国語分科会
言語資源小委員会　資料

大規模言語モデルとテキストコーパス

２０２４．９．２６

相澤彰子
（国立情報学研究所）

本日の内容

- 大規模言語モデルにおける文脈と語の意味
- 大規模言語モデルの意味理解
  - 語義あいまい性
  - 意味の最小単位の扱い
- 大規模言語モデルの訓練とテキストコーパス
- コーパスに由来する大規模言語モデルのリスク
  - バイアス
  - テキスト復元
- テキストコーパスと大規模言語モデル

大規模言語モデルにおける文脈と語の意味

「分布仮説」と「文脈類似度」

- Words that are used and occur in the same contexts tend to purport similar meanings
  - Harris, Z. (1954). "Distributional structure". Word. 10 (23): 146-162.
- A word is characterized by the company it keeps
  - Firth, J.R. (1957). "A synopsis of linguistic theory 1930-1955". Studies in Linguistic Analysis. Oxford: Philological Society: 1-32. Reprinted in F.R. Palmer, ed. (1968). Selected Papers of J.R. Firth 1952-1959. London: Longman.

文脈と意味

1. 私は毎日たくさんの野菜を食べるようにしています。
2. このレストランでは地元産の野菜を使った料理が人気です。
3. 昨日の市場で新鮮な野菜をたくさん買いました。
4. 野菜は健康に良いので、サラダにたっぷり入れています。
5. 彼は野菜をあまり食べないので、栄養バランスが心配です。
6. 庭で野菜を育てるのは楽しいですね。
7. 今晩の夕食には五種類の野菜を使ったスープを作る予定です。
8. 彼女は野菜不足を解消するために、スムージーに野菜をたくさん入れています。
9. 私のお母さんはいつも野菜をたくさん使った料理を作ってくれます。
10. このスーパーの野菜はいつも新鮮で、品揃えが豊富です。
11. 野菜を食べると、体が軽く感じます。
12. 子供たちに野菜の大切さを教えるために、一緒に野菜を植えています。
13. 野菜中心の食生活を始めてから、健康が良くなったと感じます。
14. 野菜はビタミンやミネラルが豊富なので、毎日の食事に欠かせません。
15. 彼は野菜を炒めるのが得意で、いつも美味しいです。
16. 週末は地元の農家から直接買う新鮮な野菜を楽しみにしています。
17. 野菜の色々な調理方法を学ぶことで、食事がもっと楽しくなります。
18. 野菜をたくさん含む食事は、ダイエットにも効果的です。
19. 野菜売り場は、色とりどりの野菜でいっぱいです。
20. 今日は野菜を使った新しいレシピに挑戦してみます。

Examples generated by ChatGPT

文脈の近さによって意味の近さを測る

↓

「文脈語」のベクトルを作成する

語の意味を計算機で捉えるためには？

文脈ベクトル

例. 野菜＝(5300, 4400, 3100, ……)

（コーパス中で何回共起したか？）

使う　食べる　作る　…

例. 果物＝(2300, 2200, 1100, ……)

例. 豚肉＝(2900, 3200, 100, ……)

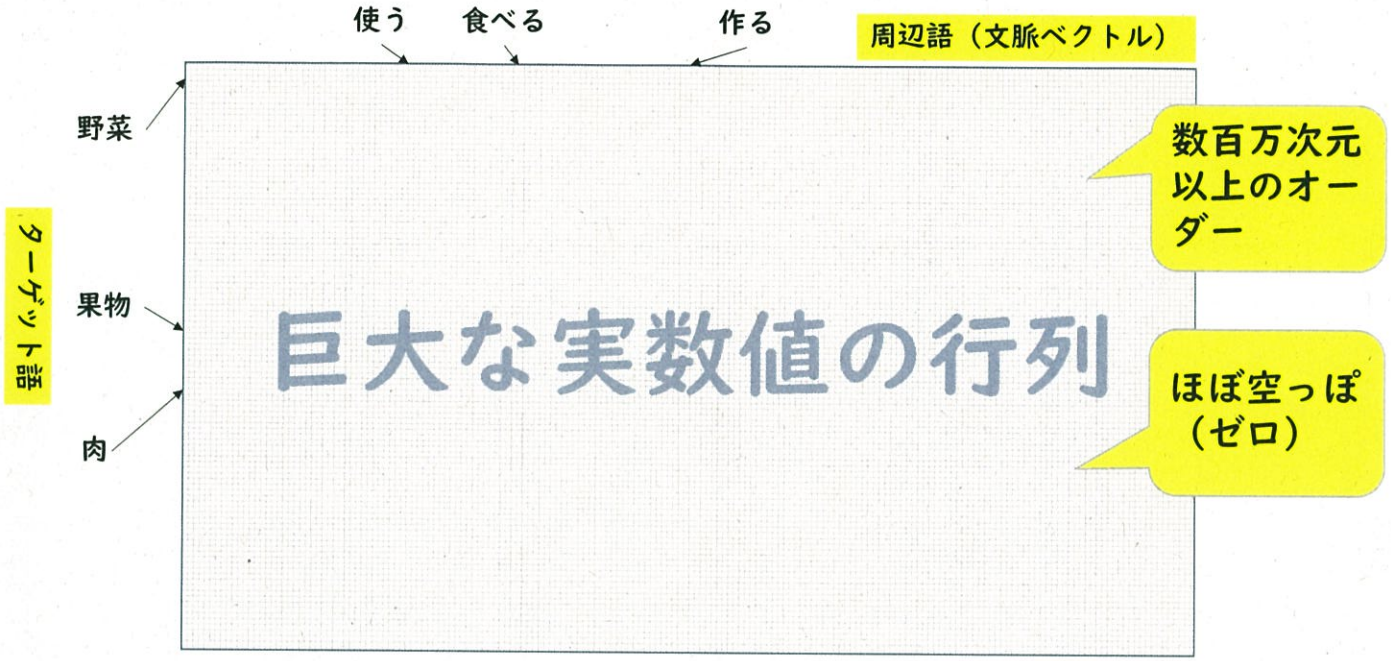
	使う	食べる	作る
野菜	5300	4400	3100
果物	2300	2200	1100
豚肉	2900	3200	100

「意味が近い」
＝「文脈ベクトルが似ている」

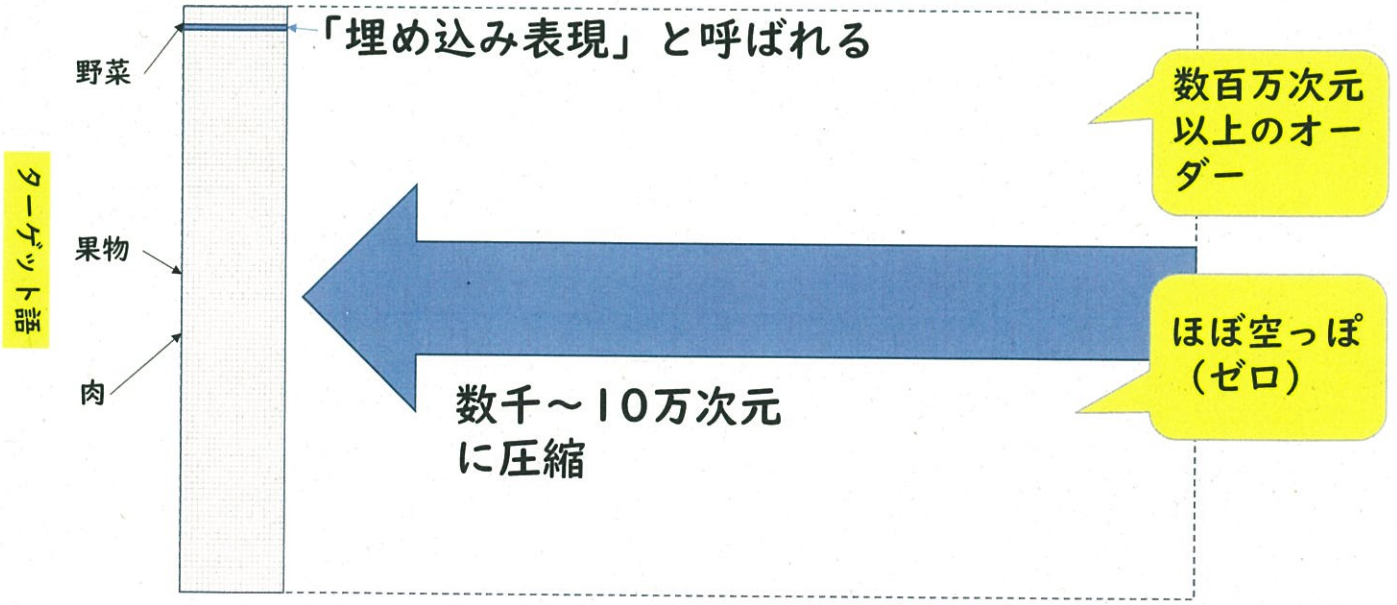
言語資源小委員会資料（NII相澤）

2024/9/26

文脈ベクトルによる意味の表現

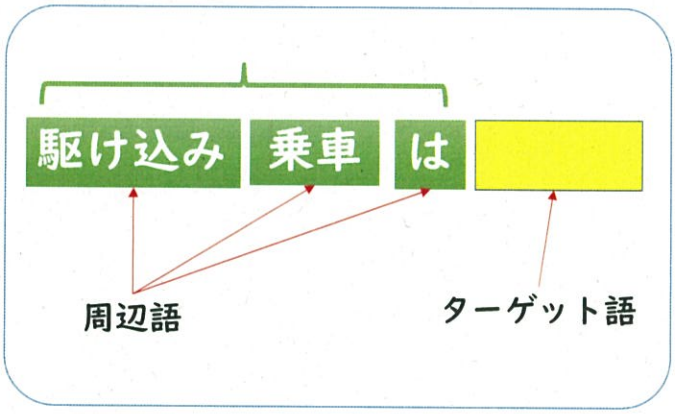


文脈ベクトルによる意味の表現



言語モデルの「訓練」

・素材は大量の「テキスト」
・目標は次の単語（「トークン」）をなるべく正確に予測すること



次に来る単語を予測

入力を与えられると、言語モデルが、それに続く文章を自動的に生成する

大規模言語モデルの意味理解

語義あいまい性と意味の最小単位の扱い

語義のあいまい性

同じ言葉が異なる意味を表すことがある

“mouse”

表層的な表現は同じ

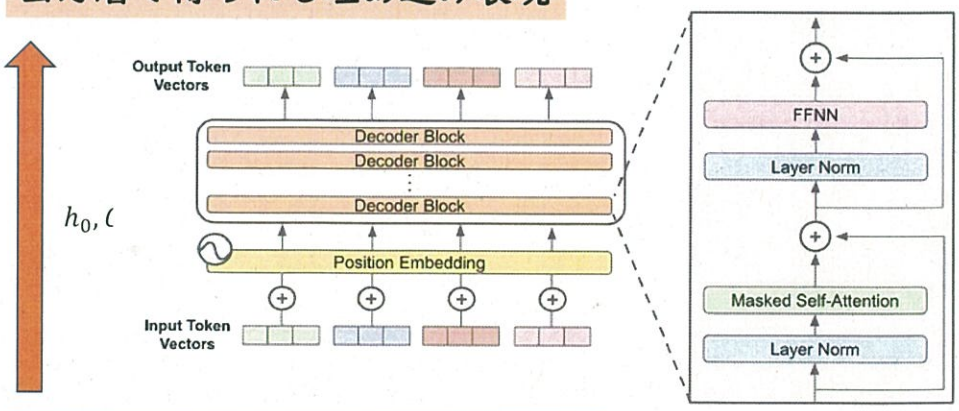
語義が違っても同じ「埋め込み表現」が割り当てられる。

言語モデルの語義あいまい性解消

図は以下から引用
https://cameronrwolfe.substack.com/p/decoder-only-transformers-the-workhorse

文脈情報が反映されている

出力層で得られる埋め込み表現



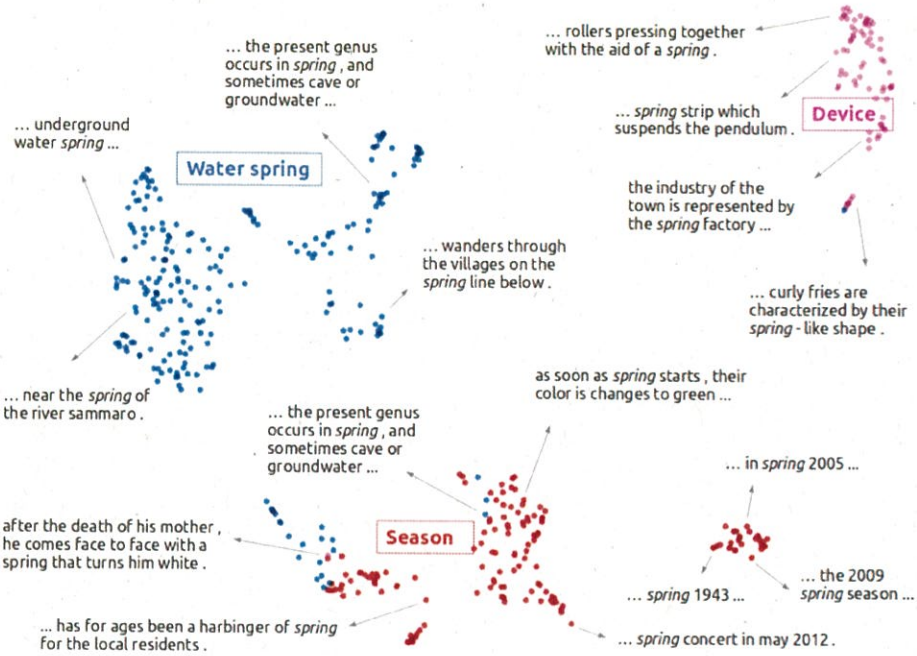
入力層に入力される埋め込み表現

文脈による語義の変化

図は以下から引用

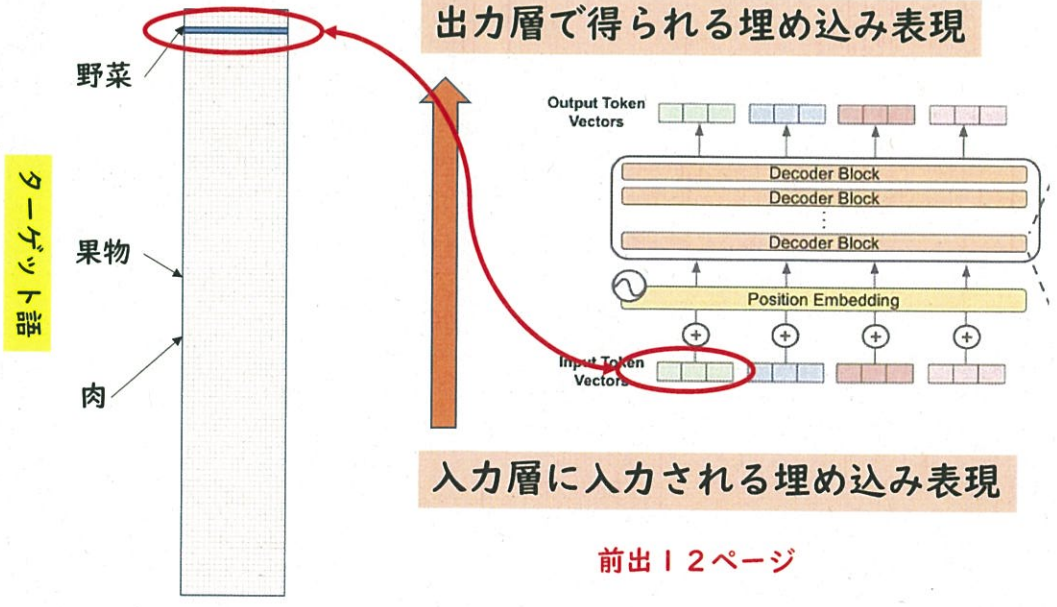
Daniel Loureiro, et al. “Analysis and Evaluation of Language Models for Word Sense Disambiguation.” Computational Linguistics (2021) https://arxiv.org/abs/2008.11608

言語モデルBERTにおける単語“spring”の文脈による変化を可視化



「トークン」

前出８ページ



前出１２ページ

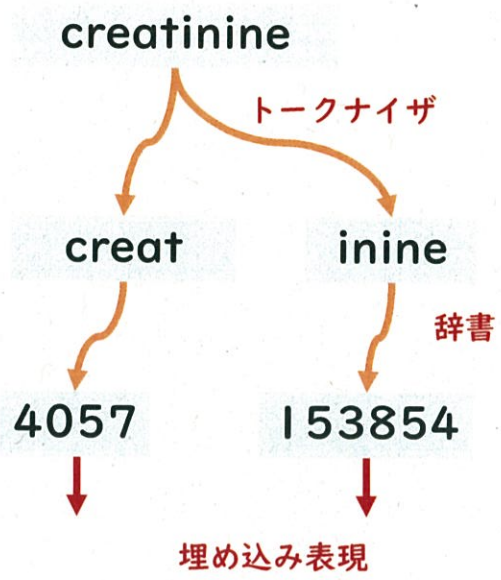
「トークン」
＝言語モデルへの入力の単位

「語彙」
＝トークンの集合

「辞書」
＝トークンと埋め込み表現の対応表

「トークナイザー」
＝入力テキストのトークンへの分割器

トークナイザ



GPT-4o のトークナイズ

https://tiktokenizer.vercel.app/

英語

Her creatinine was mildly elevated, but she did not have any known renal disease, thereby ruling out uraemic encephalopathy.

Token count
26

Her creatinine was mildly elevated, but she did not have any known renal disease, thereby ruling out uraemic encephalopathy.

25614, 4057, 153854, 673, 151020, 49217, 11, 889, 1770, 2242, 625, 679, 1062, 5542, 75879, 11089, 11, 42469, 42842, 842, 190571, 60582, 90410, 79379, 84176, 364

トークナイザー

日本語

クレアチニン値は軽度に上昇したが、既存の腎疾患がないため、尿毒症性脳症の可能性は除外された。

Token count
43

クレアチニン値は軽度に上昇したが、既存の腎疾患がないため、尿毒症性脳症の可能性は除外された。

7787, 12016, 10398, 19285, 20968, 4025, 69555, 5205, 113670, 8913, 5280, 4286, 2181, 229, 23085, 6632, 1395, 92055, 19565, 3385, 16512, 236, 107849, 84443, 6632, 47592, 103912, 1395, 91854, 49299, 76062, 8360, 14598, 111, 76062, 3385, 42043, 8360, 5205, 18593, 10727, 188599, 3414

GPT-4o のトークナイズ

https://tiktokenizer.vercel.app/

英語

Her creatinine was mildly elevated, but she did not have any known renal disease, thereby ruling out uraemic encephalopathy.

Token count
26

Her creatinine was mildly elevated, but she did not have any known renal disease, thereby ruling out uraemic encephalopathy.

25614, 4057, 153854, 673, 151020, 49217, 11, 889, 1770, 2242, 625, 679, 1062, 5542, 75879, 11089, 11, 42469, 42842, 842, 190571, 60582, 90410, 79379, 84176, 364

トークナイザーの性能
テキストをなるべく少ないトークンで表現

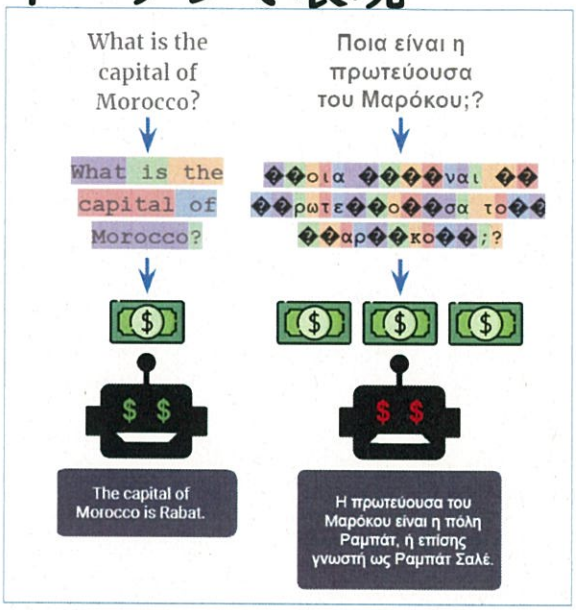
LLMのコストと利便性に直結

OpenAI の APIの課金は入力と出力のトークン数に比例

GPT-4-turboは入力 1Mトークンあたり10ドル、出力1Mトークンあたり30ドル

入出力できるテキストの上限はトークン数で決まる

GPT-4 8Kは 8,192トークン、GPT-4 32Kは32,768トークン



Ahia+ "Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models." http://arxiv.org/abs/2305.13707.

トークナイザーの性能
言語間で不平等が生じている

Ahia+ "Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models." http://arxiv.org/abs/2305.13707.

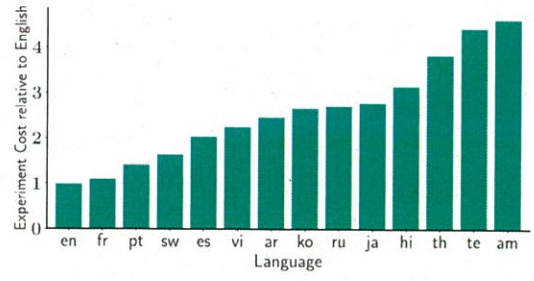


Figure 5: Average cost of prompt + generated tokens for XLSUM evaluations relative to English.

日本語は英語の2.8倍のトークン数が必要

その分、プロンプトに収まるテキスト量が少ない

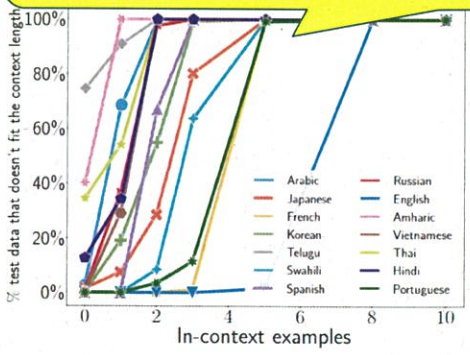
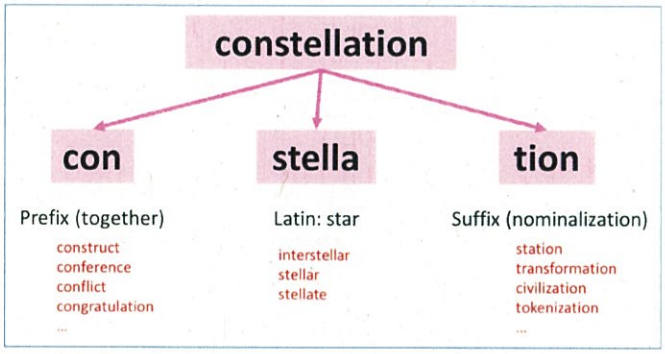


Figure 6: Percentage of test examples per language in XLSUM that do not successfully fit into the context length of ChatGPT. We can fit more few-shot examples in Latin script languages than in other languages.

統計的アプローチの限界

トークン ≠ 形態素

形態素＝意味の最小単位



形態素から語の意味を推測する

トークン＝効率の良い符号

GPT-4o
https://tiktokenizer.vercel.app/

Token count
2

constellation

トークナイザーの課題

- 言語モデルは言語を区別していない（多言語、プログラム言語、…）
- ウェブ文書は、対象言語の単語分布を忠実に再現していない
- 常用漢字など漢字の登録をある程度優先する必要がある
- 数字、記号列、インデントや改行などへの対応も必要

↓

語彙構築の難しさ

トークナイザーの課題

・言語モデルは言語を区別していない（多言語、プログラム言語、…）
・ウェブ文書は、対象言語の単語分布を忠実に再現していない
・常用漢字など漢字の登録をある程度優先する必要がある
・数字、記号列、インデントや改行などへの対応も必要

語彙構築の難しさ

大規模言語モデルの訓練とテキストコーパス

大規模言語モデルの規模

- 予測に有効な手掛かりを巨大な変数の値（重み、パラメタ）であらわす

  たとえば数千億～１兆個

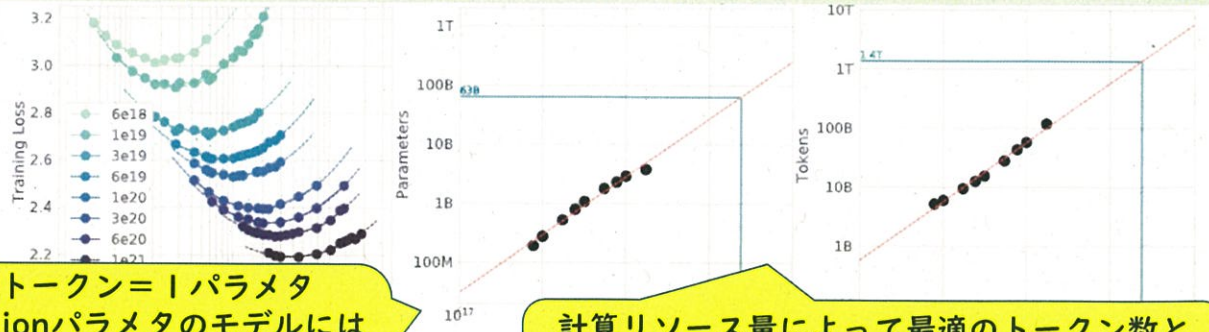
- 膨大な量のテキストについて、なるべく正確に予測できるように変数の値を調整

  たとえば数兆トークンなど

訓練用コーパスの大きさの目安

基本的には経験則：Chinchilla Scaling Lawが有名

計算資源、訓練に使うコーパス、モデルのサイズの関係



20トークン＝１パラメタ
70Billionパラメタのモデルには
1.4Trillionトークンのデータ
（＝Amazon US Kindle Store
のすべての本より多い※）

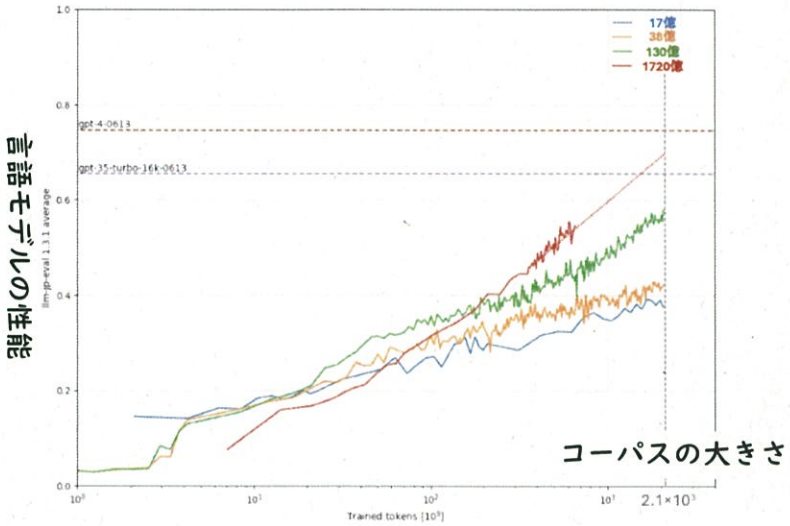
https://lifearchitect.ai/the-sky-is-bigger/

計算リソース量によって最適のトークン数と
パラメタ数が決まる！
大規模なモデルには巨大なコーパスが必要

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, et al. 2022. "Training Compute-Optimal Large Language Models." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2203.15556.

訓練用コーパスの大きさとモデル性能

実際に試してみると、17億～1720億パラメータ数のすべてのモデルで、コーパスを増やすほど性能が上がっている。



https://llmc.nii.ac.jp/topics/llm-jp-172b/

コーパスの質の問題

・訓練に使われるコーパスの大半はWeb由来であるが、「質」も重要であることが知られている

・Web文書から抽出したテキストにはノイズが多いため、コーパスの質を上げるための工夫あれこれ
  ・自然言語（またはプログラミング言語）らしくないものを除外
  ・重複した文書を除く（性能低下の原因になる）
  ・有害情報を削除

Token Crisis

- 強力な大規模言語モデル構築のためには良質かつ大量のテキストが必要
  - Meta Llama 3: 15兆トークン
  - ウェブ空間でさえ足りなくなっているという説も
- 大規模言語モデルを使って自動生成したテキストを訓練に使う
  - Microsoft Phi: 教科書レベルの質のテキスト（13億パラメータモデルに対して60億トークン+GPT3.5で10億の自動生成、ただしプログラムコードに特化）

- https://ai.meta.com/blog/meta-llama-3/
- Xue, Fuzhao, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. 2023. "To Repeat or Not To Repeat: Insights from Scaling LLM under Token-Crisis." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2305.13230.
- Gunasekar, Suriya, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, et al. 2023. "Textbooks Are All You Need." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2306.11644.

コーパスに由来する大規模言語モデルのリスク

バイアスとテキスト復元

テキスト生成におけるリスクはさまざま

- Discrimination, toxicity, and exclusion
  - （特定のグループに対する差別、攻撃的な言明、マイナーな言語の軽視）
- Factual errors, misinformation, and disinformation
  - （事実誤認，誤った情報，偽情報）
- Privacy violations
  - （プライバシー侵害）
  - （あるいは法律・倫理的に問題がある情報の出力，たとえば反社会的な情報など）

Kumar, Sachin, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. "Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey." In *Proceedings of EACL-2023*, 3299-3321.

リスク例：職業に対するバイアス

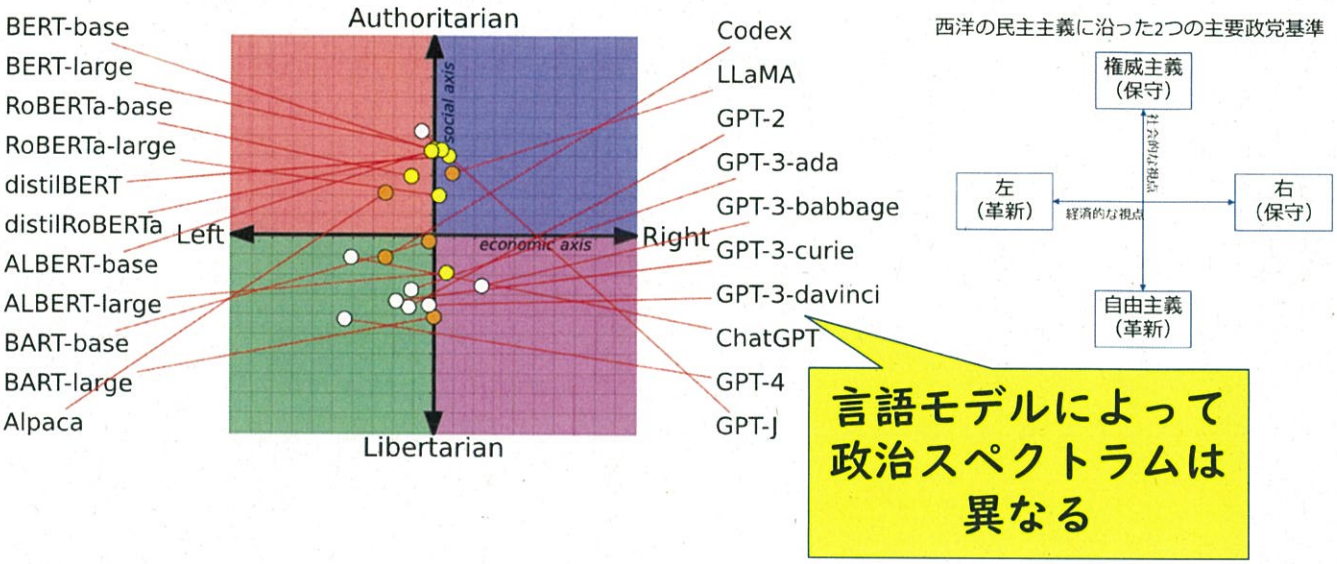
例文：The advisor met with the advisee because she wanted to get advice about job applications.

- 代名詞`She‘の参照先は？
  - Advisor
  - Advisee

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." Advances in Neural Information Processing Systems 33:1877-1901.

リスク例：政治的嗜好に対するバイアス

Shangbin Feng, Chan Young Park, Yuhan Liu and Yulia Tsvetkov: "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models" ACL 2023 Best paper award.



オリジナルテキストの復元問題

・言語モデルはどれくらい暗記しているか？　途中まで文章を与えたときの出力を観察（Kiyomaru+, INLG 2024）

　・何度も出てくるデータほど続きを抽出しやすいか?
　・プロンプトが長いほど続きを抽出しやすいか?
　・終盤に学習したデータほど続きを抽出しやすいか?

・調べるためには学習したテキストを把握している必要がある
・モデルの大きさによる影響も

Kiyomaru, Hirokazu, Issa Sugiura, Daisuke Kawahara, and S. Kurohashi. 2024. "A Comprehensive Analysis of Memorization in Large Language Models." International Conference on Natural Language Generation, 584–96.

テキストコーパスと大規模言語モデル

コーパス中の語

・国立国語研究所のKOTONOHAコーパス中の「野菜」を検索

・文脈＋ターゲット語

https://shonagon.ninjal.ac.jp/

検索結果

6874 件の結果が見つかりました。そのうち 500 件を表示し

表示番号	前文脈	検索文字列	後文脈	執筆者	生年代
1	に来る若奥さん、おかずだけポリ袋につめてテイクアウトするお婆さん、電話注文の肉や	野菜	をバイクで届けにくるお兄さんなども姿を見せ、昼間はのんびりとした時が流れる。　こ	清水　ケンソー(著)	1940
2	年間労働時間の面で他産業従事者とほぼそん色のない水準に達しているとみられる稲作，	野菜	作，酪農の各部門の個別経営及び協業経営体２，５６８経営体を対象に実施したアンケー		
3	レンソウは葉長２５ｃｍ前後で収穫しましょう。葉をジュースにするケール小さな葉もの	野菜	の菜園。手前からラディッシュ、小カブ、キョウナ、ホウレンソウ２果菜類づくり肥料	成松　次郎(著)	
4	ｃ約１％の酢水を沸騰させたのち，野菜を３分間ゆでる。●結果・塩を入れてゆでると，	野菜	の色がよくなる。・酢を入れてゆでると，野菜の色がわるくなる。（しょうゆを入れて煮		

Spring の語義（WordNet）

https://wordnet.princeton.edu/

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: Spring Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **spring**, springtime (the season of growth) *"the emerging buds were a sure sign of spring"; "he will hold office until the spring of next year"*
- S: (n) **spring** (a metal elastic device that returns to its shape or position when pushed or pulled or pressed) *"the spring was broken"*
- S: (n) **spring**, fountain, outflow, outpouring, natural spring (a natural flow of ground water)
- S: (n) **spring** (a point at which water issues forth)
- S: (n) give, **spring**, springiness (the elasticity of something that can be stretched and returns to its original length)
- S: (n) leap, leaping, **spring**, saltation, bound, bounce (a light, self-propelled movement upwards or forwards)

書き言葉均衡コーパス

https://clrd.ninjal.ac.jp/bccwj/

概要 Introduction to BCCWJ

『現代日本語書き言葉均衡コーパス』（BCCWJ）は、現代日本語の書き言葉の全体像を把握するために構築したコーパスであり、現在、日本語について入手可能な唯一の均衡コーパスです。書籍全般、雑誌全般、新聞、白書、ブログ、ネット掲示板、教科書、法律などのジャンルにまたがって１億430万語のデータを格納しており、各ジャンルについて無作為にサンプルを抽出しています。

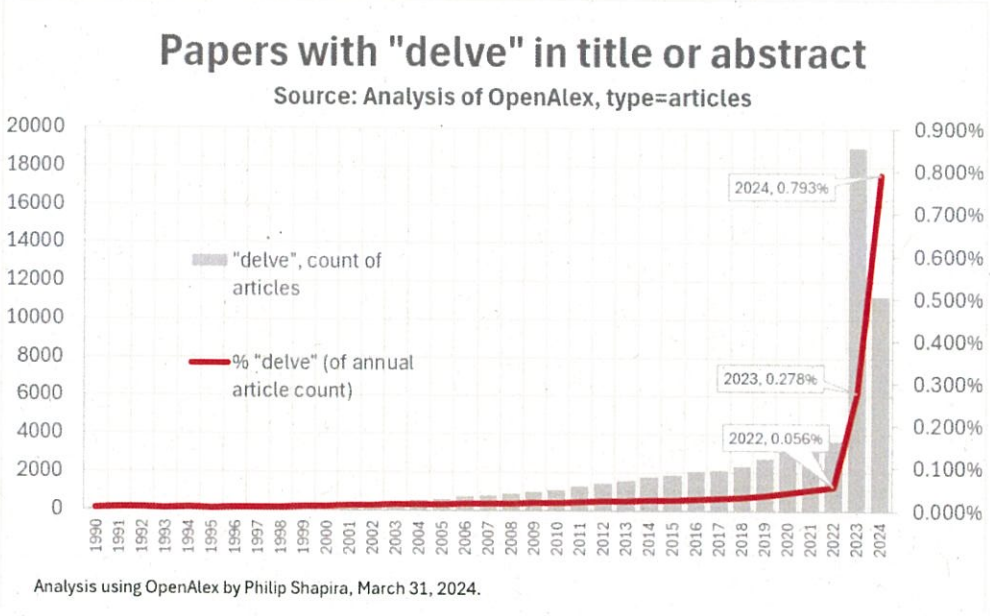
すべてのサンプルは長短ふたつの言語単位を用いて形態素解析されており、さらに文書構造に関するタグや精密な書誌情報も提供されています。著作権処理も施されていますので、安心して使っていただけます。

言語モデル構築における良質なコーパスの使われ方の例

- 品質フィルタリング
  - アノテーションなしのテキスト
  - 事前学習用のテキストの品質を担保するための参照コーパスとして用いる
- モデルの能力評価
  - 人手によるアノテーション
  - 言語モデルのメタ言語能力（例：形態素解析）の測定
- トークナイザの語彙
- 言語空間のモニタリング
  - 計算的に語の意味の変遷や変化を捉える

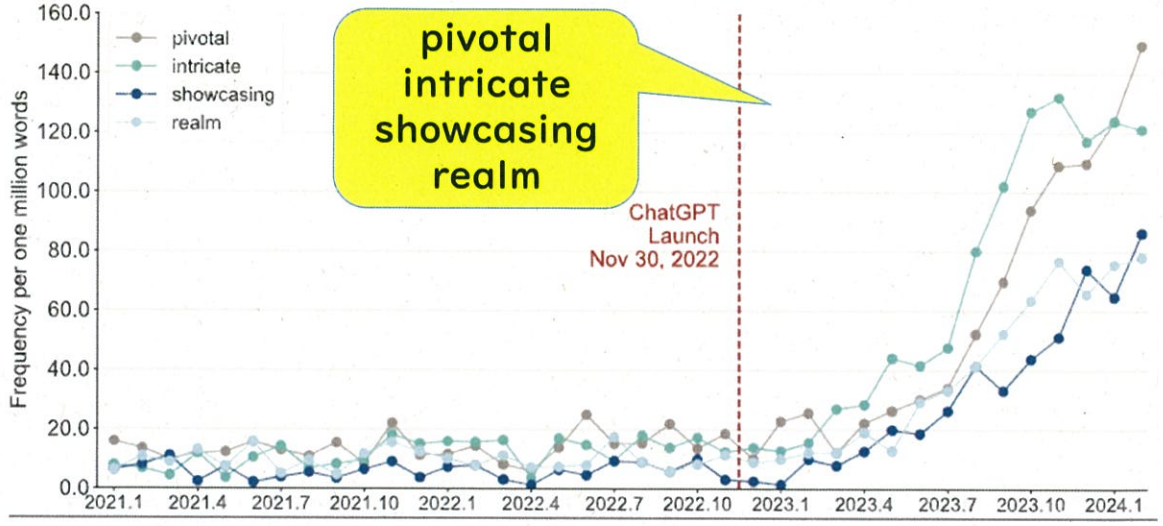
一時期話題になった delve 問題

https://pshapira.net/2024/03/31/delving-into-delve/



言語資源小委員会資料（NII相澤）
2024/9/26

arxiv.org掲載論文における頻出語の経時変化



Liang, Weixin, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, et al. 2024. "Mapping the Increasing Use of LLMs in Scientific Papers." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2404.01268.

生成テキストの再利用

Llama-3 ブログから　https://ai.meta.com/blog/meta-llama-3/

Training data

To train the best language model, the curation of a large, high-quality training da
In line with our design principles, we invested heavily in pretraining data. Llama
over 15T tokens that were all collected from publicly available sources. Our train
seven times larger than that used for Llama 2, and it includes four times more co
upcoming multilingual use cases, over 5% of the Llama 3 pretraining dataset co
quality non-English data that covers over 30 languages. However, we do not exp
of performance in these languages as in English.

To ensure Llama 3 is trained on data of the highest quality, we developed a series of d ning
pipelines. These pipelines include using heuristic filters, NSFW filters, semantic dedu cation
approaches, and text classifiers to predict data quality. We found that previous generations of
Llama are surprisingly good at identifying high-quality data, hence we used Llama 2 to generate
the training data for the text-quality classifiers that are powering Llama 3.

Llama 3で使用するテキスト品質分類器の学習データを生成するためにLlama 2を使用した。

「Model Collapse」

- “We find that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear. “

- 生成したテキストをモデルの学習に再利用することを繰り返すことにより、テキスト本来の多様性が失われて行く現象を“model collapse”と命名

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. “AI Models Collapse When Trained on Recursively Generated Data.” *Nature* 631 (8022): 755–59.

Q&A