kNN 言語モデルは低頻度語の予測に役立つか？

西田悠人 [1]　森下睦 [2]　出口祥之 [1]　上垣外英剛 [1]　渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学,　[2]NTT コミュニケーション科学基礎研究所
{nishida.yuto.nu8, deguchi.hiroyuki.db0, kamigaito.h, taro}@is.naist.jp
makoto.morishita@ntt.com

概要

検索拡張言語モデルの 1 つである kNN 言語モデルは，推論時に任意のテキストデータから構築された大規模なデータストアに直接アクセスして近傍事例を活用することで，文脈を適切に把握し，言語らしさを高精度に予測可能であることが報告されている．データストアの明示的な記憶を利用することによって低頻度語の予測性能が改善することが，kNN 言語モデルの性能向上の要因の 1 つであるという仮説が提唱されてきたが，この仮説の定量的な検証は行われてこなかった．本研究では，低頻度語に対する kNN 言語モデルの振る舞いを定量的に分析し，これまでの仮説に反して，kNN 言語モデルは低頻度語の予測性能の改善に寄与しないことを示した．

1　はじめに

検索拡張言語モデル (retrieval-augmented language models) は，ベースとなる言語モデルと外部のデータストアから検索された近傍事例を組み合わせて予測確率を計算する，言語モデリングの新たなパラダイムであり，言語らしさを高精度に予測可能であることが報告されている [1, 2, 3, 4]．検索拡張言語モデルの 1 つである kNN 言語モデル [3] は，ベースとなる訓練済み言語モデルの各時刻の予測確率を，データストア内の近傍事例から計算された kNN 確率で補間する．データストアは任意のテキストデータの各トークンをベースモデルでベクトル化することで構築され，各時刻の kNN 確率は当該時刻のベースモデルの隠れ表現とデータストア内のベクトルとの距離に基づいて検索された近傍トークンを用いて計算される．上記の過程を経ることで，kNN 言語モデルは推論時にデータストアを介してテキストデータに直接アクセスすることができる．

このように明示的な記憶を利用することで，低い頻度で現れるトークンの文脈をより良く予測できることが kNN 言語モデルの性能向上に寄与する 1 つの要因であるという仮説が提唱され [3]，Khandelwal ら [3] によって仮説の妥当性を示唆する定性的な例が報告されている．しかし，既存の検証において kNN 言語モデルは主にテストデータ全体に対する PPL などの抽象的な指標で評価され，低頻度語のみに焦点を当てた定量的な分析はなされていない．

本研究では，次に示す 3 つのリサーチクエスチョンを起点として，低頻度語の予測における kNN 言語モデルの振る舞いを分析する．

RQ1　低頻度語に対する kNN 確率はベースモデルが割り当てる確率より高いか？

RQ2　近傍事例として検索できない低頻度語はどれくらいの割合で存在するか？

RQ3　データストア内のトークンの分布はそのトークンの頻度によってどのように異なるか？

分析の結果，これまでの仮説に反して，kNN 言語モデルは低頻度語の予測性能の改善に寄与しないことが示された．

2　kNN 言語モデル

kNN 言語モデルは検索拡張言語モデルの 1 つであり，事前に任意のテキストデータからデータストアを構築し，推論時にデータストアから検索した近傍事例を利用して入力テキストの言語らしさを予測する．データストアは，ベースとなる訓練済み言語モデルにテキストデータを入力し，データ中の各トークンをバリュー，対応する隠れ状態ベクトルをキーとして保持することで構成する．

データストア構築　データストアに格納するテキスト集合を $\mathcal{D} = \{\boldsymbol{d}^1, \cdots, \boldsymbol{d}^N\}$ とする．ただし，$\boldsymbol{d}^n = (d_1^n, \cdots, d_{|\boldsymbol{d}^n|}^n) \in \mathcal{V}^{|\boldsymbol{d}^n|}$ であり，$\mathcal{V}$ は語彙である．時刻 t において，言語モデルは文脈系列として $\boldsymbol{x} = (d_1^1, \cdots, d_{|\boldsymbol{d}^1|}^1, \cdots, d_1^N, \cdots, d_{|\boldsymbol{d}^N|}^N)$ の部分列 $\boldsymbol{x}_{<t}$ を受け取り，目的トークン x_t を予測する．文脈系列 $\boldsymbol{x}_{<t}$ に対して言語モデルの D 次元隠れ状態ベクトル

を計算する関数を $\boldsymbol{f}: \mathcal{V}^{t-1} \to \mathbb{R}^D$ とすると，データストア $\mathcal{S} \subseteq \mathbb{R}^D \times \mathcal{V}$ は式 (1) で表される．

$$\mathcal{S} = \{(\boldsymbol{f}(\boldsymbol{x}_{<t}), x_t)\}_{t=1}^{|\boldsymbol{x}|} \tag{1}$$

ここで，データストア $\mathcal{S}$ の大きさは入力データのトークン数と同一となる．

kNN 確率の計算　テストデータを入力した際の時刻 t におけるベースモデルの隠れ状態ベクトル $\boldsymbol{q}_t = \boldsymbol{f}(\boldsymbol{x}_{<t})$ を，目的トークン x_t に対応するクエリとして，データストア $\mathcal{S}$ から k 近傍 $\mathcal{N} \subset \mathcal{S}$ を検索する．クエリ $\boldsymbol{q}_t$ と k 近傍点の距離から kNN 確率 p_{kNN} を式 (2) のように計算する．なお，$\mathrm{dist}(\cdot,\cdot)$ は距離関数，τ は softmax 関数の温度パラメータである[1)]．

$$p_{\mathrm{kNN}}(x_t|\boldsymbol{x}_{<t}) \propto \sum_{(\boldsymbol{k},v)\in\mathcal{N}} \mathbb{1}_{x_t=v} \exp\left(\frac{-\mathrm{dist}(\boldsymbol{k},\boldsymbol{q}_t)}{\tau}\right) \tag{2}$$

kNN 確率は，クエリからの距離が大きいトークンに対しては低く，距離の小さいトークンに対しては高く割り当てられる[2)]．なお，k 近傍 $\mathcal{N}$ に含まれないトークンの kNN 確率は 0 となる．

予測確率の補間　文脈 $\boldsymbol{x}_{<t}$ が与えられたとき，目的トークン x_t の予測確率 p は，kNN 確率 p_{kNN} とベースモデルの予測確率 p_{LM} の線形補間により，式 (3) のように計算する．ここで，λ は kNN 確率の重みを決定するハイパーパラメータである．

$$p(x_t|\boldsymbol{x}_{<t}) = \lambda p_{\mathrm{kNN}}(x_t|\boldsymbol{x}_{<t}) + (1-\lambda) p_{\mathrm{LM}}(x_t|\boldsymbol{x}_{<t}) \tag{3}$$

kNN 言語モデルは，推論時にデータストアを介してテキストデータに直接アクセスすることができる．そのため，低い頻度で現れるトークンをより良く予測でき，このことが kNN 言語モデルの性能向上に寄与する 1 つの要因であると指摘されている [3]．

3　低頻度語の予測の分析

kNN 言語モデルは明示的な記憶であるデータストアを利用することで，低頻度語の予測性能が向上することが指摘されている [3]．本研究では，この仮説の定量的な検証を目的とし，低頻度語を対象に分析を行う．なお，本稿において低頻度語は，データストア内での頻度が通常の kNN 言語モデルで使用される近傍数 $k = 16$ よりも低いトークンを指す．

1)　オリジナルの kNN 言語モデルでは温度パラメータ τ は導入されていないが，本稿では，kNN 確率の分布の滑らかさの制御のために導入する．温度パラメータは kNN 言語モデルの変種である kNN 機械翻訳 [5] でも利用されている．

2)　温度パラメータ τ が大きいほど，距離による確率の割り当て方に差がなくなる．極限 $\tau \to \infty$ を考えると，あるトークンに対する kNN 確率はその距離に依存せず，k 近傍トークン全体における当該トークンの相対頻度と等しくなる．

表 1　各データの文数

	オリジナル	再分割
訓練データ	1,165,029	1,157,595
開発データ	2,461	3,360
テストデータ	2,891	4,074

表 2　低頻度語の出現数: テスト・開発ともに，再分割後の低頻度語の数はオリジナルデータを大きく上回る．

頻度	テストデータ		開発データ	
	オリジナル	再分割	オリジナル	再分割
0	11	151	1	190
1 ~ 5	31	333	13	349
6 ~ 10	24	439	8	241
11 ~ 15	16	264	19	335

3.1　実験設定

ベースモデル　ベースとなる言語モデルとして gpt2-medium [6] を用いた．

データセット　wikitext-103 [7] を利用した．低頻度語の振る舞いを分析するために，wikitext-103 の訓練データを次の手順で再分割し，低頻度語が多く含まれるような評価データを作成した．

1. ベースモデルのサブワードトークナイザを用いて訓練データをトークン化する
2. $n = 1, 2, 3, 4$ にわたって，データ内の頻度が 1 である n-gram の占める割合が大きい順に，オリジナルデータと同程度の分量になるまで文書を重複なく抽出する
3. 抽出した文書をランダムに開発・テストデータに割り振り，残りの文書を訓練データとする

オリジナルのデータと再分割後のデータの文数および低頻度語の出現数をそれぞれ表 1 と表 2 に示す．

kNN 言語モデル　実装には knn-transformers [8] を用いた．kNN 確率の補間係数 λ は，[3] に倣って 0.25 とした．データストアのキーおよびクエリには最終層の feed forward network (FFN) の入力部分の隠れ表現を用いた．式 (2) の距離関数には平方ユークリッド距離を用いた．データストアはオリジナルおよび再分割後の訓練データを用いてそれぞれ構築した．

3.2　事前実験: 全体の PPL

kNN 言語モデルの分析を行う前に，低頻度語を多く含むデータでの kNN 言語モデルの有効性を調査するために，ベースモデルと kNN 言語モデルの性能をオリジナルおよび再分割後のデータを用いて比較する．モデルの性能はテストデータの PPL によって

表 3　テスト性能 (PPL) の比較

	オリジナル	再分割後
ベースモデル	18.25	21.90
kNN 言語モデル	12.90	19.60

評価する．kNN 言語モデルの近傍数 k と温度 τ は，開発データの PPL が最低となるものを選択した[3]．

実験結果を表 3 に示す．オリジナルと再分割後のデータの両方で kNN 言語モデルによって性能が改善しており，低頻度語を多く含むデータにおいても kNN 言語モデルが有効であることが示された．しかし，再分割後のデータにおける改善の幅はオリジナルのデータと比較して小さく，低頻度語が多く含まれていることによってその有効性が制限されていることが示唆された[4]．

3.3　kNN 確率の分析

kNN 言語モデルによって低頻度語の予測性能が向上するという仮説が正しければ，入力系列中で観測された低頻度語における kNN 確率はベースモデルの予測確率（LM 確率）を上回るはずである．本節では，目的トークンが低頻度語の場合と高頻度語の場合の kNN 確率をそれぞれ LM 確率と比較することで，kNN 言語モデルの振る舞いを分析する．

図 1 に，頻度上位 100 トークン（高頻度語）と頻度 0 を除く低頻度語に対する kNN 確率と LM 確率の期待値の比較を示す．図より，近傍数 k の値に依らず，低頻度語に対する kNN 確率は LM 確率より非常に小さく，kNN 確率と LM 確率の補間によって低頻度語の予測性能は悪化することが明らかとなった．また，高頻度語に対する kNN 確率は LM 確率より大きい傾向が観察された．このことから，kNN 言語モデルによる性能向上は低頻度語の予測性能の向上に依るものではなく，むしろ高頻度語の予測性能の向上に依るものであることが示唆された．低頻度語に関する更なる分析として，付録 B.2 ではトークン頻度および温度ごとの kNN 確率と LM 確率の期待値を比較し，低頻度語に対する kNN 確率はトークンの頻度に関わらず常に LM 確率を下回り，温度 τ を大きくするとより顕著に kNN 確率が小さくなることを観察した．

3)　ハイパーパラメータの詳細は付録 A に記載した．以降の実験は，この設定を踏襲する．

4)　付録 B.1 では，kNN 言語モデルの性能とハイパーパラメータの関係を調査し，オリジナルと再分割後のデータの双方で，温度 τ が大きいほど性能が悪化し，適切な温度のもとでは近傍数 k は大きいほど性能が向上する傾向を観察した．

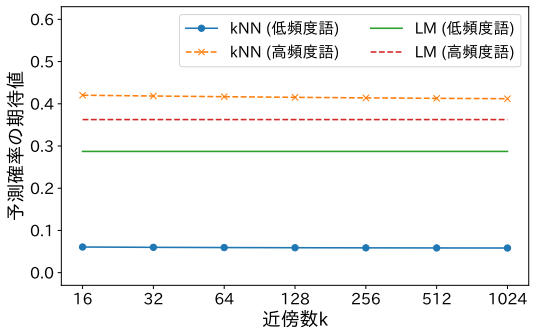

図 1　kNN 確率と LM 確率の比較: 低頻度語に対する kNN 確率は LM 確率より小さく，高頻度語は逆の傾向を示す．

3.4　kNN ヒット率の分析

低頻度語に対する kNN 確率はなぜ低いのだろうか？本節では，低頻度語は近傍探索でヒットしづらいためであるという仮説を立て，分析を行う．

仮説の検証のため，クエリ近傍 $\mathcal{N}$ に目的トークンが 1 件以上含まれる割合（kNN ヒット率）を調査する．図 2 に，高頻度語と頻度 0 を除く低頻度語の kNN ヒット率を示す．近傍数 k を大きくすると，頻度に関わらず kNN ヒット率が高くなる傾向が見られた．高頻度語に対する kNN ヒット率は近傍数 $k = 16$ のもとで 82.5%，$k = 1024$ で 99.2%であり，非常に高い割合でヒットする．対照的に，低頻度語では近傍数 $k = 16$ でわずか 15.5%であり，$k = 1024$ のもとであっても 38.1%と非常に低い[5]．このように，低頻度語に対しては，過半数のケースで目的トークンが近傍探索で 1 件もヒットせず，kNN 確率が 0 となる．このことが，前節で述べた低頻度語の kNN 確率の期待値の低さの要因であると考えられる．

また，付録 B.4 では近傍数 k を大きくするほど近傍トークン集合に占める目的トークンの割合（kNN 相対頻度）が低くなることを観察した．すなわち，kNN ヒット率と kNN 相対頻度にはトレードオフ関係があり，このことが図 1 で観察された近傍数 k を大きくしても kNN 確率の期待値は向上しないという傾向を引き起こしていると考えられる．

3.5　データストアの分布の調査

なぜ低頻度語は近傍探索でヒットしづらいのだろうか？本節では，低頻度語はデータストア上でまばらに分布していることがその要因であるという仮説を立て，この仮説を検証する．

5)　付録 B.3 では，低頻度語に対する kNN ヒット率とトークン頻度の関係を調査し，トークンの頻度が大きいほどヒット率が高くなる傾向を観察した．

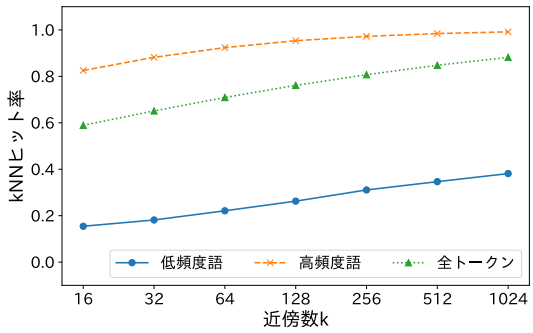

図 2 kNN ヒット率: 低頻度語に対する kNN ヒット率は 15.4%~38.1%と低く，高頻度語では 82.5%~99.2%と高い．

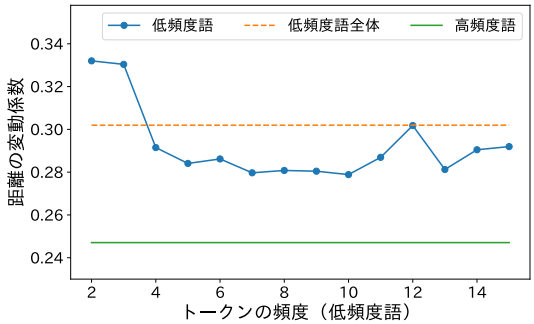

図 3 各トークンのセントロイドまでのユークリッド距離の変動係数: 低頻度語は高頻度語に比べて変動係数が大きく，よりまばらに分布していることが分かる．

仮説の検証のため，データストア上の各トークンのばらつきを調査する．あるトークンをバリューに持つデータ点のキーのセントロイドから各データ点のキーまでのユークリッド距離を測り，その距離の変動係数[6]によってそのトークンのばらつきを定量化する．図 3 に，低頻度語と高頻度語の変動係数を示す．低頻度語は高頻度語に比べて変動係数が大きく，よりばらつきが大きいことが示唆された．よって，低頻度語の kNN ヒット率が低いことは，データストア上で低頻度語はまばらに分布しており，近傍探索が困難であることに起因すると考えられる．

3.6 本章のまとめ: なぜ kNN 言語モデルは低頻度語の予測に役立たないのか？

本章では，1 章で示した 3 つのリサーチクエスチョンを起点として，低頻度語に対する kNN 言語モデルの振る舞いを分析した．分析によって，3 つのリサーチクエスチョンに対してそれぞれ以下のことが明らかとなった．

A1 低頻度語に対する kNN 確率はベースモデルの予測確率より低く，高頻度語では低頻度語の逆の傾向を示す

A2 小さい近傍数 k のもとでは低頻度語のほとんどは近傍探索でヒットせず，大きい近傍数 k のもとでは近傍集合に占める低頻度語の割合が低い

A3 低頻度語は高頻度語と比較してデータストア上の分布のばらつきが大きい

以上の理由により，kNN 言語モデルは低頻度語の予測性能の向上に寄与せず，低頻度語を多く含むデータでは kNN 言語モデルの有効性は限定的であることが示された．

4 関連研究

kNN 言語モデル [3] は検索拡張言語モデルの 1 つであり，事前に任意のテキストデータからデータストアを構築し，推論時にデータストアから近傍事例を検索して利用する．kNN 言語モデルは明示的な記憶であるデータストアを利用することで，低頻度語の予測性能が向上することが指摘されている [3] が，この仮説の定量的な検証は行われてこなかった．

Xu ら [9] は，kNN 言語モデルに対して様々な ablation study を行うことで，kNN 確率の計算に FFN の出力とは異なる表現を用いることによるアンサンブルの効果と近似近傍探索による正則化の効果が性能向上の要因であることを発見した．ただし，低頻度語に対する kNN 言語モデルの振る舞いは分析しておらず，上記の仮説の直接的な検証は行われていない．kNN 言語モデルやその派生である kNN 機械翻訳 [5] の性能を向上させるために，近傍数 k や補間係数 λ を適応的に決定する手法 [10, 11, 12] や 2 つの異なるデータストアを用いて距離を再計算する手法 [13] が提案されているが，これらの手法は本研究で明らかとなった低頻度語の予測に対する負の影響を明示的に解決するものではない．

5 おわりに

kNN 言語モデルは低頻度語の予測性能を改善することが指摘されてきたが，この仮説の定量的な検証は行われてこなかった．本研究では，低頻度語に対する kNN 言語モデルの振る舞いを分析することで上記の仮説を検証した．実験により，これまでの仮説に反して，kNN 言語モデルは低頻度語の予測性能の改善に寄与せず，むしろ高頻度語の予測性能の改善が kNN 言語モデルの成功の要因の 1 つであることを明らかにした．低頻度語の予測性能を改善するには，データストアの構築やクエリの計算の方法を工夫する必要があり，これは残された課題である．

6) 標準偏差を平均で除した値で，分布のばらつきを示す．

謝辞

本研究の一部は JSPS 科研費 JP22J11279, JP22KJ2286 の助成を受けたものです.

参考文献

[1] Edouard Grave, Moustapha Cisse, and Armand Joulin. Unbounded cache model for online language modeling with open vocabulary. **arXiv preprint arXiv:1711.02604**, 2017.

[2] Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. Generating sentences by editing prototypes. **Transactions of the Association for Computational Linguistics**, Vol. 6, pp. 437–450, 2018.

[3] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In **International Conference on Learning Representations**, 2020.

[4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In **International conference on machine learning**, pp. 2206–2240. Proceedings of Machine Learning Research, 2022.

[5] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In **International Conference on Learning Representations**, 2021.

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[7] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. **arXiv preprint arXiv:1609.07843**, 2016.

[8] Uri Alon, Frank Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. Neuro-symbolic language modeling with automaton-augmented retrieval. In **International Conference on Machine Learning**, pp. 468–485. Proceedings of Machine Learning Research, 2022.

[9] Frank F. Xu, Uri Alon, and Graham Neubig. Why do nearest neighbor language models work? **arXiv preprint arXiv:2301.02828**, 2023.

[10] Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. Adaptive nearest neighbor machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 368–374, Online, August 2021. Association for Computational Linguistics.

[11] Qingnan Jiang, Mingxuan Wang, Jun Cao, Shanbo Cheng, Shujian Huang, and Lei Li. Learning kernel-smoothed machine translation with retrieved examples. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7280–7290, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[12] Hui Jiang, Ziyao Lu, Fandong Meng, Chulun Zhou, Jie Zhou, Degen Huang, and Jinsong Su. Towards robust k-nearest-neighbor machine translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 5468–5477, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[13] Rishabh Bhardwaj, George Polovets, and Monica Sunkara. Adaptation approaches for nearest neighbor language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 1135–1146, Toronto, Canada, July 2023. Association for Computational Linguistics.

A　実験設定の詳細

3 章における実験において，kNN 言語モデルの補間係数 λ は [3] に倣って 0.25 とした．また，近傍数 k と温度 τ は，$k \in \{16, 32, 64, 128, 256, 512, 1024\}$ および $\tau \in \{1, 10, 100, 1000\}$ の範囲でグリッドサーチを行い，オリジナルおよび再分割後の双方で開発データの PPL が最低となる $k = 1024, \tau = 1$ を選択した．

B　追加の分析

B.1　kNN 言語モデルのチューニング

3.2 節での分析に加えて，kNN 言語モデルの近傍数 k や温度 τ による性能への影響を調査した．実験結果を図 4 に示す．温度 τ は値が大きいほど性能が悪化する傾向が見られた．また，最適な温度 $\tau = 1$ のもとでは，[3] で報告された傾向と同じく，近傍数 k は大きいほど性能が向上することが観察された．

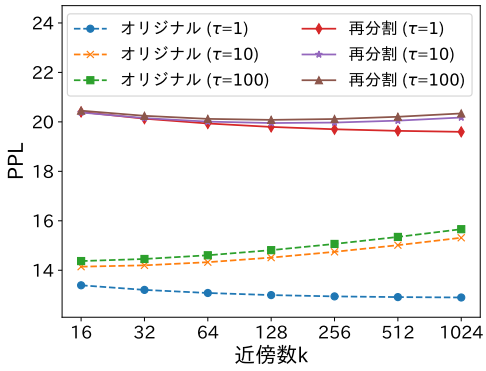

図 4　kNN 言語モデルの近傍数 k および温度 τ と性能の関係

B.2　トークン頻度および温度と kNN 確率

3.3 節での分析に加えて，トークン頻度および温度による低頻度語の kNN 確率への影響を調査した．実験結果を図 5 に示す．全ての頻度において，kNN 確率は LM 確率を常に下回ることが観察された．また，温度 τ が大きいときにはより顕著に低頻度語の kNN 確率が低くなることから，温度パラメータを大きくすることは低頻度語の予測確率の向上に寄与しないことが示唆された．

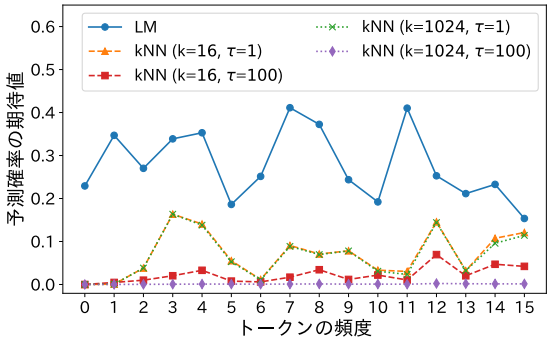

図 5　低頻度語に対する kNN 確率

B.3　トークン頻度と kNN ヒット率

3.4 節の分析に加えて，低頻度語に対する kNN ヒット率とトークン頻度の関係を調査した．実験結果を図 6 に示す．トークンの頻度が大きいほどヒット率が向上する傾向を観察した．

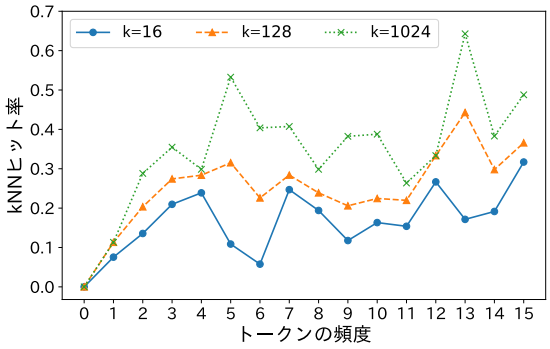

図 6　低頻度語に対する kNN ヒット率

B.4　kNN 相対頻度

3.4 節の分析に加えて，近傍数 k と近傍トークン集合に占める目的トークンの割合（kNN 相対頻度）の関係を調査した．実験結果を図 7 に示す．低頻度語および高頻度語，全トークンのすべてにおいて，近傍数 k を大きくすると，kNN 相対頻度が低下する傾向が観察された．

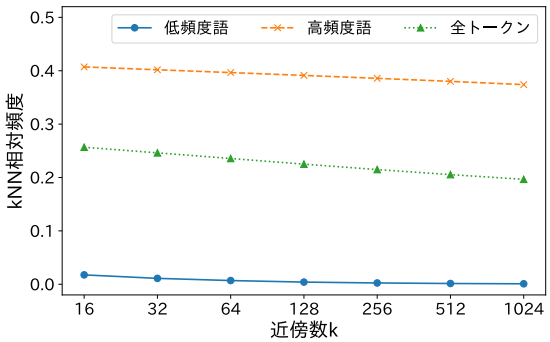

図 7　kNN 相対頻度