURE
PAST
PRESENT

NWJC-BERT による文脈化単語埋め込み
さまざまな「今」

- 「今でも／今では」

なかった。彼女がおれを訴えるとさえ言わなかったならば、殺さずにすんだ。あれは無駄な殺しだった。	【 いま 】		でも後悔している。 しかし、清瀬組長は許せなかった。おれの釈明に一切耳を傾けず、詰めた指
で「もうかりまっか？」と聞けば「ボチボチでんなあ（まあまあです）」が挨拶のように交わされていた。	【 今 】		でもビジネスマンがこの言葉を使っているところを見ると、浸透度合いがうかがえる。そして「ボチボチ」はうまくいっている合図
、そう信じていた。私は従容と死に就いた彼を見て、その言葉に偽りはないと知った。	【 今 】		でも彼には感謝している。 人の一生は、小さな輪の連鎖だ。ひとつひとつの輪の形
年間に何度も警察に逮捕されているのだが、彼女はいつもケロッとしています」（同前）	【 今 】		では、とても七十四歳とは思えない逞しさを発揮しているというのだ。作家の藤本義一さんもこう
正弘さんが腕を伸ばしてきた時、一度、「痛いから」と言ったが、わかってもらえなかった。	【 今 】		では大学生の息子がサークル活動などで外泊する日に「またかな」と思うと憂うつになる。そんな日は

- 「○○は今、」「○○も今、」

。内外で高い評価を受けている現代美術家、宮島達男さんの数字は何を語りかけているのだろう。その発想は	【 いま 】		、終末期医療の場でも注目されているという。（聞き手　早瀬廣美）住民共作　間もなく大阪府
占められた時代であり、若年の労働者はその中心部分を担っていました。この労働市場のピラミッド構造は、	【 今 】		、急速に解体しつつあります。その変化は、グローバリゼーションによって海外に工場が移管されることによって加速度
の男性像は力士・レスラー・ボクサーですから。篠崎　あ、それは私も同じ。私の究極の理想像も「	【 今 】		、原始時代にワープしても知恵と力を使って自分で獲物をとってこれる人」だもの。宮崎　…（
無党派だけでは不安。組織票（連合）は捨てられない」と、同党からの出馬を決めた。柳沢氏は	【 今 】		、知事選の時のような、無党派層のうねりを感じていない。 小泉純一郎首相とのツーショット写真を

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所
NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

NWJC-BERT による文脈化単語埋め込み
さまざまな「今」

- 「今から」

働いていて、この町のことは何でも知っている男です。この時間ならまだオフィスにいるはずだから、	【	今	】	から行ってみましょう」 河合が席から立った。思い立つとジッとしていられない性分らしい。「ついて来
組のカップルがいた。日本人の血が入っているという若いコックが直接やってきて、たどたどしい日本語で	【	今	】	から出来るメニューの説明をした。おそらくひびきの英語に肝を潰したさっきの男が差し向けたのだろう。出てき
よると、運転士は今年八月二十五日午後四時ごろ、酒に酔った状態で前上司に電話して「	【	今	】	から行くから待っていろ」などと暴言を吐いたうえ、前上司の職場の市営地下鉄貝塚駅事務所に出向き、

- （いまとなっては、いまのとこ、いまの時代、いまのまま）

を賭して抵抗したことに、彼女の高木に対する愛の証しを見たような気がする。 だが、	【	いま	】	となっては、その愛の証しが恨めしい。要するに、狂犬に噛まれたので、香保には責任はない。
挟んだ。「なかなかの男前じゃないか」 マックは唇をとがらせた。「ほかに怪しい人間は？」「	【	いま	】	のとこ、こいつだけだ」リッジオが答えた。「あとは、単独でエレベーターに乗ってきた女が二人。十代
大勢のおともだちにイエス様のことを伝える手段があるなら、ポケモンでなくてもよいと思っています。しかし、	【	今	】	の段階では、このカードを用いることが、おともだちを集めるにはよいと考えています。彼らがイエス様
中、約千四百万人の有権者がどれだけ投票所に足を運ぶのかが焦点となっている。「	【	今	】	のような状況では、全体の投票率が五十％を超えれば選挙は成功だ」。イスラム教シーア派有力政党「
、当然の帰結である。 商品に魅力があれば、購買力はおのずとついてまわるものなのだ。 加えて	【	いま	】	の時代、瞬時に売ってしまう、スピーディさを研磨していくことが不可欠だ。須田も、「中古・再生住宅の
以外の電話会社、政府などに対するご要望、ご意見があれば、何でもお聞かせください。多くの人は	【	いま	】	の速度に不満 Q1を見てみると、最も多い通信環境がISDNの六十四kbps（三十七％）、続い
ていればいい。 つまり、現行の皇室典範を一部改正すれば、女帝を回避できるというのである。「	【	いま	】	のままの皇室典範が維持される限り、雅子妃か秋篠宮妃が男子を生まないと皇統が絶えてしまう。 雅子
あったのです。 今夜、ほうせきをもらいに行く。いくら用心しても、だめだよ。二十めんそう さては、	【	今	】	の子どもは、二十めんそうの手下だったのかと、たけしくんは、門の外をにらみつけました。「おにい
のか。それを知るためには、まず子供たちの心の世界としっかりと向き合うことが必要だろう。そうすれば、	【	今	】	の子供たちが自分たちよりもずっと進化した存在であることを理解し、彼らを大人社会による抑圧から解き放っ

NWJC-BERT による文脈化単語埋め込み
さまざまな「今」

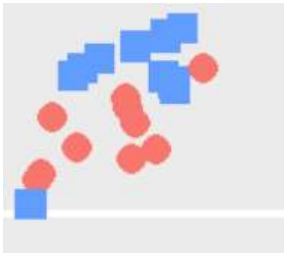
- 「今協議」

、“決裂”も辞さないという警告の意味合いも含まれている。 ケリー次官補の発言は短いながらも、	【	今	】	協議に臨む米国の基本姿勢がすべて盛り込まれ、北朝鮮に対する核廃棄要求だけでなく、米国が与える代償、将来の関係
については北朝鮮が存在を認めず―と基本的な部分では双方の主張の隔たりはなお大きい。 米国は	【	今	】	協議で、完全妥結には至らないまでも一定の進展があることを強く期待しており、それが実現しない場合
戦略を描いており、ケリー発言には当然こうした意図を北朝鮮に暗に伝える狙いも込められている。 米国は	【	今	】	協議を通じ、北朝鮮に対し各国とともにあらゆる手段でその態度変更を促す構えだ。 日本が拉致問題を提起する
公正取引委員会の山田昭雄事務総長は十二日の定例記者会見で、官製談合防止法案について、「	【	今	】	国会中に成立していただきたい」と述べ期待感を明らかにした。与党3党は、発注者である公務員

- 「今ひとつ」

。形状デザインの「業界標準」となっている。3G周波数を探せ さて、このようにモバイルデータ通信市場が	【	いま	】	ひとつブレイクしないにもかかわらず、二千年頃から「次世代3Gサービス向けの周波数を割り当ててくれ」と
通常、自然住宅とか健康住宅を唱えている会社は、コストが高すぎるか、デザインセンスがないことにより、	【	今	】	ひとつ踏み込めないのですが、総合的に超えていた点には、本当に参ってしまったのです。 ただ

NWJC-BERT による文脈化単語埋め込み「今度」



- 現在　：PM25_00084

ゴミ箱に捨てたりしている。しかしどうもまだ気がのらないようで，しばらく漫然と机を眺めていたが，【今度】は机のいちばん下の引き出しを開けた。中から何枚か，ＢｅＯＳのＣＤ－ＲＯＭが出てくる。

- 未来：PN4g_00006

が、トータル百四十四（七十一、七十三）のスコアで初優勝。この春から沖学園高に進学し、「【今度】は高校で日本一になりたい」と大きな目標を掲げている。ゴルフを始めたのは佐賀県武雄市立朝日

NWJC-BERT による文脈化単語埋め込み

「先」

1.
ものの先端の方。
2.
物・作用等が向かう所。

- 過去 ：PN1e_00002

医療側の痛み不足」　政府の経済財政諮問会議（議長・小泉純一郎首相）が九日開かれ、厚生労働省が【先】に公表した医療制度改革試案に対し、本間正明・阪大教授ら四人の同会議民間議員が意見書を提出し

- 未来：PM25_00084

原稿を入れ始めるのが@日からなんですよね。わりと急な企画なもので。メーカーにマシンを返すのは【先】なんで後処理のことは悩まなくていいんですが。…アップデートの最中に止まっちゃうって感じですか。ええ，

NWJC-BERT による文脈化単語埋め込み その他

- 現在 PM26_00004

安全かつ低侵襲で留置可能な尿道ステント留置術を必要とする機会が増えると予想されます。Ｔｈｅ　Ｃｌｉｎｉｃａｌ　【Ｃｕｒｒｅｎｔ】加齢黄斑変性症の新しい治療法---光線力学療法尾花明大阪市立大学大学院医学研究科視覚病態学助教授

- 現在 PN2e_00004

だが「彼女の穴を埋める存在は見当たらない」（ＵＳＡ【トゥデー】紙）のが実情で、ローブ氏の権勢が高まるのは確実だ。　ローブ氏の最大使命は大統領の再選だ。

- 未来　PM51_00077

裏方のお仕事が多いので、動きやすくてお客様に対して失礼にならない服装を心がけてます。この【トゥモロー】ランドのサテンワンピースは、胸元の切り替えとハイウエストマークで脚長に見えます。この冬は、ファーバッグが大活躍です

事前学習モデル GPT/GPT-2/GPT-3

- 文生成系の事前学習モデル
  - 穴埋め問題ではなく、次の単語予測
    （方式としては BERT の単方向モデル）

NWJC-GPT-2
『国語研日本語ウェブコーパス』1,287,504,831文による
訓練済み GPT-2 モデル

言語解析ではなく言語生成に利用可能

【参考】
朝日新聞社メディアラボ (2020/11/18)
https://cl.asahi.com/api_data/gpt2-demo.html
LINE / NAVER (2020/11/25)
https://linecorp.com/ja/pr/news/ja/2020/3508

NWJC-GPT-2が生成するテキスト

通常の訓練結果：

前 から いい 店 だっ た ん だろう

前 アップ さ せ て ください

削除 要請 の メール は 非常 に 承っ て おり ます

削除 要請 メール は 領収 書 を 当せん サイト 社内 に 配布 いたし て おり ます

ウェブデータ＋短歌での訓練結果：

生くる 事 われ の 有能 不 条 の うち に 羨し 乳 射す 畝 越え に けり

あした 天 は かがやく 手のひら を 頬 に 洗い やま ず 頭 なか に 血まひて 来 た

木の下 に 鋭く 美しく 束 突き 激しき 足取り の 葉先 つつ も

切ら れ ざり し て 切ら れ し 者 あり て その 切 里 を 潜め む と す

おのづから 花 開く よ は 低き うつろ さ あり し か 熟れ て ゆく 前日 の 朝

NWJC-GPT-2 の問題点

- NWJC は収集時に収集対象の統制を行っていないため、
  元テキストに性的な表現・暴力的な表現が含まれている可能性がある
  結果、どんな表現が出力されるかわからない
- 解析のためのモデル(NWJC-BERT) は元テキストの属性の影響は少ないが、
  生成のためのモデル (NWJC-GPT-2) は元テキストの属性の影響が大きい

国立国会図書館
インターネット資料収集保存事業について

NDL WARP の活用について

- 深層学習のための質のよい大規模テキストデータ
  WARP（国立国会図書館インターネット資料収集保存事業）は
  収集時に収集対象の統制されているため
  言語処理（さらには法律・政治関連言語処理）への活用が見込まれる

- 言語の変遷の調査の可能性
  言語研究において「国会会議録検索システム」はよく利用される
  言語の性差・世代差・通時的変遷・方言の調査
  アーカイブしつづけることで未来の言語研究者の重要なデータとなる

まとめ

- 『国語研日本語ウェブコーパス』
  - 検索ツール『梵天』
  - 構築方法
- 『国語研日本語ウェブコーパス』を用いた深層学習モデル
  - 『NWJC2vec』　　単語埋め込み
  - 『NWJC-BERT』　　事前学習モデル（文脈化単語埋め込み）
  - 『NWJC-GPT-2』　　文生成モデル

- 国立国会図書館
  インターネット資料収集保存事業について