

研究背景

大規模言語モデル (Large Language Models, LLM) の進化

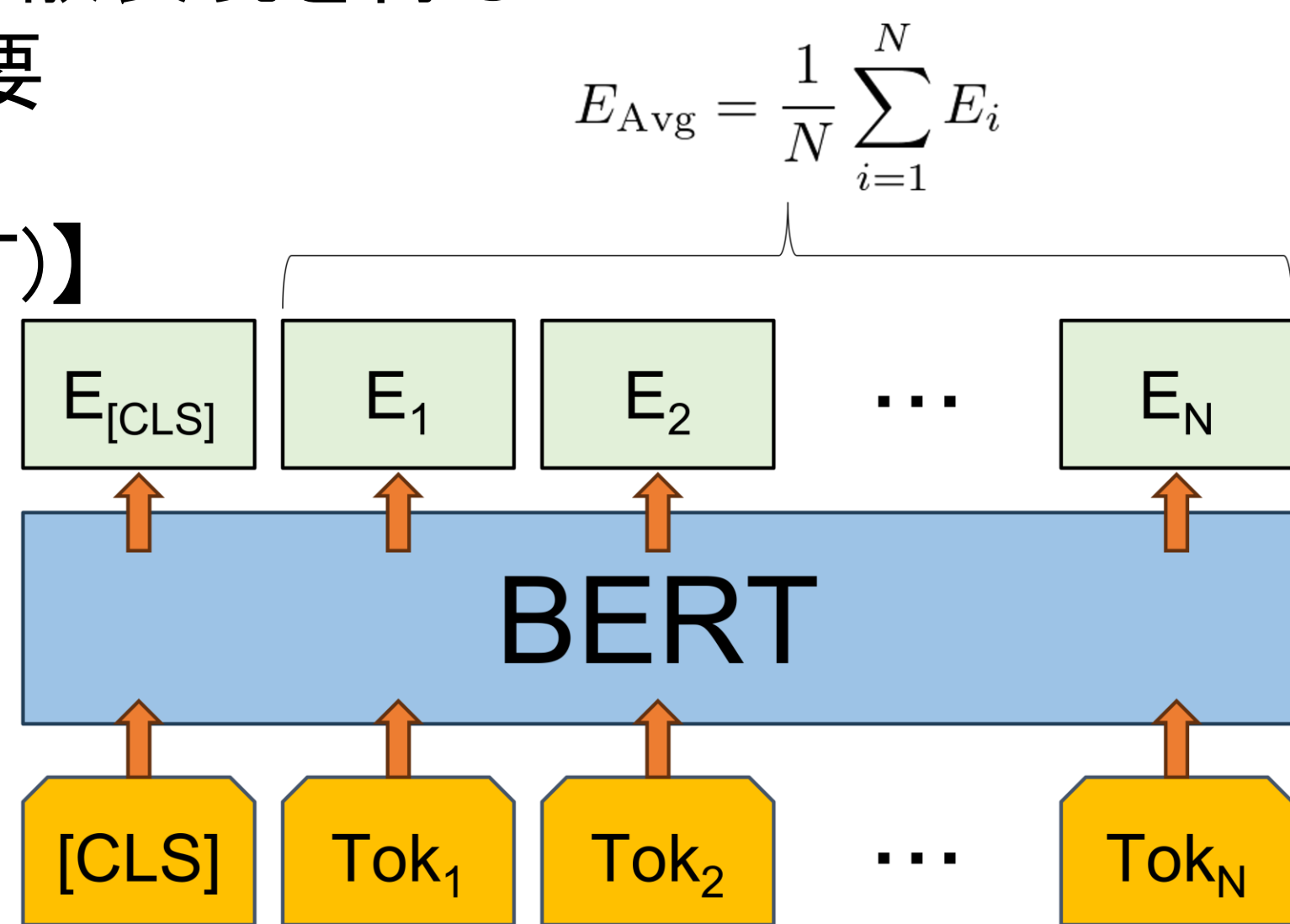
Transformer 構造を持つ BERT や
GPT (Generative Pre-trained Transformer) などを
活用した LLM の商業利用への需要拡大

⇒文章全体の適切な分散表現を得るための
プーリング戦略が重要

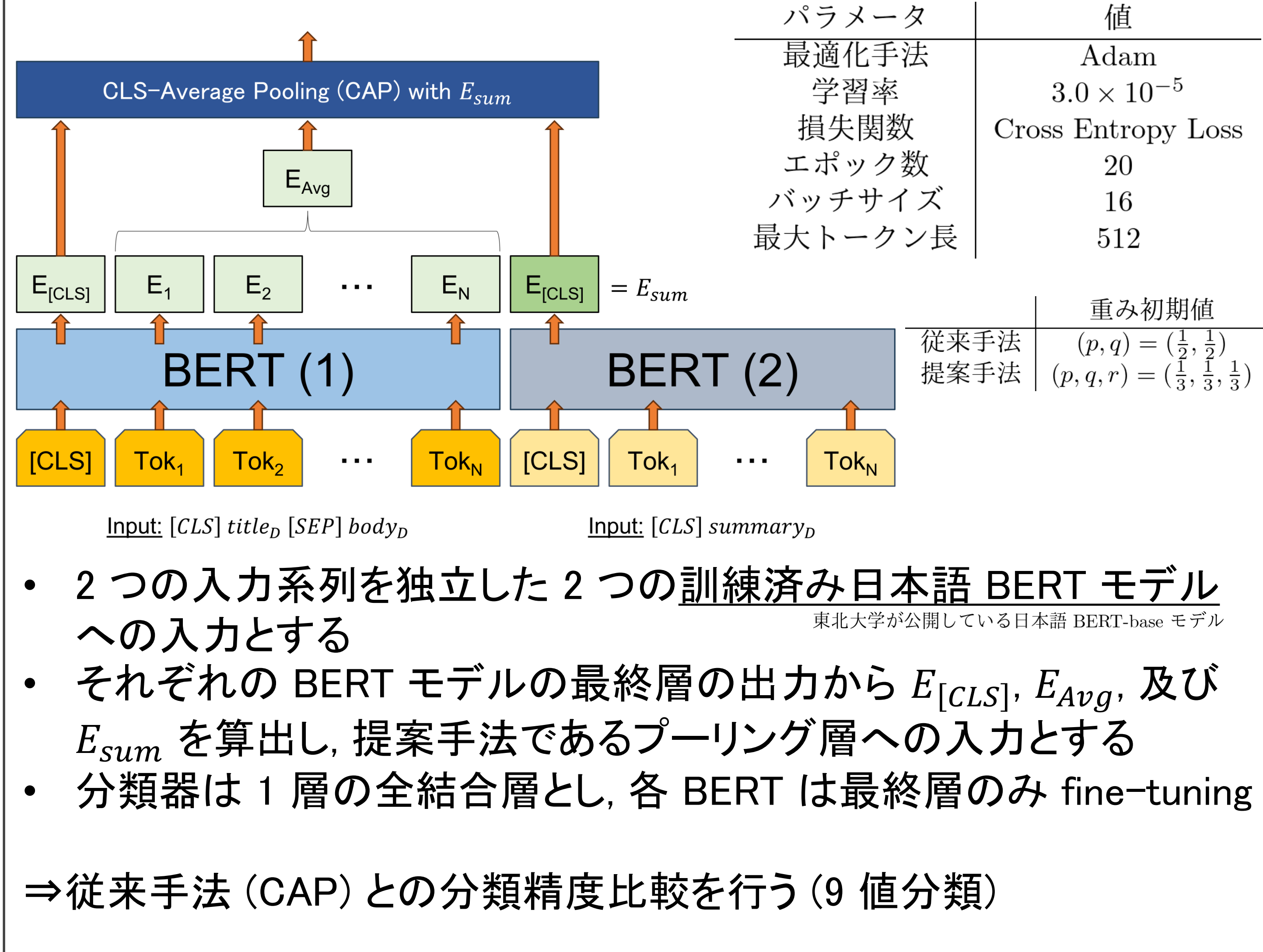
【プーリング例 (BERT)】

1. $E_{[CLS]}$
2. E_{Avg}
3. E_{max}

⇒改良の余地あり



数値実験



従来研究

CLS-Average Pooling (CAP) 層の導入 [Hidenori Yamato, 2023]

学習可能なパラメータ $p, q (\geq 0)$ を用いて
 $E_{[CLS]}$ 及び, E_{Avg} の重み付き和
 $pE_{[CLS]} + qE_{Avg}$ を文の分散表現とする

⇒テキスト分類タスクにおいて, $E_{[CLS]}$, E_{Avg} のみを用いた場合よりも高い性能を発揮

提案手法

CAP 層における要約ベクトル項 E_{sum} の追加

学習可能なパラメータ $p, q, r (\geq 0)$ を用いて
 $E_{[CLS]}$, E_{Avg} , 及び E_{sum} の重み付き和
 $pE_{[CLS]} + qE_{Avg} + rE_{sum}$ を文の分散表現とする

実験設定

【livedoor ニュースコーパスデータセットを用いたテキスト分類】

データセットに含まれる各記事データ D は
記事タイトル $title_D$, 記事本文 $body_D$,
カテゴリーラベル $label_D \in \{0, 1, \dots, 8\}$ を持つ

要約文 $summary_D$ の生成

$Summary_D = \mathcal{L}(prompt_D)$

$prompt_D = \begin{cases} \text{"role": "system", "content": "あなたは優秀な AI アシスタントです。"} \\ \text{"role": "user", "content": "以下は「title_D」というタイトルの"} \\ \text{記事の本文です。内容を要約し、} \\ \text{結果のみを出力してください。"} \\ \text{"role": "user", "content": "body_D"} \end{cases}$

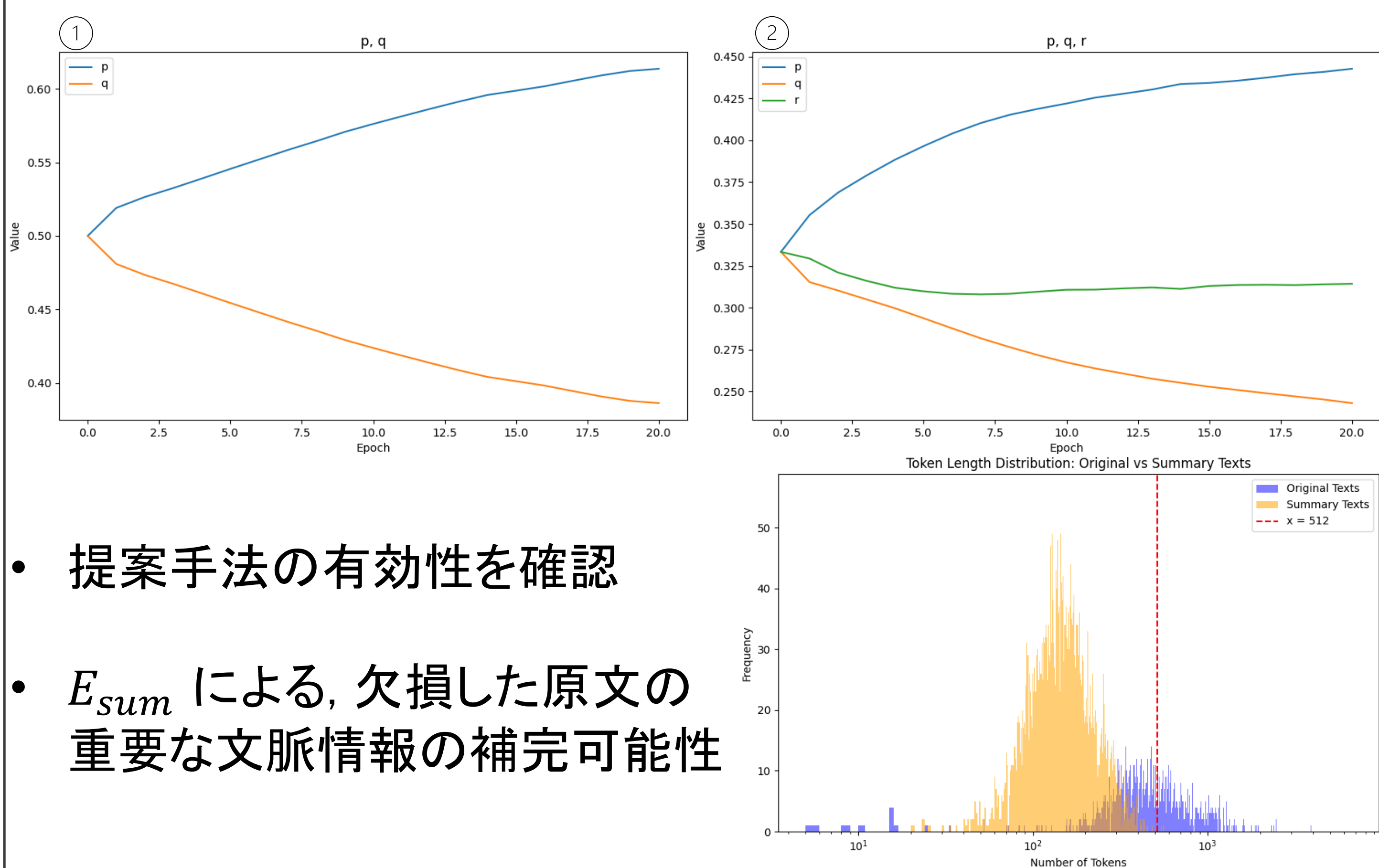
\mathcal{L} : 大規模言語モデル API (e.g., ChatGPT API, PLaMo API)

データ分割

各カテゴリーの記事を訓練データ, 検証データ, テストデータ
として 8:1:1 に分割 (全データ数: 7367 件)

実験結果

	Accuracy	F1 (weighted)
① 従来手法 (CAP)	0.9565	0.9560
② 提案手法 (CAP with E_{sum})	0.9687	0.9683



提案手法の有効性を確認

 E_{sum} による, 欠損した原文の
重要な文脈情報の補完可能性

今後の課題

- 要約文の妥当性, 生成手法改善の検討
- 他のデータセットやタスクを用いた提案手法の有効性の検討
- 最適な学習パラメータやアーキテクチャの探索

	label 0	label 1	label 2	label 3	label 4	label 5	label 6	label 7	label 8	Total
訓練	684	695	690	421	700	645	713	716	631	5895
検証	88	80	88	42	93	100	77	105	63	736
テスト	98	95	86	48	77	97	80	79	76	736

- label 0: 独女通信 (870 件)
- label 1: IT ライフハック (870 件)
- label 2: 家電チャンネル (864 件)
- label 3: livedoor HOMME (511 件)
- label 4: MOVIE ENTER (870 件)
- label 5: Peachy (842 件)
- label 6: エスマックス (870 件)
- label 7: Sports Watch (900 件)
- label 8: トピックニュース (770 件)