日本語埋め込みモデル評価ベンチマークの構築

Shengzhe Li, 大萩雅也, 李凌寒
SB Intuitions株式会社

目次

1. 埋め込みモデルとは

埋め込みモデルとは

一言で言えば，テキスト(単語/文)を埋め込み(embedding)表現に変換するモデルを指す

- 単語レベル: Word2Vecなど
- 文レベル: SentenceBERT

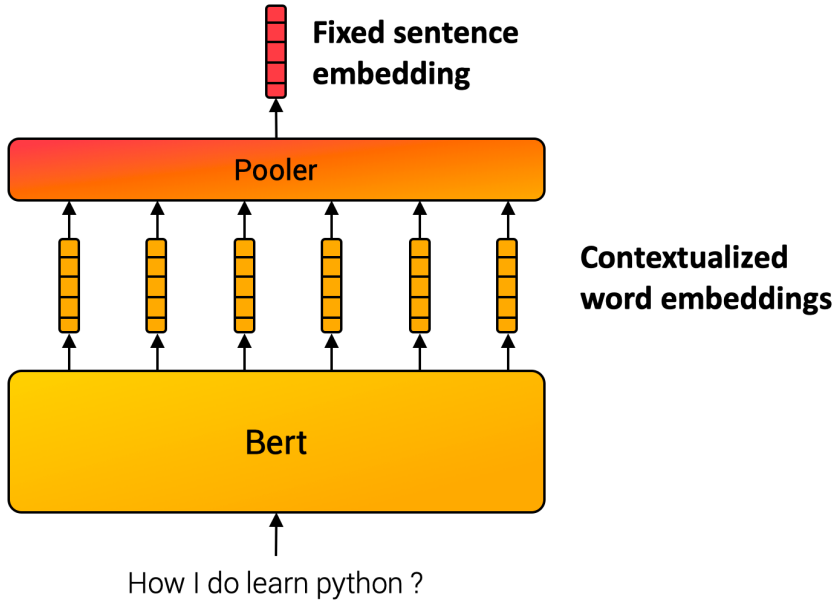


Figure: https://huggingface.co/blog/1b-sentence-embeddings

埋め込みモデルはなぜ重要

- 類似文検出
- クラスタリング
- 情報検索
- …

例：最近流行っているRAG

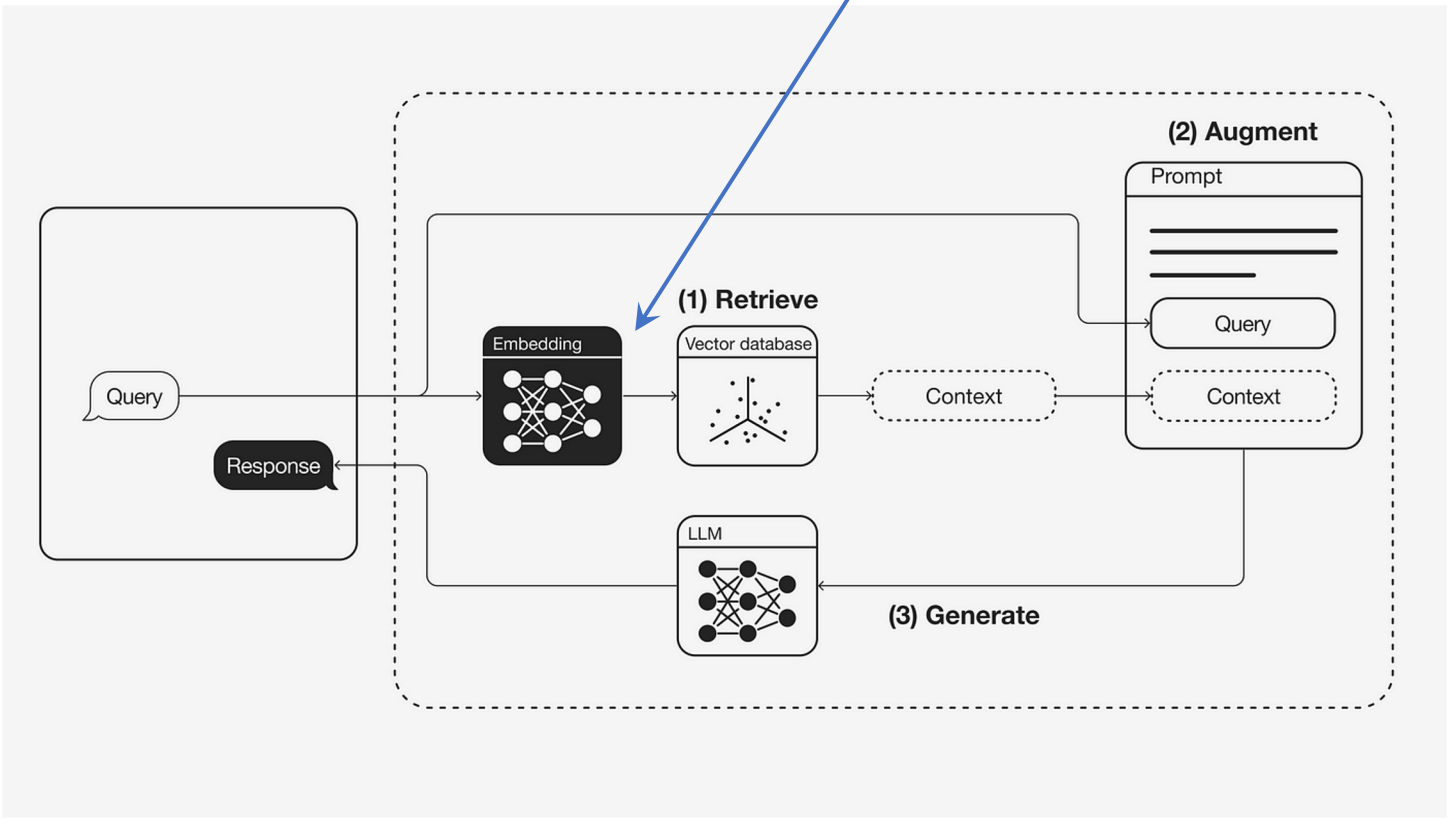


Figure: https://towardsdatascience.com/retrieval-augmented-generation-rag-from-theory-to-langchain-implementation-4e9bd5f6a4f2

文埋め込みモデル構築手法の進化

SentenceBERT (Reimers+, 2019)

- 類似する文章のペアを学習データとし，似た文章同士のベクトルも類似度が高いようにBERTを微調整する手法

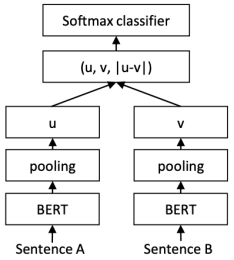


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

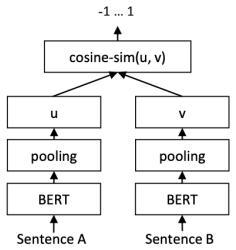


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

SimCSE (Gao+, 2021)

- Contrastive learning (対照学習)を用いてBERTをfinetuningする

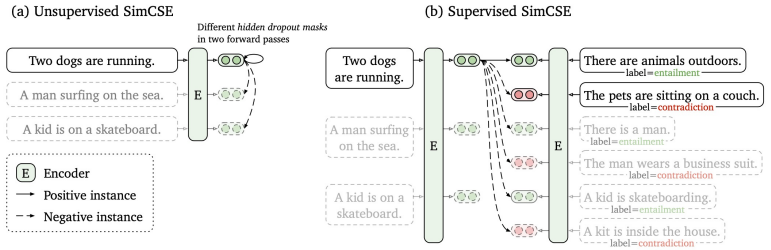
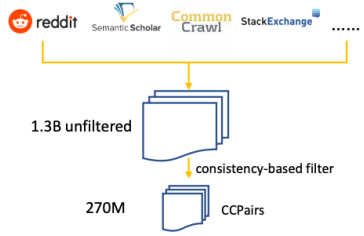


Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different hidden dropout masks applied. (b) Supervised SimCSE leverages the NLI datasets and takes the entailment (premise-hypothesis) pairs as positives, and contradiction pairs as well as other in-batch instances as negatives.

E5 (Wang+, 2022)

- 大規模なqueryとpassageのペア（主にweb crawling）を収集し，事前訓練データセットとして、埋め込みモデルを構築する



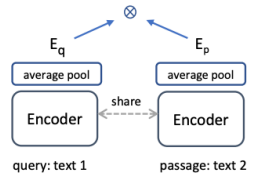


Figure 1: Overview of our data curation pipeline and model architecture.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
Gao, T., Yao, X., & Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821.
Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., ... & Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

2. MTEB: 英語での評価ベンチマーク

MTEBとは

英語*ではMTEB (**M**assive **T**ext **E**mbedding **B**enchmark)という評価ベンチマークが存在しており，埋め込みモデルの評価に広く使われている。

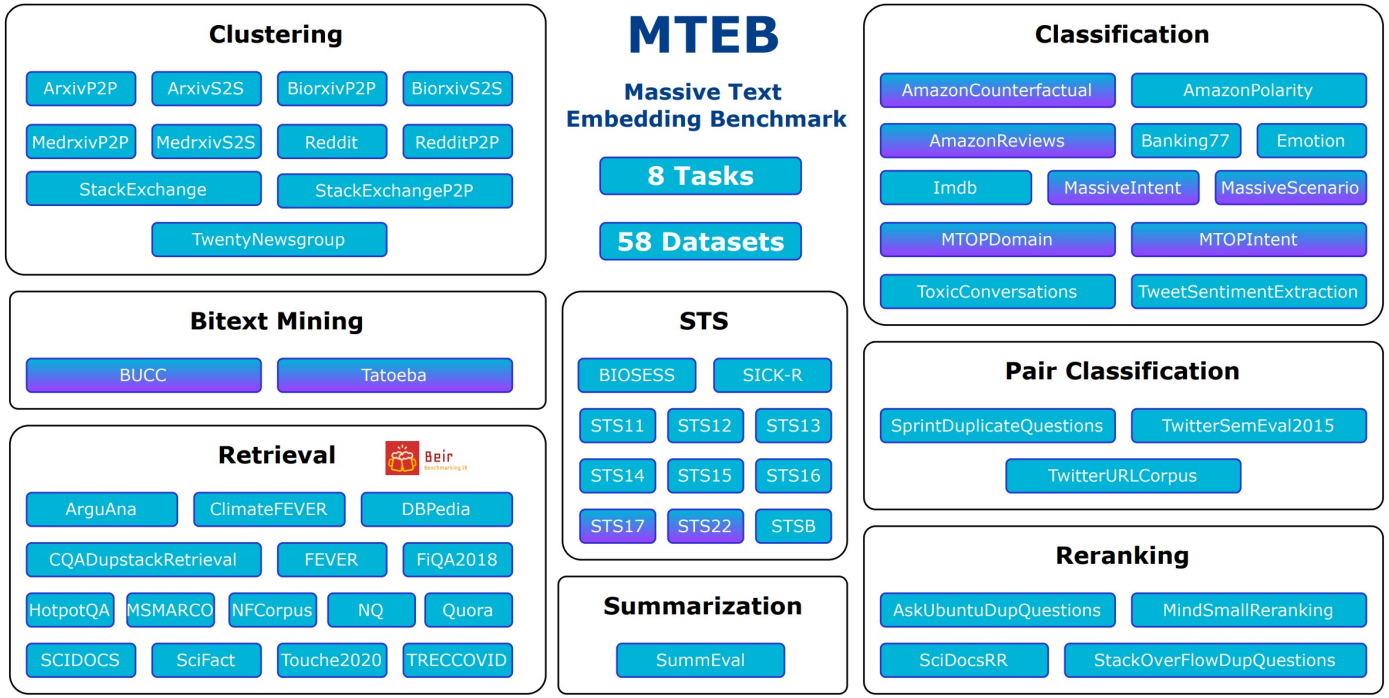


Figure 1: An overview of tasks and datasets in MTEB. Multilingual datasets are marked with a purple shade.

* 一部のタスク (論文発表当初，図の中紫色で表記されたもの。現在は拡充されている。) は多言語対応。

MTEBの設計理念

Diversity	Simplicity
• タスク種類の多様化 • 各タスクにおいてデータセットが複数あること • 文レベルと段落レベル，長いと短い文など様々なテキスト形式があること	• 簡単なAPIで実行できること
Extensibility	**Reproducibility**
• 新データセットが容易に追加できること	• 容易に再現できること

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

MTEBを構成するタスク

タスク名	データセット数	概要
Clustering	11	(意味, 話題などにより)近い文書を同一クラスタに，遠い文書が異なるクラスタにまとめる
Classification	12	各文書もしくはドキュメントの埋め込みを使い，どれほど綺麗にクラス分類ができるかを測る
STS	10	文書ペアの類似度を測る
Pair Classification	3	文書のペアに対して割り当てられたラベルを予測する
Retrieval	15	queryに対して，corpus内から適切なpassageを取り出す
Reranking	4	query, relevant documents, irrelevant documentsが与えられたときに，queryからの類似度をもとにこれらのdocumentを並び直すタスク
Summarization	1	あるドキュメントを人手要約したものと機械要約したものの類似度を測る
Bitext Mining	2	意味が同じ文書を複数言語で表したものを同一と看做せるかを測る
合計	58	

3. 日本語埋め込みモデル評価ベンチマークの構築

現状の日本語埋め込みベンチマークの問題点

- MTEBのような統一的 & タスク横断的なベンチマークが存在しない
- 結果として...
  - 各モデルごとに違うベンチマークでの結果が報告されがち
    - 例えば
      - cl-nagoya: JSICK (test), JSTS (dev)
      - pksha: JSTS (dev), AIO 3
  - どのモデルがどのタスクに強いのかの分析が難しい
    - STSタスクに強いからといって検索タスクにも応用可能かはわからない
    - 数とドメインが限られた評価タスクで出した評価結果ではモデルの性能の汎用性がわからない
    - モデル間の性能比較が難しい

本発表の貢献

- MTEBの日本語版である **JMTEB** (**J**apanese **M**assive **T**ext **E**mbedding **B**enchmark)を構築，公開
  - HuggingFace: https://huggingface.co/datasets/sbintuitions/JMTEB
  - 多様なタスク上での横断的な評価を実現
- JMTEBの評価用スクリプトを公開
  - JMTEB上でそれぞれのモデルを簡単に評価できるようなコードを公開
  - GitHub: https://github.com/sbintuitions/JMTEB

dataset

code

JMTEBの概要

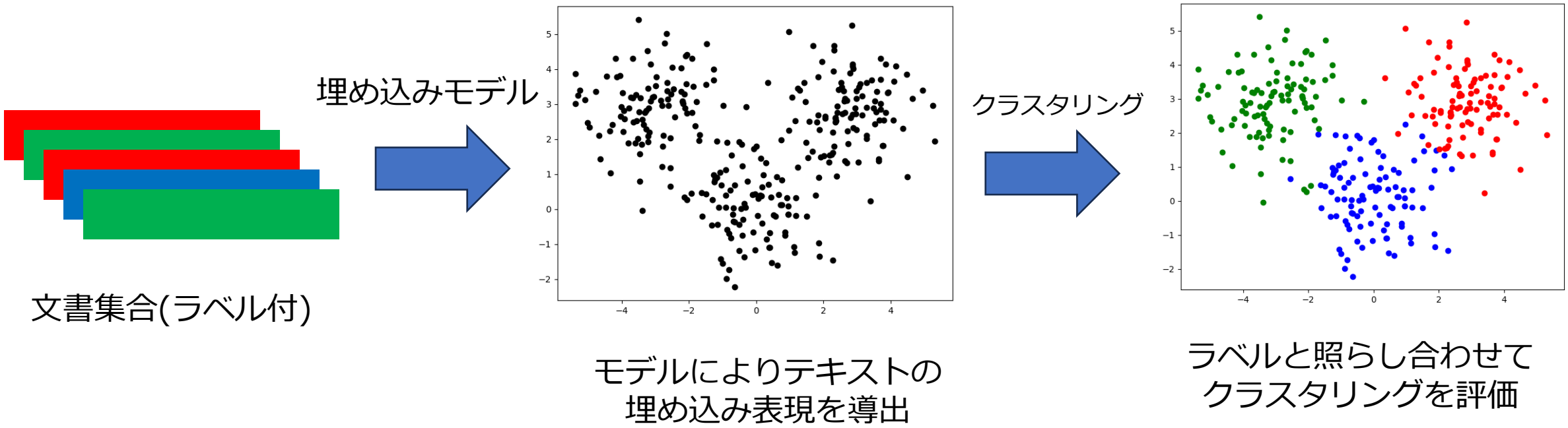
- 多様なタスクで評価するため5つのタスクを設定
- 多様なドメインで評価するために様々なドメイン(e.g. 商品レビュー，論文情報)からデータセットを収集

タスク名	データセット数（3/10現在）	タスク概要
Clustering	2	意味が近い文章を同じクラスタに、遠い文章を異なるクラスタにまとめる
Classification	4	文章がどのクラスに属するかを分類する
STS	2	二つの文章の類似度を測るタスク
Pair Classification	1	二つの文章のペアに割り当てられたラベルを予測する
Retrieval	6	特定のクエリから関連するドキュメントを検索する

Appendix. Aにて，データセットごと詳細な統計データを紹介する。
Appendix. Bにて，評価指標の詳細について紹介する。

JMTEB: Clustering

- 意味が近い文書を同一クラスタにまとめるタスク



応用先: レビューの分析，顧客分析

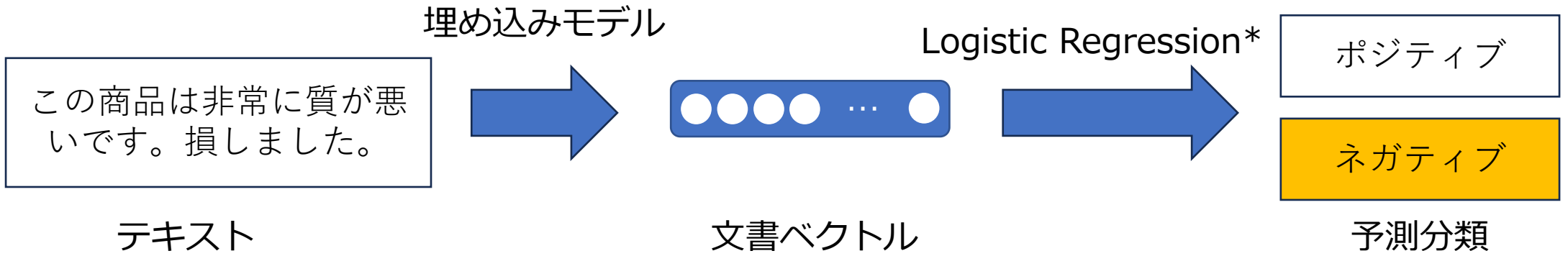
JMTEB: Clustering

- 収集データ

データセット	ラベル数	データ数 train/dev/test	中身
Livedoor-News	9	5,163/1,106/1,107	Livedoorニュースデータセット。2012年に株式会社ロンウイットにより収集した，トピック別のニュース記事。
MewsC-16-ja	12	-/992/992	Wikinewsから作った**M**ultilingual Short Text **C**lustering Dataset for N**ews** in **16** languagesの日本語部分。記事のトピックにより12分類されている。

JMTEB: Classification

- テキストをそれぞれが属するクラスに分類するタスク



応用先: レビューの分析，顧客分析

* このタスクでは埋め込みは固定した上でlogistic layerの学習を行うためtraining dataを必要とする

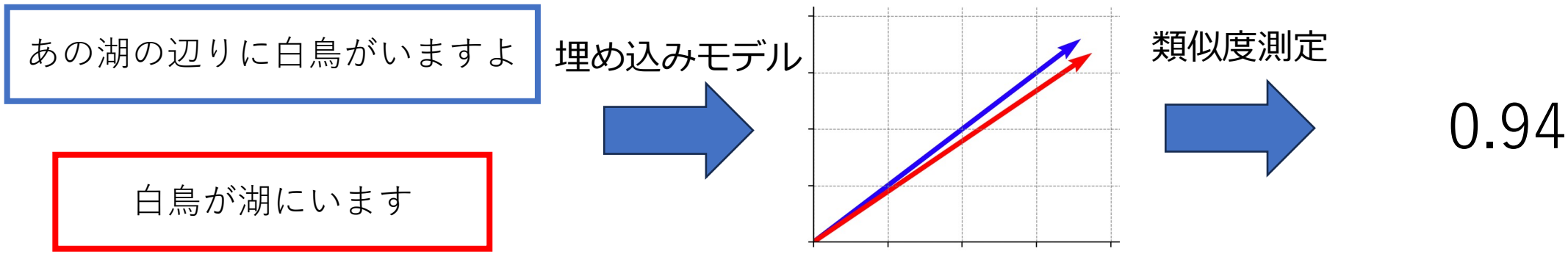
JMTEB: Classification

- 収集データ

データセット	ラベル数	データ数 train/valid/test	中身
AmazonCounterfactualClassification	2	5,600/466/934	Amazonの商品レビュー文に事実と反するものを検出 (詳細)
AmazonReviewClassification	5	200k/5k/5k	Amazon商品のレビュー文のスター数(1~5)を予測
MassiveIntentClassification	60	11,514/2,033/2,974	Alexaのユーザ発話の意図を60種類に分類
MassiveScenarioClassification	18	11,514/2,033/2,974	Alexaのユーザ発話のシナリオを18種類に分類

JMTEB: STS

- 二つの文書の類似度(0~1の間の小数値)を測るタスク



応用先: 文書間の一致判定

JMTEB: STS

- 収集データ

データセット	データ数 train/test	概要
JSTS	12,451/1,457	二つの文書間の類似度を測る
JSICK	7,941/1,986	

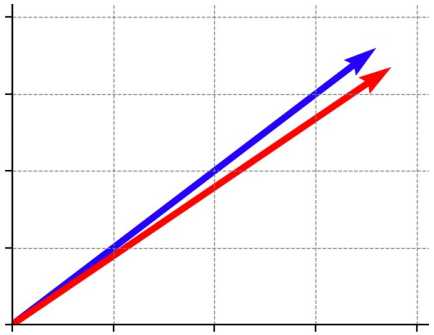
JMTEB: Pair classification

- 二つの文書の組み合わせがどのようなクラスに属するかを分類する

あの湖の辺りに白鳥がいますよ

白鳥が湖にいます

埋め込みモデル



類似度を閾値で判定

言い換え

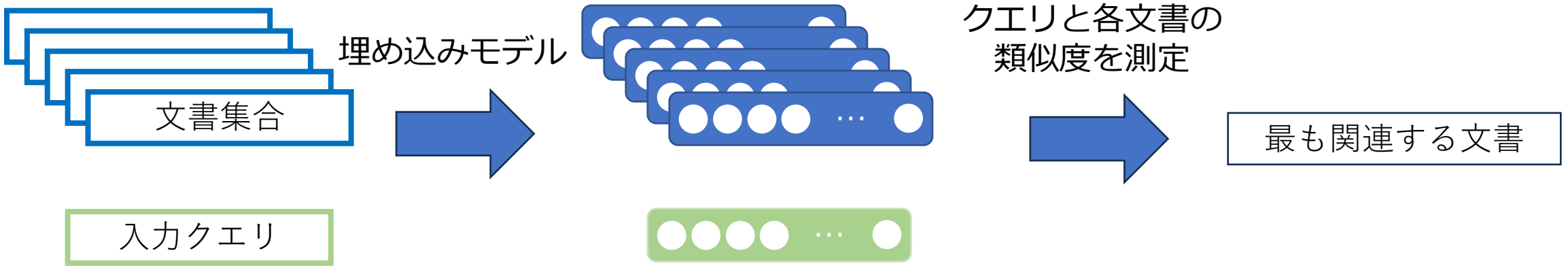
言い換えでない

JMTEB: Pair classification

データセット	ラベル数	データ数 train/dev/test	中身
PAWS-X-ja	2	49,401/2,000/2,000	PAWS-X(言い換え表現判定データセット)の日本語部分

JMTEB: Retrieval

- 文書集合から入力クエリに最も関連する文書を導き出す

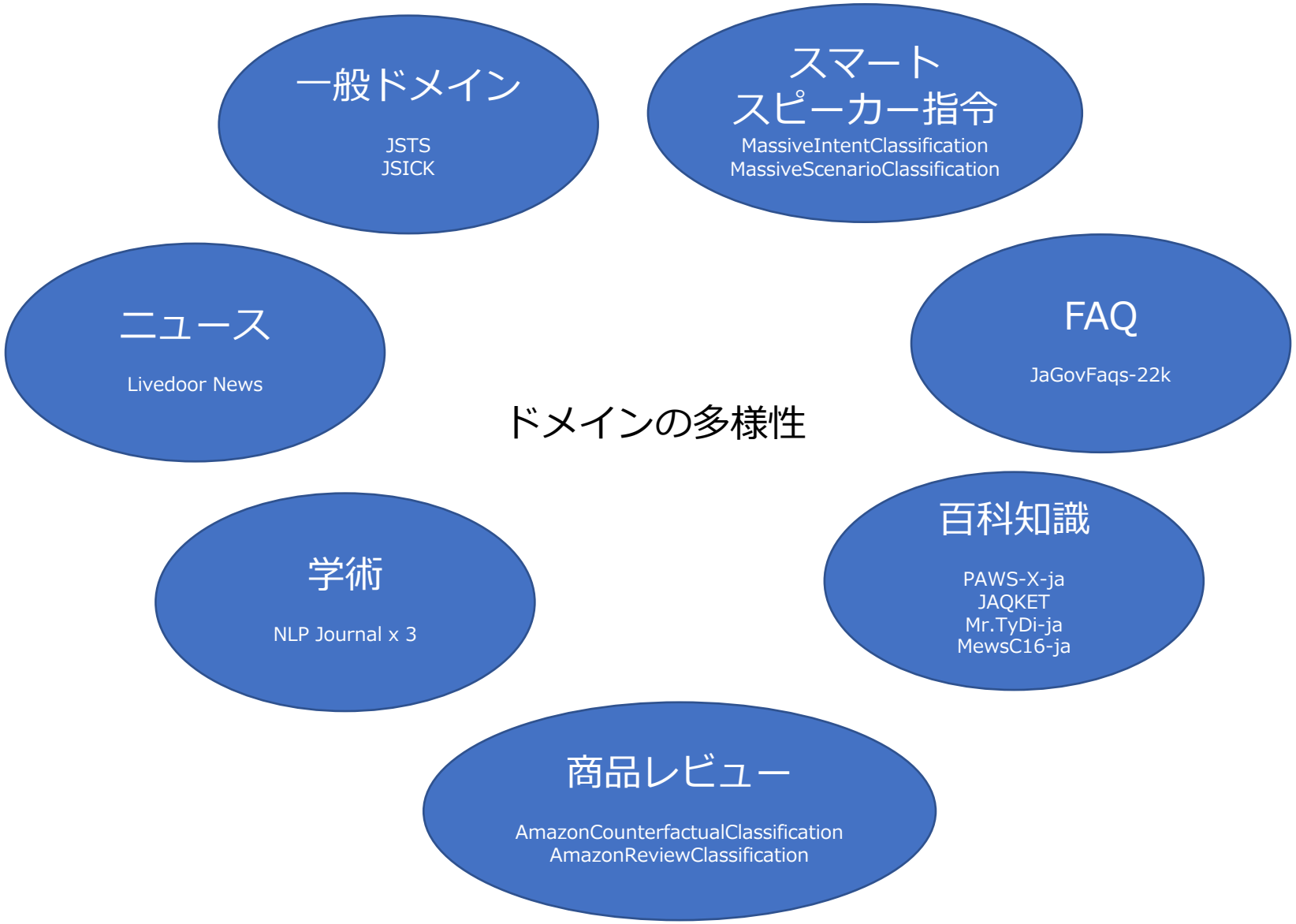


応用先: FAQ, RAG (検索つき生成)

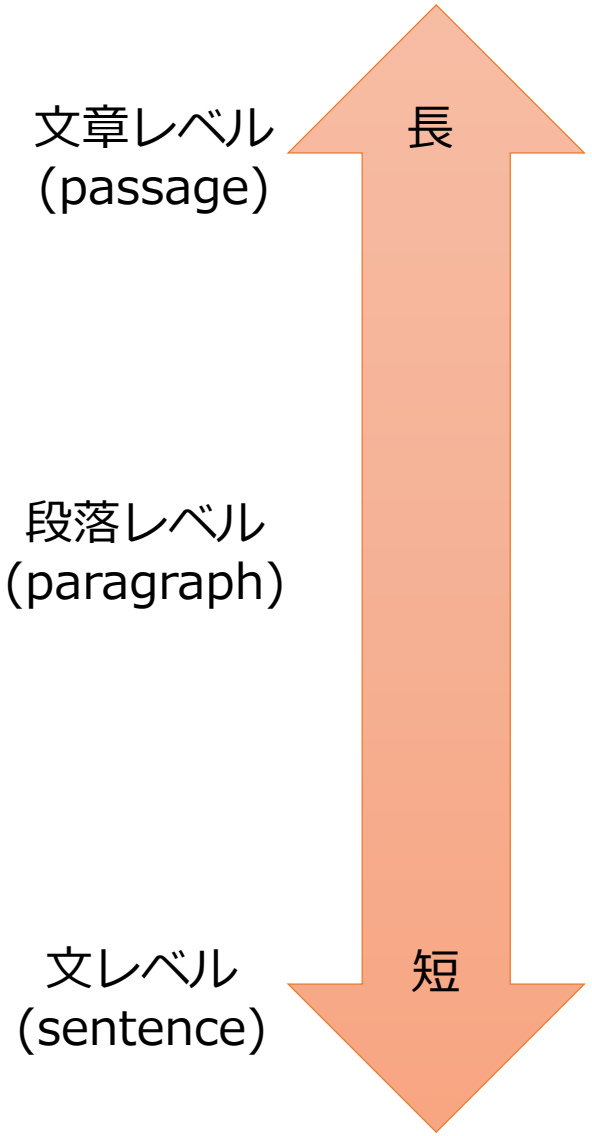
JMTEB: Retrieval

データセット	query件数 dev/test	corpus件数	概要
JAQKET (AIO ver.1)	995/997	114,229	クイズ題材のQA。問題文を入力クエリとし，関連する文書をwikipediaの記事から検索する。
Mr.TyDi-ja	928/720	7,000,027	問題文をコーパス(約700万文)に照らし合わせ，答えが含まれる文を抽出する。
JaGovFaqs-22k	3,419/3,420	22,794	日本の官公庁のWebサイトに掲載されている「よくある質問」を抽出したもの。
NLP Journal title-abs	-/504	504	言語処理学会論文誌LaTeXコーパスから抽出し，我々が構築した，論文のタイトル，概要文とイントロ (1st section)。Query (title or abstract)を与え，同じ論文のabstract or introductionをコーパスから最大類似度基準で抽出する。
NLP Journal title-intro			
NLP Journal abs-intro			

データセット・ドメインの多様性



テキスト形式の多様性



JAQKET (corpus)
NLP Journal (introduction)
Livedoor News

NLP Journal (abstract)

Mr.TyDi-ja (corpus)
JaGovFaqs-22k (corpus)
MewsC16-ja
AmazonReviewClassification

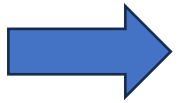
PAWS-X-ja
AmazonCounterfactualClassification

JAQKET (question)
JaGovFaqs-22k (question)
JSTS, JSICK
NLP Journal (title)
Mr.TyDi-ja (question)
MassiveIntentClassification
MassiveScenarioClassification

JMTEB: 評価コード

- 各データセット上での埋め込みモデルの結果を簡単に分析できる評価コードを商用利用可能で公開: https://github.com/sbintuitions/JMTEB
- 以下のようなコマンドだけで，評価結果まで出力することができる

```
poetry run python main.py \
  --embedder SentenceBertEmbedder \
  --embedder.model_name_or_path "sbintuitions/model-name" \
  --save_dir "results"
```



```
{
  "Clustering": {
    "livedoor_news": {"v_measure_score": 0.5163},
    "mewsc16_ja": {"v_measure_score": 0.5270}
  },
  "STS": {
    "jsts": {"spearman": 0.8169},
    "jsick": {"spearman": 0.7798}
  },
  ...
}
```

JMTEB: まとめ

- MTEBの日本語版を作成
  - 合計5タスク，15データセットで様々な側面から埋め込みモデルを評価できる
  - 今後も継続的にデータセットを追加していく予定
- 評価コードも公開する
- JMTEBを用いることでどういう分析ができるかを次の章で解説

4. JMTEBによる既存の埋め込みモデルの分析

全体の評価結果

Model	Classification平均	Clustering平均	STS平均	PairClassification平均	Retrieval平均	15タスクの平均
cl-nagoya/sup-simcse-ja-base	73.47	**52.98**	82.05	62.62	49.81	61.69
cl-nagoya/sup-simcse-ja-large	73.73	50.10	**83.44**	62.53	38.11	56.88
cl-nagoya/unsup-simcse-ja-base	73.07	44.58	78.85	62.44	40.85	56.44
cl-nagoya/unsup-simcse-ja-large	74.66	46.20	80.78	62.51	41.34	57.54
pkshatech/GLuCoSE-base-ja	**76.83**	49.53	78.72	**66.50**	59.29	65.74
pkshatech/simcse-ja-bert-base-clcmlp	71.30	44.50	76.96	62.51	37.15	54.23
sonoisa/sentence-luke-japanese-lite	76.11	36.29	80.62	62.46	40.21	56.13
oshizo/sbert-jsnli-luke-japanese-base-lite	72.83	51.04	76.60	62.44	42.55	57.62
MU-Kindai/Japanese-SimCSE-BERT-base-sup	72.76	46.68	74.66	62.44	41.47	56.33
MU-Kindai/Japanese-SimCSE-BERT-base-unsup	73.30	46.02	77.50	62.57	46.77	58.89
MU-Kindai/Japanese-SimCSE-BERT-large-sup	73.47	43.28	78.28	62.51	41.29	56.48
MU-Kindai/Japanese-SimCSE-BERT-large-unsup	73.13	43.56	78.99	62.50	47.68	59.08
MU-Kindai/Japanese-DiffCSE-BERT-base	73.77	44.52	75.50	62.47	42.18	56.71
MU-Kindai/Japanese-MixCSE-BERT-base	61.85	43.43	77.05	62.51	42.76	53.83
sentence-transformers/LaBSE	72.66	44.38	76.56	62.42	39.61	55.51
intfloat/multilingual-e5-small	67.62	51.88	80.08	62.42	67.60	66.83
intfloat/multilingual-e5-base	69.30	48.91	79.84	62.42	68.41	67.17
intfloat/multilingual-e5-large	72.89	47.62	79.70	62.42	**71.11**	**69.02**

評価タスクを拡張すると何が変わる

分析例1: 直感が必ず正しいわけではない

model	STS		Clustering	Classification	Retrieval	
	JSTS	JSICK	Livedoor News	AmazonReview	JAQKET	NLP Journal title-abs
cl-nagoya/sup-simcse-ja-base	80.85	82.83	**55.01**	40.97	**50.21**	**66.36**
cl-nagoya/sup-simcse-ja-large	**83.08**	**83.80**	46.50	41.85	39.84	27.72
cl-nagoya/unsup-simcse-ja-base	78.95	78.49	52.25	**44.70**	39.47	57.16
cl-nagoya/unsup-simcse-ja-large	81.41	80.15	45.17	44.64	34.62	64.76

- 直感 + STSのみで評価する（赤い枠内の結果しか知らない）場合の知見
  - largeがbaseより性能良い
  - supがunsupより性能良い
- ベンチマークで評価する（全面的に評価を行う）場合の知見
  - largeがbaseより良いとは限らない（一部のタスクではlargeの性能が逆に劣る）
  - supがunsupより良いとは限らない（catastrophic forgettingの可能性）

評価タスクを拡張すると何が変わる

分析例2: 訓練データが多ければ多いほど、性能が高い

Model	Classification	Clustering	STS	Retrieval	13タスク平均
(A) pkshatech/GLuCoSE-base-ja	**76.83**	49.53	**78.37**	**65.10**	**65.69**
(B) oshizo/sbert-jsnli-luke-japanese-base-lite	72.83	**51.04**	76.60	49.34	57.62

＊ (A)モデルが訓練に入ったPAWS-XとMr.TyDi-jaタスクは比較から除去した。

- 両方とも，studio-ousia/luke-japanese-base-liteをバックボーンとして訓練された
- 両モデルの訓練データセット
  - (A) mC4, MQA, JNLI, JSNLI, PAWS-X, JSeM, MoritzLaurer/multilingual-NLI-26lang-2mil7, JSICK, Mr.Tidy
  - (B) JSNLI
- 考察
  - 訓練データが多ければ多いほど，性能が高いという傾向が見られる
  - 特に，(A)の訓練では検索データを用いたため，検索での性能は(B)よりはるかに優れている

＊ 本ページの考察では，厳密な局所管理化は行っていない。

評価タスクを拡張すると何が変わる

分析例3: 訓練データセットの選び方によって，得意なタスクが変わる

Model	Classification	Clustering	STS	Retrieval	13タスク平均
(A) pkshatech/GLuCoSE-base-ja	**76.83**	**49.53**	78.37	65.10	65.10
(C) intfloat/multiligual-e5-base	69.30	48.91	**79.83**	**76.60**	**76.60**

＊ (A)モデルが事前訓練に入ったPAWS-Xと，(A)(C)が共に訓練に使用したMr.TyDi-jaタスクは比較から除去した。

- 訓練データセットの違い
  - (A) mC4, JNLI, JSNLI, PAWS-X, JSICK(一般), Mr.Tidy, MQA(検索)などある
  - (C) 第一段階(検索の弱教師あり対照学習)にfiltered mC4, CC News, Stackexchangeなどのquery-docペア, 第二段階(検索の教師あり学習)にMS MARCO, Trivia QA, SQuAD, Quora, Mr.TyDiなどを使用。英語データをメインとし，多言語データに日本語データが含まれる
- 考察
  - (C)モデル(E5)が検索タスクに大幅に優れている一つの要因は， 多くの検索データで学習させたからと考えられる
  - Classification, Clusteringでは日本語の意味に関する知識が求められるため，(A)の方が全て日本語データを用いたことに対して，(C)モデルがマルチリンガルであるため日本語データセットの比率が低いため，(A)の方が性能高い
  - (C)モデルがquery-docペアで学習させたから，同じく文ペアの形式であるSTSタスクでも高い性能を示していると思われる

＊ 本ページの考察では，厳密な局所管理化は行っていない。

評価タスクを拡張すると何が変わる

分析例4: ベクトル正規化の影響 (1)

```python
class SentenceBertEmbedder(TextEmbedder):
    """SentenceBERT embedder."""

    def __init__(
        self,
        model_name_or_path: str,
        batch_size: int = 32,
        device: str | None = None,
        normalize_embeddings: bool = False,
    ) -> None:
        self.model = SentenceTransformer(model_name_or_path)
        self.batch_size = batch_size
        self.device = device
        self.normalize_embeddings = normalize_embeddings

    def encode(self, text: str | list[str]) -> np.ndarray:
        return self.model.encode(
            text,
            convert_to_numpy=True,
            batch_size=self.batch_size,
            device=self.device,
            normalize_embeddings=self.normalize_embeddings,
        )

    def get_output_dim(self) -> int:
        return self.model.get_sentence_embedding_dimension()
```

- **normalize_embeddings** – Whether to normalize returned vectors to have length 1. In that case, the faster dot-product (util.dot_score) instead of cosine similarity can be used.

$$\mathbf{e} \leftarrow \frac{\mathbf{e}}{|\mathbf{e}|}$$

- 「距離」の求め方が色々あるが，正規化したら，距離の度量 (cosine, dot product, Euclidean distance) が全てcosine類似度と等価になる
- 文をエンコードする時，正規化すべきか？

評価タスクを拡張すると何が変わる

分析例4: ベクトル正規化の影響 (2)

- 正規化するかしないか，タスクのパフォーマンスに影響があるかどうか，調べてみた。

Model	正規化	Classification平均	Clustering平均	STS平均	PairClassification平均	Retrieval平均	15タスク平均
18モデル*の平均スコア	有	69.47	48.94	79.44	62.78	51.80	60.55
	無	**71.52**	48.67	79.53	62.80	**52.53**	**61.36**
cl-nagoya/sup-simcse-ja-base	有	67.38	54.03	81.84	62.54	46.50	58.85
	無	73.47	52.98	82.05	62.62	49.81	61.66
pkshatech/GLuCoSE-base-ja	有	74.35	47.94	78.37	66.31	59.18	64.76
	無	76.83	49.53	78.72	66.50	59.29	65.69

*全てのモデルでは、訓練時cosine similarityで算出された損失関数が用いられる。

- 考察: STS，ClusteringとPairClassificationが顕著な差がないが、ClassificationとRetrievalにおいては明らかに正規化しない方がスコア高い。
  - 15タスク中，”正規化あり>正規化なし“タスクは2つ (AmazonReivewClassificationとMewsC16-ja)のみある
  - その理由は，正規化したらベクトルのスケール情報が消えてしまうと考えられる

結論: 大体のタスクでは、正規化する必要がない。

まとめ

- 日本語埋め込みモデル評価ベンチマーク**JMTEB**を構築・公開した
  - データセット: https://huggingface.co/datasets/sbintuitions/JMTEB (再掲)
  - 評価スクリプト: https://github.com/sbintuitions/JMTEB (再掲)

dataset

code

- JMTEBを用いてモデルの性能分析をしてみた
- これからも，ベンチマークを継続的に拡張する予定

ご清聴ありがとうございました

Appendix A. データセットの詳細

* 特に説明がなければ，統計データがスコアの算出ベースであるtest setに基づいて導出されたものである。

A.0.1 データセットのライセンス

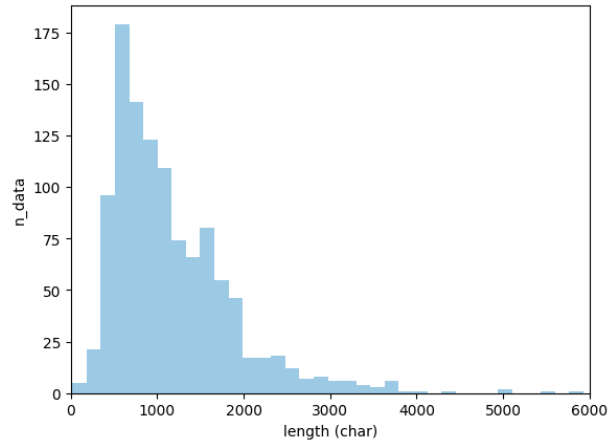
ライセンス	データセット
Apache-2.0	MrTyDi-ja
CC-BY-SA-4.0	AmazonCounterfactualClassification MassiveIntentClassification MassiveScenarioClassification MewsC-16-ja JSTS JSICK JAQKET NLP Journal x 3 JaGovFaqs-22k
CC-BY-ND-2.1	Livedoor News
No warranty (使用自由)	PAWS-X-ja
Non-commercial	AmazonReviewClassification

A.0.2 Data splitの取り扱い

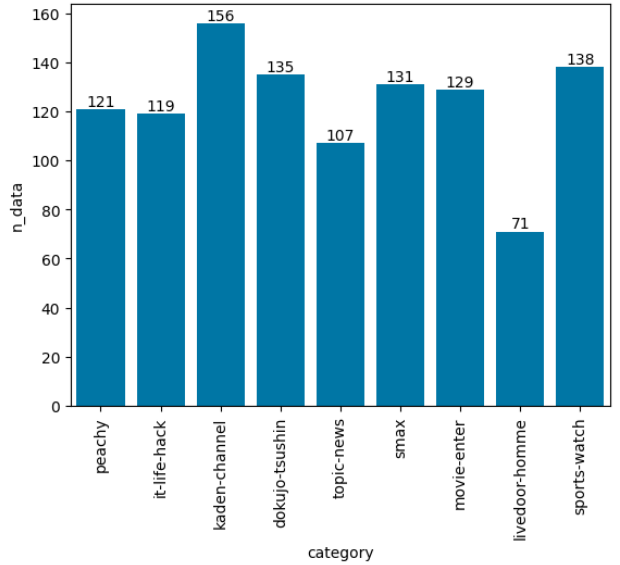
- 大前提
  - 評価スコアを**テストセットのみ**で算出する
  - Classificationではlogistic regressionを学習する必要があるため，train setで学習し，dev setでparameter tuning用とする
  - Classification以外のタスクでは，埋め込みモデルの評価ではtrain setを用いない
- 元データセットがtrain, dev, test set揃っている場合はそのまま流用する
- そうではない場合
  - Classificationでは，devがない場合，train setからtest setと同じ数のデータを無作為で取り出し，dev setとする
  - JGLUEにあるSTSでは，test setが公開されず，dev setしか公開されない場合， dev → testとスライドする
  - Retrievalでは，test setのみ利用する

A.1 Clustering – Livedoor News

Label	Text
it-life-hack	Youtubeやニコニコ動画といった動画系サービスやニコニコ生放送、USTREAMやJustinといったユーザー生放送サービスでは、動画をローカルPCに保存する方法が限られる。できないわけではないが、PCの知識がないと難しい。具体的には、サーバー上にアップされている動画をダウンロードするわけだが、今後は違法ダウンロードの罰則化が始まるので、Youtubeやニコ動にアップされた動画を無許可でダウンロードする行為はNGになってしまう。 … (総計961文字)
movie-enter	今年9月、200万部を突破した大人気コミックを原作者・久保ミツロウ書き下ろしによるオリジナルストーリーで映画化した『モテキ』。21億を超える興行収入を記録した大ヒット映画が2012年3月23日、ブルーレイ＆DVD化され発売されることが決定した。映画は、原作のラストから１年後を描くオリジナルストーリー。ついに伝説の“セカンドモテキ”がやってくる。 … (総計898文字)
smax	シャープのハイスペックスマートフォン「AQUOS PHONE Xx 」は買いだ！既報の通り、ソフトバンクモバイルおよびウィルコムは29日、2012年夏以降に発売する予定の夏モデル発表会「ソフトバンクモバイル・ウィルコム新商品発表会 2012 Summer」を都内で開催し、高速通信サービスULTRA SPEEDに対応したシャープ製のスマートフォン「AQUOS PHONE Xx SoftBank 106SH(アクオス・フォン・ダブルエックス・ソフトバンク・イチゼロロクエスシチ)」（以下、106SH）を7月上旬以降に発売すると発表した。展示会場で106SHを試すことができたので写真と動画で紹介する。 …(総計2491文字)
sports-watch	W杯アジア最終予選が来月3日より幕を開ける。18日放送、日本テレビ「NEWS ZERO」では、日本代表FWとして活躍が期待される、シュトゥットガルト所属・岡崎慎司が生出演を果たした。ドイツでの生活を「最初よりは慣れたと思いますし、会話とか。コミュニケーションの取り方とかも掴んできた」と話す岡崎だが、シュトゥットガルトのチーム内では“キレキャラ”だという。 …(総計863文字)



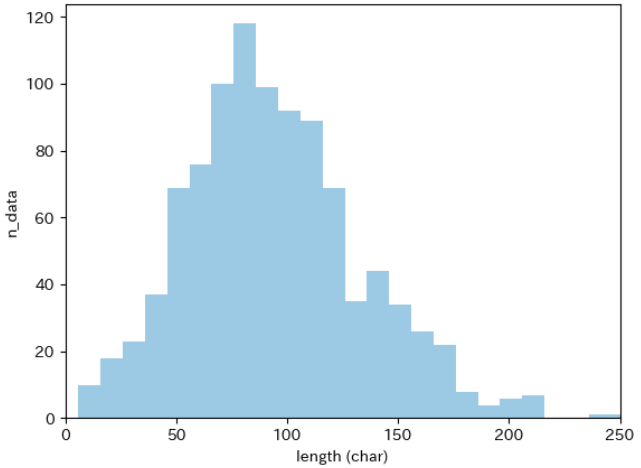
文の長さ分布



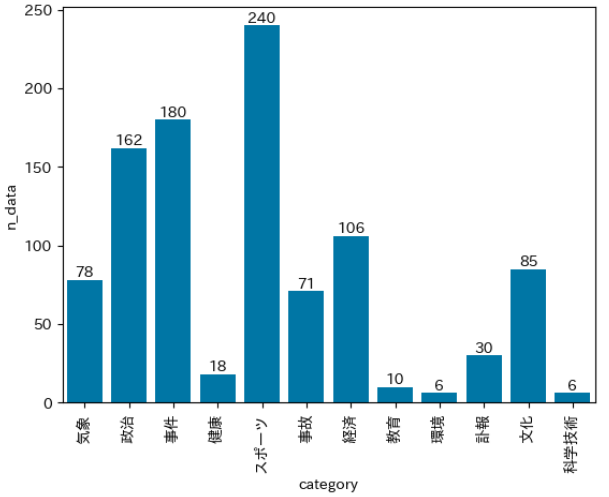
ラベル分布

A.1 Clustering – MewsC16-ja

Label	Text
スポーツ	この1月31日に投開票される日本サッカー協会の会長選で、その立候補者を選出する管理委員会は1月8日に、原博実専務理事とFIFA（国際サッカー連盟）理事も務める田嶋幸三副会長の2人に絞り込まれたことを明らかにした。
気象	朝日新聞によると、7月27日（　以下同）群馬県館林市（たてばやし）を中心に発生し、大きな被害をもたらした突風について、前橋地方気象台は7月28日、これを竜巻と断定したことを発表した。
経済	読売新聞、朝日新聞によると、村上ファンド主宰者・村上世彰氏が阪神電鉄の株式公開買い付け (TOB) に応じる方針であることを、阪急ホールディングス関係者が3日明らかにした。
教育	東京地方裁判所（資料）共同通信によると、東京都教育長が、国旗に向かって起立したり国歌を歌ったりすることを拒否したとして東京都立高等学校、盲ろう養護学校の教諭らを処分したことについて、東京地方裁判所は21日、起立したり歌ったりする義務がないとして、東京都と東京都教育委員会（都教委）に、国歌を歌わないことなどによる処分を禁じ、東京都にひとりあたり3万円の損害賠償を命じた。
事件	産経新聞・日刊スポーツによると、福島県郡山市は13日、勤務時間中に職場のパソコンで、インターネット上の百科事典「ウィキペディア」の業務とは関係ない項目への投稿や閲覧を400回以上行った為、30代の男性職員を減給懲戒処分とした。



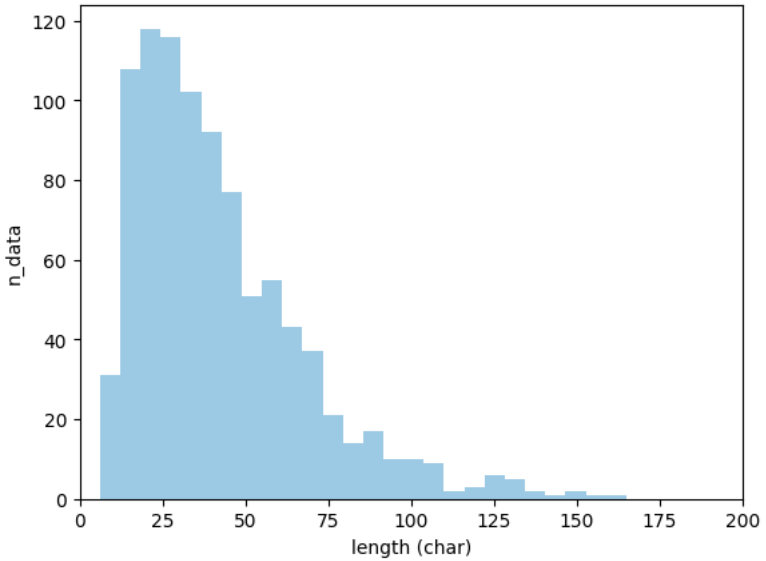
文の長さ分布



ラベル分布

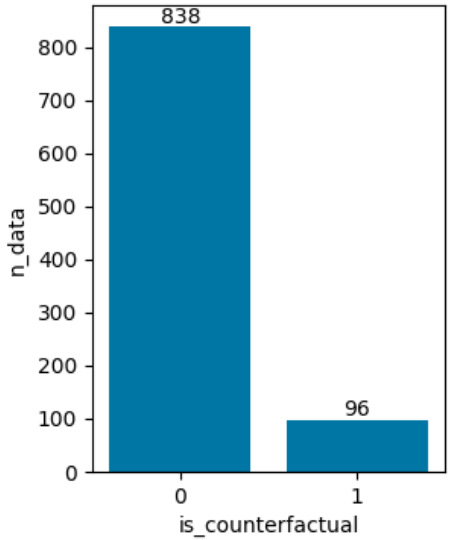
A.2 Classification – AmazonCounterfactualClassification

Label	Text
0	単純に楽しめる作品だと思います。
0	音質は想像以上に良かったです。
0	これもミリタリー仕様なのでしょうか、とても使いやすいです。
0	髪も毛糸ですし、顔も過去最高に似ていると思います。
1	ボタンの位置が別の位置にあれば星5個だったんですが。
1	長期間嵌めていなくても止まらないこと、なのでソーラーが欲しかった。



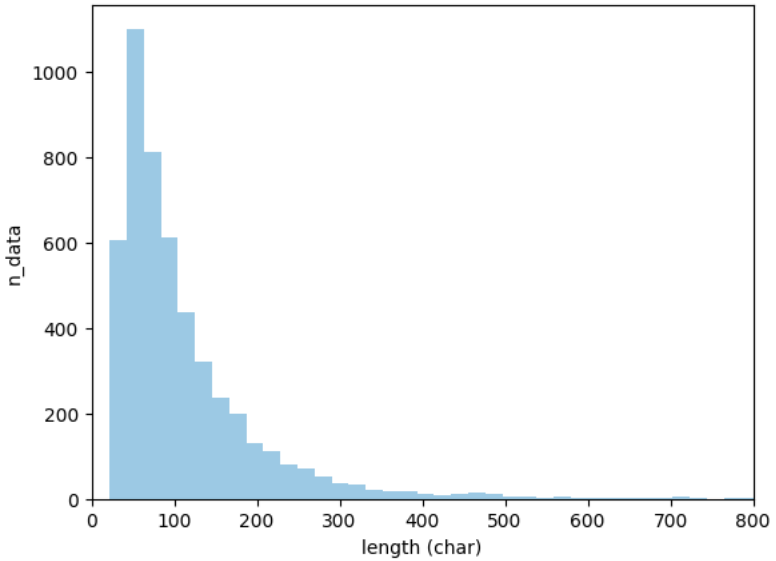
文の長さ分布

ラベル分布

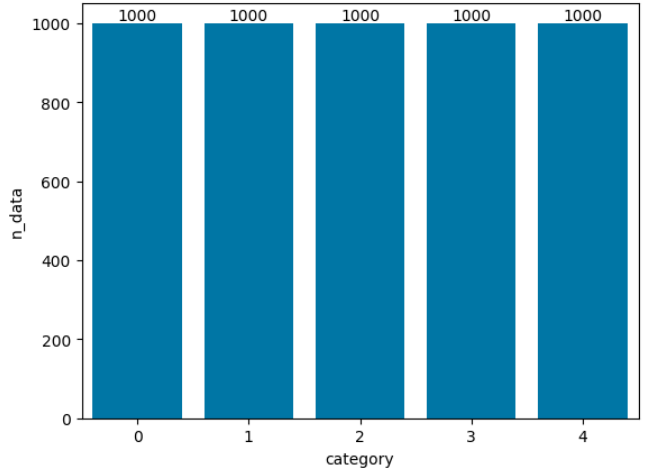


A-2 Classification – AmazonReviewClassification

Label	Text
0	すぐ壊れた たったの一か月半で壊れました。 一か月以上経ってるから返品もできず。。。泣き寝入りです。 粗悪品です。
1	縫製の改善 比較的安価。 縫製が緩く、雑、ほぐれやすい。耐久性に疑問。
2	磁力は強力。だけどねぇ・・・・ 確かに磁石は強力です。けどあまり重いモノを掛けるとフックと磁石の取り付け部分が外れます。
3	コスパ良し お風呂場で使うには良かった。コンパクトで、光るし、音もそれなりの質
4	音の臨場感！ スマホやPCでゲームをやる時に使用しています！ ヘッドセットも考えたんですがもっとライトに使えるものないかと思い探していたところ こちらのカナル型のゲーミングイヤホンを発見！ 初めはそこまで期待していませんでしたが想像以上に音の臨場感がありスマホではPUBGやモダコンを パソコンではBF5やCODをやっていますが個人的にはかなり満足いく音質でした！ マイク（取り外し可）も付いているので試しにスカイプで会話してみましたがまったく問題ありませんでした。 ただ音楽を聴く用途としては結構派手な音なので長時間聴いてると疲れるのでその用途ならほかのイヤホンのがいいかなと。 コスパはかなりいいので映画やゲーム用途でカナル型のゲーミングイヤホンを探してる方でしたらおススメです！



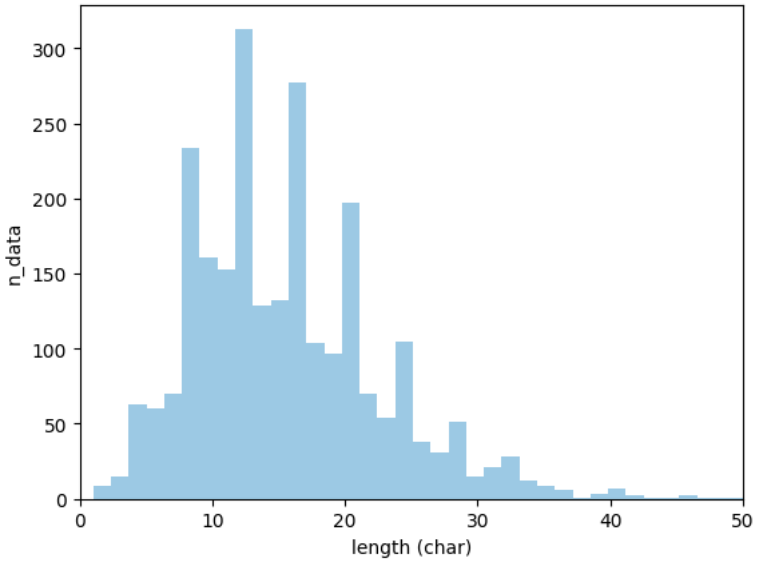
文の長さ分布



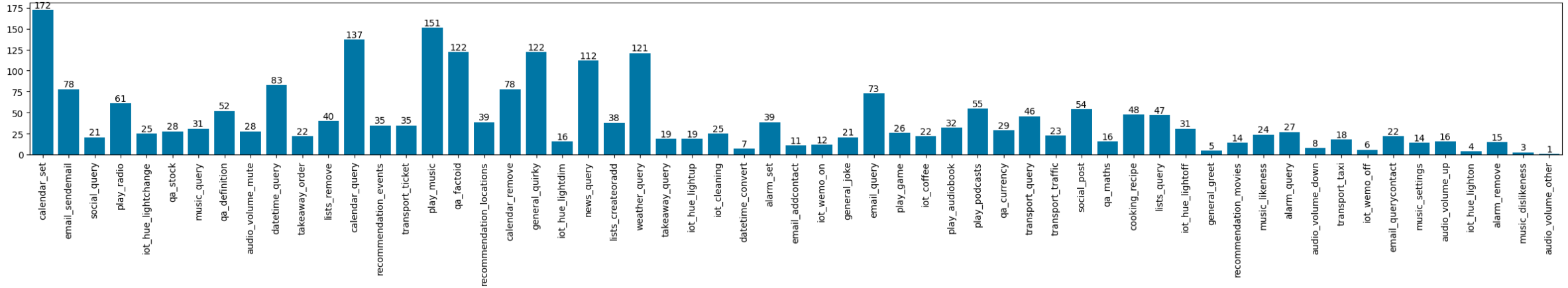
ラベル分布

A-2 Classification – MassiveIntentClassification

Label	Text
weather_query	名古屋市の天気はどうですか
recommendation_events	横浜で今週末開いている動物園を教えて
transport_query	京都までの電車の時間を教えてくれないか。
general_greet	おはよう、何かあった
datetime_query	次の月曜日は何日ですか
play_music	ラップ



文の長さ分布

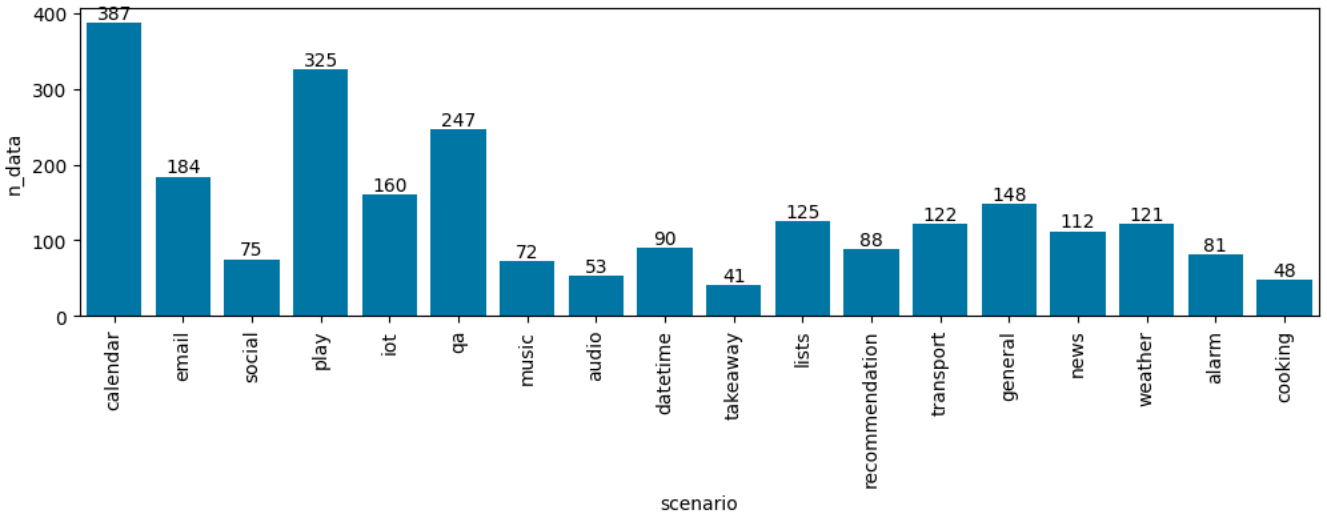


ラベル分布

A-3 Classification – MassiveScenarioClassification

Label	Text
weather	名古屋市の天気はどうですか
recommendation	横浜で今週末開いている動物園を教えて
transport	京都までの電車の時間を教えてくれないか。
general	おはよう、何かあった
datetime	次の月曜日は何日ですか
play	ラップ

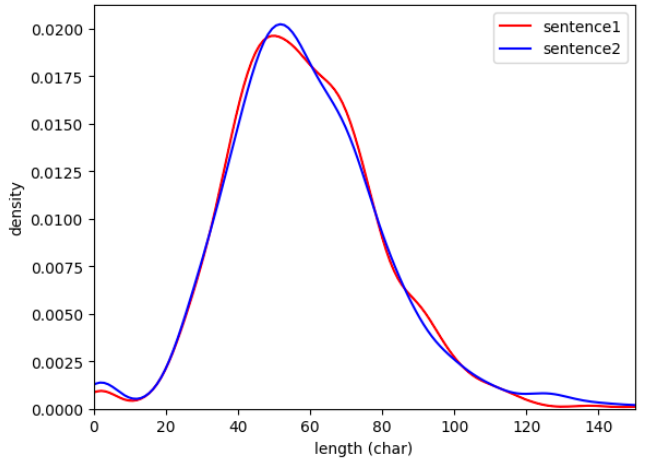
テキストは全てMassiveIntentClassification
文の長さ分布は同じ。



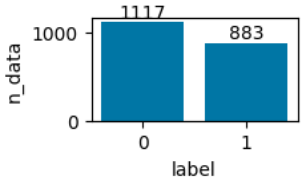
ラベル分布

A-3 Pair Classification – PAWS-X-ja

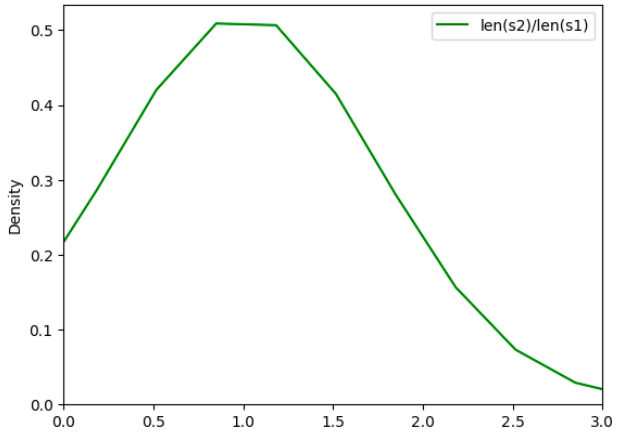
Label	Text
1	力を加えたとき、機構システムの構成要素が変形すると、その要素が弾性位置エネルギーを保持します。
	機械システムの構成要素は、システムに力が加わった際に変形を伴うと、弾性位置エネルギーを蓄える。
1	1999 年、アルバニアの家族はセルビアの家族が帰還する前に村を去りました。
	1999 年のセルビア難民の帰還前に、アルバニア人の家族は村を離れました。
0	2011 年の国勢調査では、住民の 78.8 % がルーマニア人、 17 % がロマ人、2.7 % がハンガリー人、そして 1.4 % がドイツ人でした。
	2011 年の人口調査では、住民の 78.8 % がローマ人、17 % がハンガリー人、2.7 % がルーマニア人、1.4 % がドイツ人でした。
0	この運河は、ヨーロッパとベルギーで運行可能な最古の運河の 1 つである。
	この運河はベルギーだけでなく、ヨーロッパ全体でも最古の運河の 1 つである。



文の長さ分布



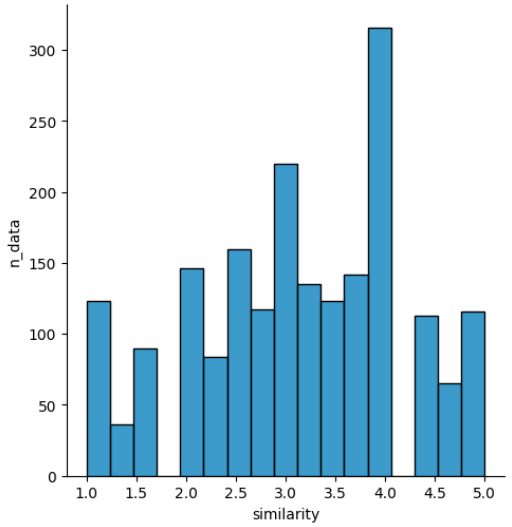
ラベル分布



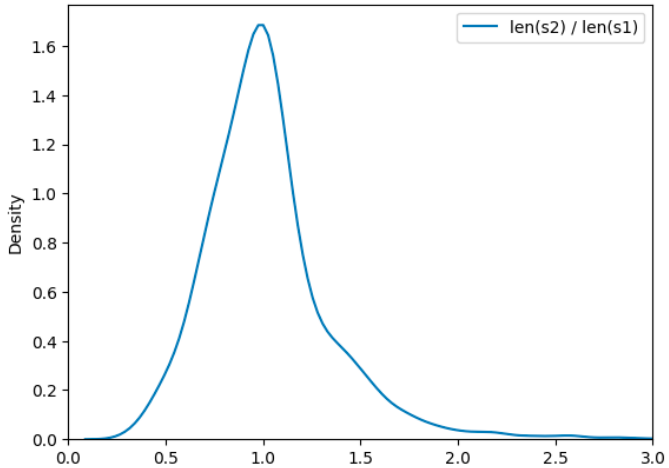
ペア中sentence2とsentence1の長さの相対比

A-4 STS – JSICK

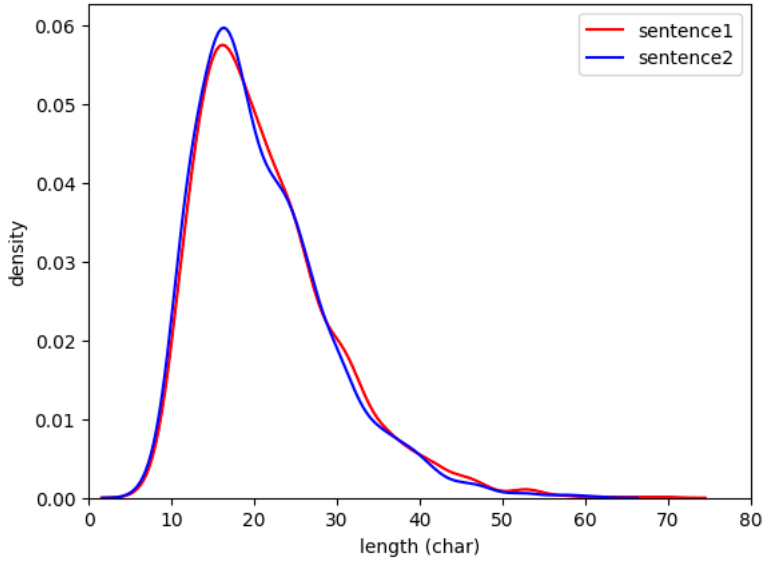
Score	Text
5.0	その人はヌードルを煮ている
	ある人がヌードルを煮ている
3.5	その飛行機は南アフリカのもので、青い空を飛んでいる
	飛行機が空を飛んでいる
2.5	ある人が猫の毛をクシャクシャにしている
	ある人が猫にブラシをかけている
1.7	猿が武術を練習している
	ある人を蹴っている猿は一匹もいない
1.0	男性が荒野を通る小道に沿って歩いている
	男性がタマネギの皮をむいている



スコア分布



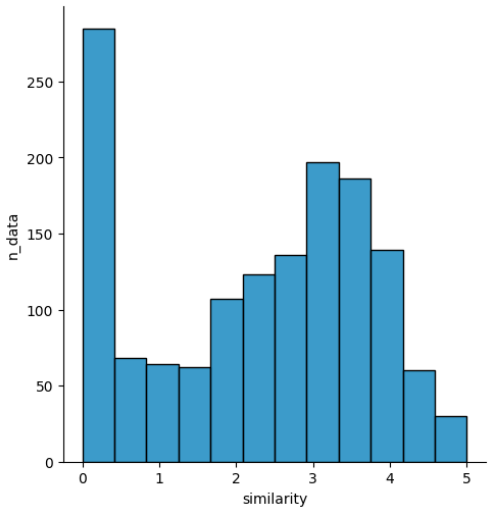
ペア中sentence2とsentence1の長さの相対比



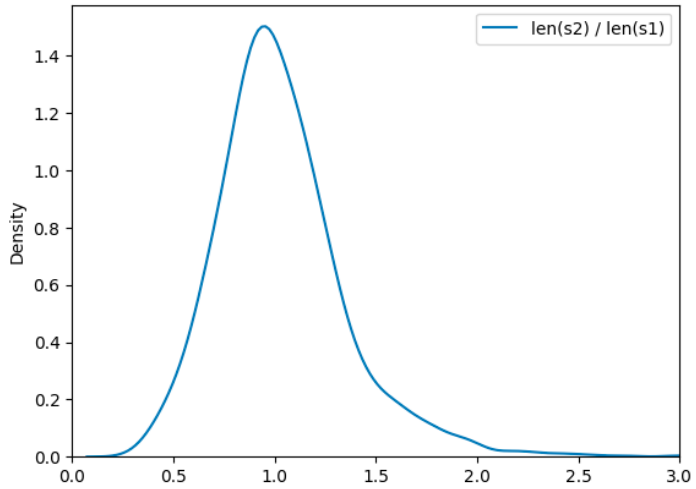
文の長さ分布

A-4 STS – JSTS

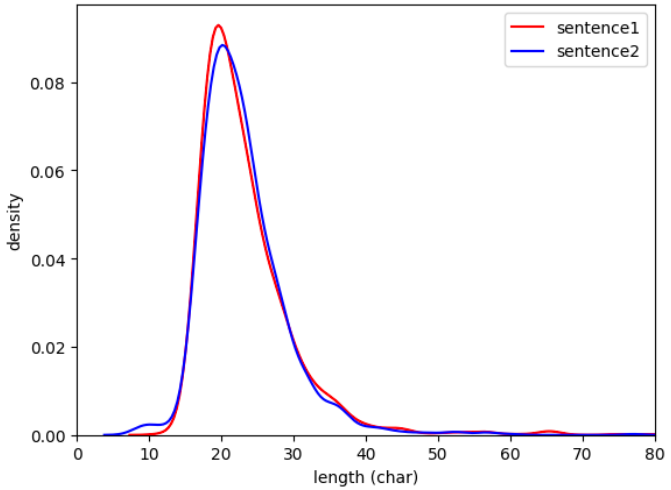
Score	Text
4.4	線路の上を列車が走行しています。
	線路を電車が走行している風景です。
3.6	スノーボーダーが雪の斜面を滑っている。
	雪山を男性がスノーボードで、下っています。
2.6	テーブルの上に、料理と飲み物が置いてあります。
	皿の上に、ケーキがあり横にグラスがあります。
1.4	洗面所の周りはクリーム色や茶色、白で配色されています。
	洗面台の上に瓶や容器が置かれています。
0.6	室内には、小さなテーブルといすがあります。
	まな板の上に包丁と食材が置かれています。



スコア分布



ペア中sentence2とsentence1の長さの相対比



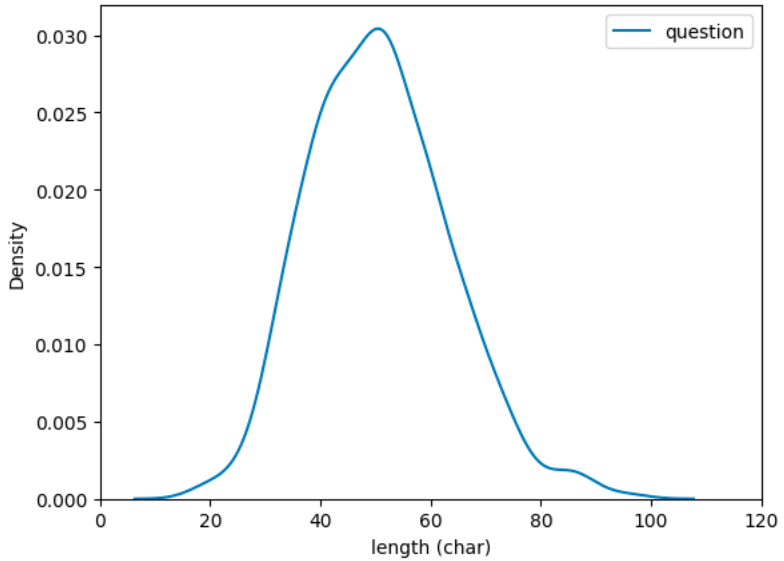
文の長さ分布

A-5 Retrieval – JAQKET

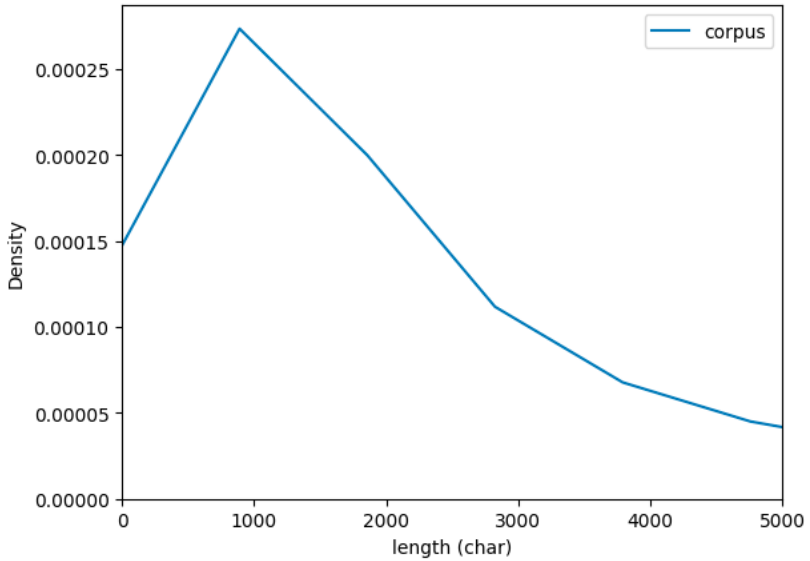
Question	Answer
宮本武蔵と佐々木小次郎が戦った巌流島、平家滅亡の地・壇ノ浦といえば、山口県の何市にある景勝地でしょう?	下関市
日本国憲法第38条1項に規定されている、自分にとって不利益な供述を強要されない権利を何というでしょう?	黙秘権
ある物質1gの温度を1°C上昇させるために必要な熱量のことを何容量というでしょう?	比熱容量
歴代アメリカ大統領のうち、最初に共和党から当選したのは誰でしょう?	エイブラハム・リンカーン
約4.6万平方メートルと広い面積を持つため、よく広大な面積の比喩として「これ何個分」などと例えられる、プロ野球・読売ジャイアンツの本拠地にもなっているドーム球場とはどこでしょう?	東京ドーム

Corpus title	Text
アメリカ合衆国	アメリカ合衆国 United States of America 国の標語：E pluribus unum（1776年 - 現在） （ラテン語:多数からひとつへ） In God We Trust（1956年 - 現在） （英語:我ら神を信ずる） 国歌：The Star-Spangled Banner（英語） 星条旗 アメリカ合衆国（アメリカがっしゅうこく、英語: United States of America）、通称アメリカ、米国（べいこく）は、50の州および連邦区から成る連邦共和国である。 アメリカ本土の48州および同国首都ワシントンD.C.（コロンビア特別区）は、カナダとメキシコの間の北…（総計56673文字）
固有地震	固有地震（こゆうじしん, Characteristic earthquake）とは、ある断層において、ほとんど同じ間隔と規模をもって、周期的に繰り返し発生する地震のこと。固有地震は震源域・規模や地震波形までも類似していることから、相似地震（そうじじしん）という呼び方もある。 また、地震は地殻内でランダムに発生するという考え方に対して、固有地震のように一定の時間的間隔をもってほぼ同じ震源域・規模の地震が発生するという学説を固有地震説と呼ぶ。…（総計3161文字）
不飽和脂肪酸	不飽和脂肪酸（ふほうわしぼうさん、unsaturated fatty acid）とは、1つ以上の不飽和の炭素結合をもつ脂肪酸である。不飽和炭素結合とは炭素分子鎖における炭素同士の不飽和結合、すなわち炭素二重結合または三重結合のことである。天然に見られる不飽和脂肪酸は1つ以上の二重結合を有しており、脂肪中の飽和脂肪酸と置き換わることで、融点や流動性など脂肪の特性に変化を与えている。また、いくつかの不飽和脂肪酸はプロスタグランジン類に代表されるオータコイドの生体内原料として特に重要である…（総計11907文字）

A-5 Retrieval – JAQKET



問題文の長さ分布



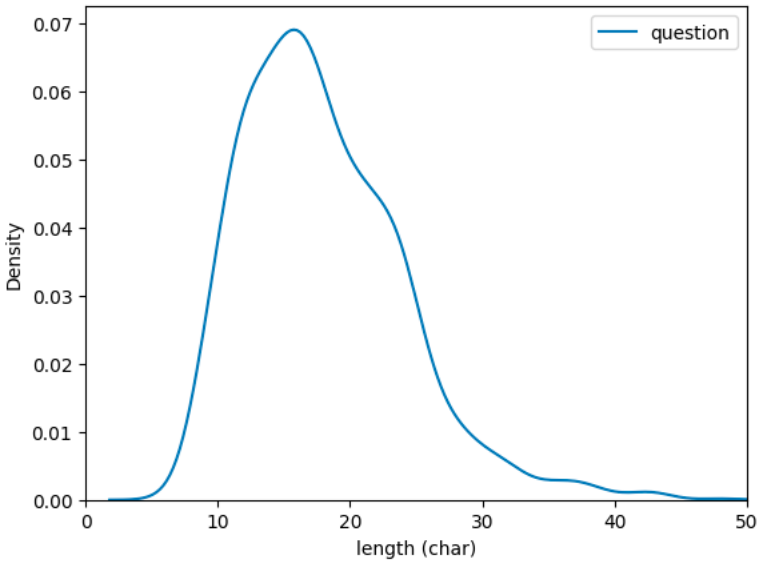
検索対象であるwikipedia文の長さ分布

A-5 Retrieval – Mr.TyDi-ja

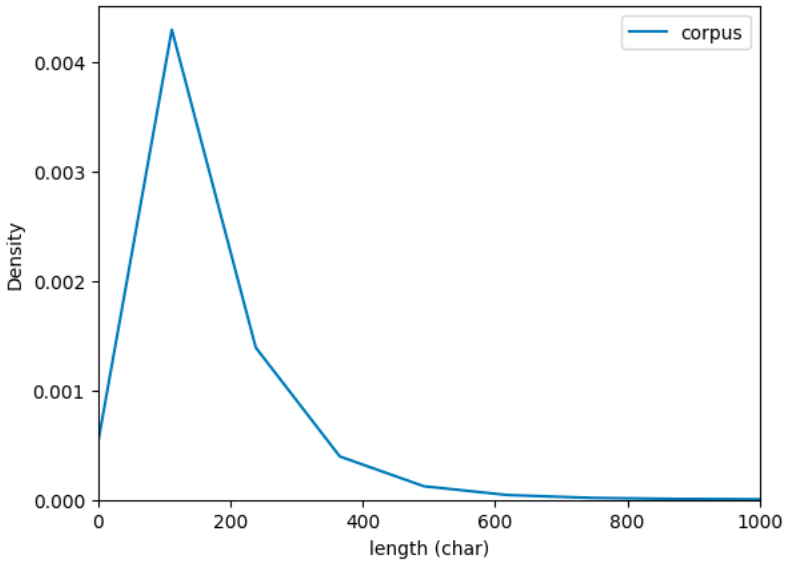
Question	Answer
国際サッカー連盟はいつ設立した？	12306#6
清の咸豊帝は何代目？	63225#0
カール=ハインツ・クラスは銃で人を撃ったことがありますか？	2883370#1, 2883370#22
ＲＮＡを発見したのは誰	3879446#8
阪急宝塚線の西宮北口駅はいつ完成した？	37496#10

Corpus index	Corpus title	Text
5#0	アンパサンド	アンパサンド (&、英語名：) とは並立助詞「…と…」を意味する記号である。ラテン語の の合字で、Trebuchet MSフォントでは、と表示され "et" の合字であることが容易にわかる。ampersa、すなわち "and per se and"、その意味は"and [the symbol which] by itself [is] and"である。
11#0	日本語	'語彙は、古来の大和言葉（和語）のほか、漢語（字音語）、外来語、および、それらの混ざった混種語に分けられる。字音語（漢字の音読みに由来する語の意、一般に「漢語」と称する）は、漢文を通して古代・中世の中国から渡来した語またはそれらから派生した語彙であり、現代の語彙の過半数を占めている。また、「紙（かみ）」「絵/画（ゑ）」など、もともと音であるが和語と認識されているものもある。さらに近代以降には西洋由来の語を中心とする外来語が増大している（「語種」の節参照）。
157379#19	13日の金曜日 (映画)	1985年公開。原題「新たなる始まり」が示す通り、当初の予定では前作で死んだジェイソンに代わる新たな殺人鬼誕生の物語として作られるはずであった。ジェイソンの特徴として語られる「歩いて追いかけてくる」ようになったのは、本作からである。

A-5 Retrieval – Mr.TyDi-ja



問題文の長さ分布

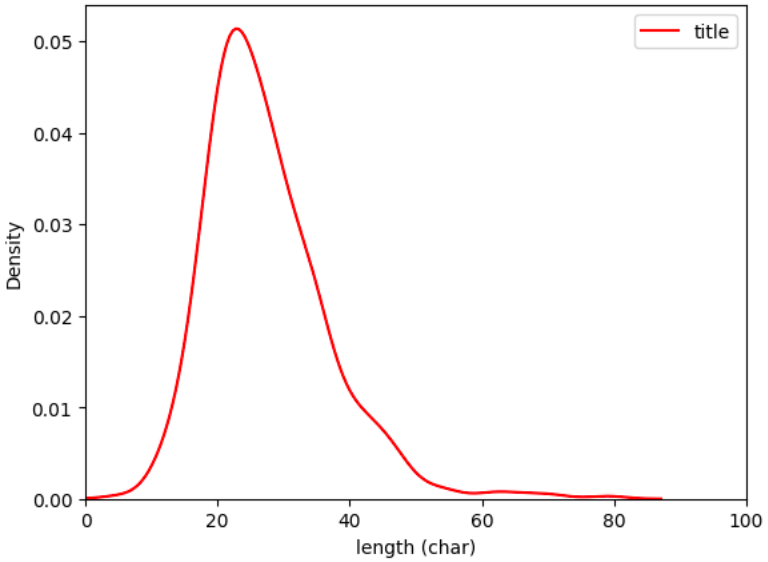


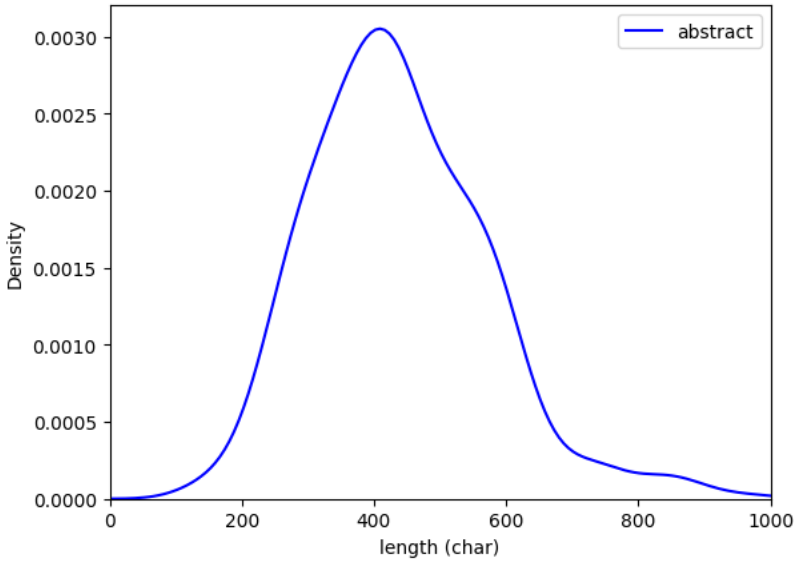
検索対象であるwikipedia文の長さ分布

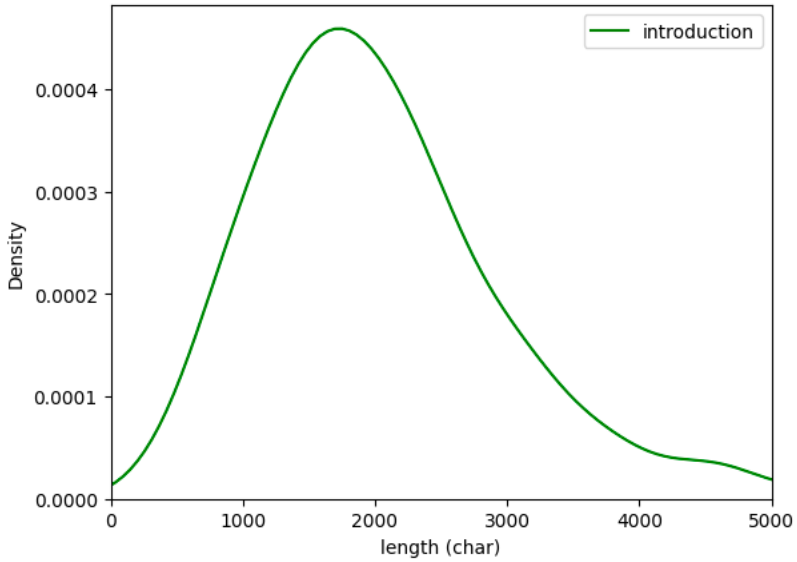
A-5 Retrieval – NLP Journal

Title	Abstract	introduction
古典の総索引からの品詞タグ付きコーパスの作成	全単語の出現箇所を与える総索引は日本の古典の研究の補助として用いられている．品詞タグ付きコーパスはコンピュータを用いた自然語研究の手段として重要である．しかし日本語古典文に関する品詞タグ付きコーパスは公開されていない．そこで総索引を品詞タグ付きコーパスに変換する方法を検討した．	¥¥label{sec:hajime}実際に使用された文例を集めたコーパスは，コンピュータによって検索できる形で準備されることにより，自然言語の研究者にとって便利で重要な資料として利用価値が高まっている．コーパスの種類としては，文例のみを集めた生コーパス（新聞記事など多数がある），文例を単語分けして品詞情報など…
複数の分類スコアを用いたクラス所属確率の推定	文書分類の多くのアプリケーションにおいて，分類器が出力するクラスに確信度すなわちクラス所属確率を付与することは有用で，正確な推定値が必要とされる．これまでに提案された推定方法はいずれも2値分類を想定し，推定したいクラスの分類スコア（分類器が出力するスコア）のみを用いている.しかし…	label{sec:hajime}自然言語処理においては，タグ付けや文書分類をはじめとするさまざまな分類タスクにおいて，分類器が出力するクラスに確信度すなわちクラス所属確率を付与することは有用である．例えば，自動分類システムがより大きなシステムの一部を構成し，…
かな漢字換言を通した日本語語義曖昧性解消の分析	本論文では，日本語語義曖昧性解消に存在する問題点を文中のひらがなを漢字に直すかな漢字換言タスクを通して明らかにする．素性について分散表現と自己相互情報量を組み合わせる手法を考案し実験を行った結果，かな漢字換言においてベースラインに比べ約2ポイント高い精度を得ることができた．…	¥¥label{section:first}語義曖昧性解消はコンピュータの意味理解において重要であり，古くから様々な手法が研究されている自然言語処理における課題の一つである¥¥cite{Navigli:2009:WSD:1459352.1459355,Navigli2012}．語義曖昧性解消の手法には大きく分けて教師あり学習，教師なし学習，半教師あり学習の3つが存在する．教師なし学習を用いるものにはクラスタリングを用いた手法…

A-5 Retrieval – NLP Journal







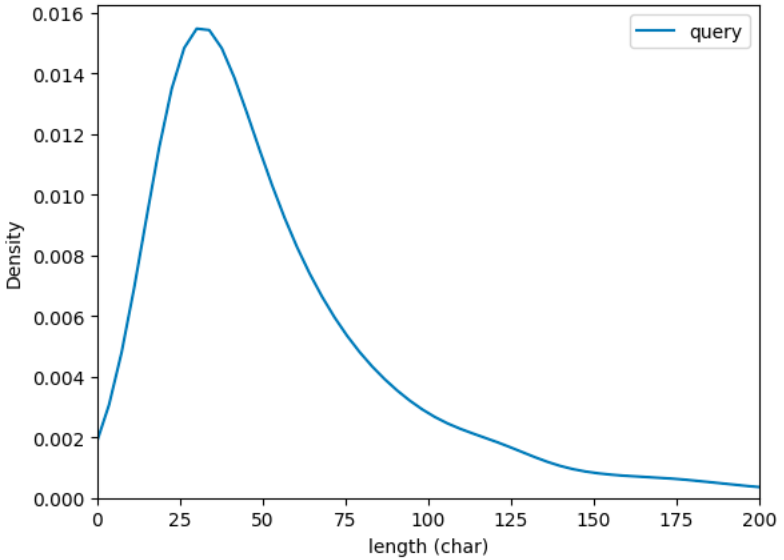
タイトル，概要，イントロの長さ分布

A-5 Retrieval – JaGovFaqs-22k

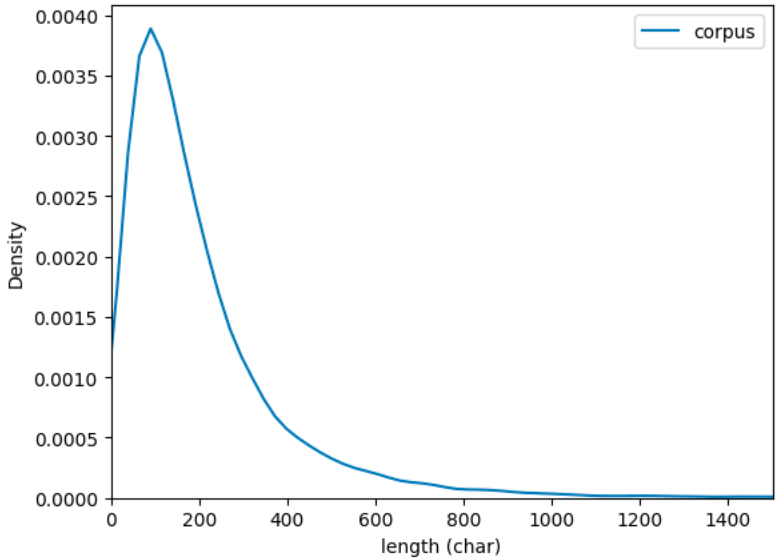
Question	Answer
国鉄清算事業団の長期債務残高はどのくらい残っていますか	15879
加入者が休職することにより資格喪失した後に復職することなく退職する場合、休職期間の２分の１を加入者期間、給付額算定期間に加算することは可能か。	9055
次のとおりグループ通算制度開始前に資本関係が変遷している場合、当社の欠損金額は、法人税法第57条第８項の規定により通算承認の効力が生じた日以後に開始する各事業年度においてないものとされることはないと考えてよろしいでしょうか。	13844
自己資本管理とは何ですか。	21855

Corpus index	Text
17	他院への里帰り分娩の場合も含め、妊婦健診のみの受診はお受けしておりません。
254	１ 防衛省の競争参加資格を取得 ２ 地方防衛局や部隊等のホームページなどで発注情報を入手 ３ 防衛省の電子入札システムに利用者登録し、入札説明書等を入手 ４ 入札に参加 という流れになります。 初めての参加の場合、まずは競争参加資格の取得手続きが必要ですので、本社（本店）の近くの地方防衛局までお問い合わせください。 （Ｉ 競争参加者の資格に関する手続きの問い合わせはこちらを参照
688	排出量の増減の情報については、任意で報告様式第２により提供することができます。個別対策の導入による排出量の削減効果の算定方法については、特段規定されていませんが、本制度の排出量の算定方法を踏まえ、個々の対策の実態に即した合理的な方法により算定する必要があります。

A-5 Retrieval – JaGovFaqs-22k



問題文の長さ分布



検索対象であるFAQ文の長さ分布

Appendix B. 評価指標の詳細

各タスクのMain Metric

Task	Main Metric	Other Metrics
Clustering	V-Measure ↑	Homogeneity↑, Completeness↑
Classification	Macro-F1 ↑	Accuracy↑
Pair Classification	Macro-F1 ↑	Accuracy↑
STS	Spearman correlation ↑	Pearson correlation↑
Retrieval	NDCG@10 ↑	Accuracy@{1,3,5,10} ↑, MRR@10↑

↑　高ければ高いほどよい指標
↓　低ければ低いほどよい指標

Clusteringの評価指標

- V-measure ↑ (main): HomogeneityとCompletenessの調和平均数, 0.0~1.0でクラスタリングの精度をはかる。高ければ高いほどよい。
  - Homogeneity ↑ :同じクラスター内に同じラベルのデータのみが入っているかを測る
  - Completeness ↑ : 同じラベルのデータは全て同じクラスタに入っているかを測る

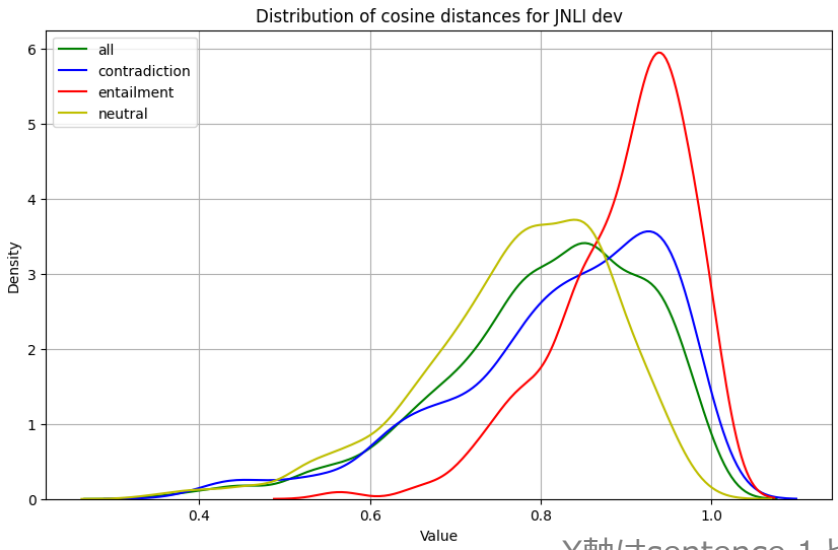
Retrievalの評価指標

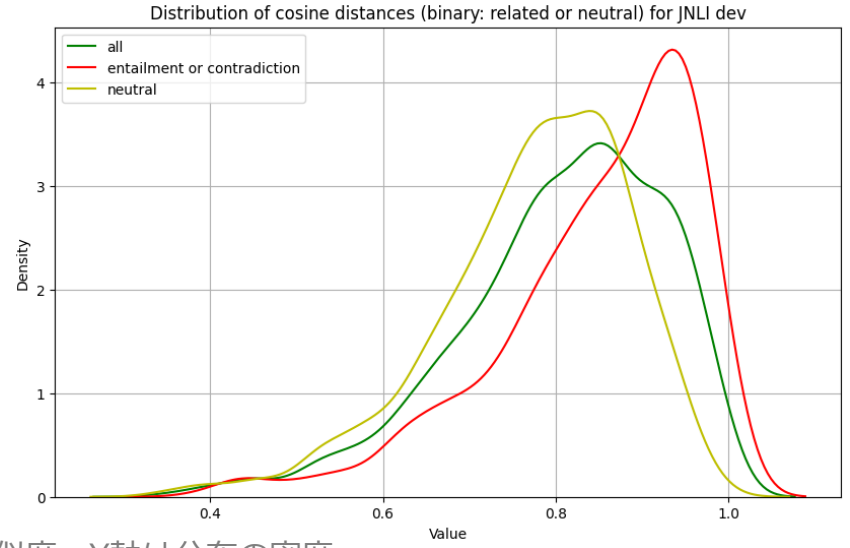
- Accuracy@k ↑ (k = 1, 3, 5, 10):上位k件の検索結果の中で，正解（関連するアイテム）が含まれる割合。
- MRR@k ↑ (k = 10):上位k件の検索結果の中での最初の正解の逆順位の平均。
- NDCG@k ↑ (main) (k = 10):上位k件の検索結果の中で，アイテムの関連度の累積ゲインを割引して正規化したもの。

Appendix C. データセット選択における失敗例

Pair Classification

- JNLI/JSNLI/JSICKは含意関係認識タスク(entailment, contradiction, neutral) であるため，類似度の閾値ベースの分類手法を調べてみた。





X軸はsentence 1とsentence 2のcosine類似度、Y軸は分布の密度。

- 考察1: 含意と矛盾の文ペアが高い類似度を持ち，中立の文ペアが少し低い類似度を持つ。
- 考察2: 類似度の分布には，はっきりした差がない。

埋め込みモデルが意味的類似度を測るため，含意関係を認識する能力を持っていないことがわかる。既存の埋め込みモデルの評価結果に有意な差が得られず，ベンチマークの趣旨と合わないため，外した。

- 本ページの図を導出したため使用した埋め込みモデルはcl-nagoya/sup-simcse-ja-base，データセットはJNLI。
- データセットをJSNLI/JSICKに，モデルを他の埋め込みモデルに切り替えても，分布図の形状がほとんど変わらない。