低資源言語音声認識における *n*-gram 言語モデルの有効性と無効性

田口智大 [1]
Department of Computer Science and Engineering
University of Notre Dame
ctaguchi@nd.edu

概要

自己教師あり学習によって多言語の音声をエンコードするように訓練された音声モデルは，対象言語に特化したコネクショニスト時系列分類（CTC）デコーダ層を訓練することで，数時間程度，あるいはそれ以下の音声データ量での低資源音声認識をも可能にした．以来，この手法を用いた低資源音声認識の成功事例は数多く報告されているものの，どのような要因が低資源言語音声認識器の性能の向上にどれだけ寄与するのかは不明のままである．そこで，本研究では CTC 層のデコーディングに着目し，21 の低資源言語を対象に，貪欲法，ビームサーチ，そしてビームサーチと *n*-gram 言語モデルの混合の各手法の有効性を比較する．実験と検証の結果，低資源言語であっても *n*-gram 言語モデルは精度の向上に概ね有効性を示すが，形態論的に複雑な語形成を有する言語に対してはむしろ精度を低下させることが示された．

1 はじめに

注意機構を用いた Transformer [1] の登場と計算資源の向上により，多言語音声認識の分野はは目覚ましく発展した．特に，HuBERT [2] や Wav2Vec2 [3] といった自己教師あり学習（self-supervised learning）を用いたモデルは，大量のラベルなし音声データを事前訓練データとしてその音声特徴を学習するように訓練することで，抽象的な文脈化埋め込み表現（contextualized embedding）を得ることができる．この手法を多言語データに応用することで，多様な音声言語の一般的な特徴を捉えた埋め込み表現を得ることが可能となり，少量の教師ありデータを用いてコネクショニスト時系列分類（CTC）を行う層を付加するファイン・チューニングを行うことで，様々な言語の音声認識モデルを訓練することができる．

ファイン・チューニングでは，事前学習済みモデルによって得られたフレーム単位の音声の文脈化埋め込み表現を語彙数（一般的には文字素の数）分の次元数を持ったベクトルに写像する層を追加し，正解ラベルが得られるような文字列の予測の確率を最大化するように，すなわち CTC 損失を最小化するように訓練を行う．

推論時には，モデルの出力のロジットをもとにデコーディングを行うが，ここでいくつかの手法が存在する．一つ目は貪欲法（greedy decoding）であり，フレーム毎のロジットのうち最も確率の高い文字を選択する．二つ目はビーム探索（beam search decoding）で，上位 β 件（ビーム幅）の候補のみを保持しながら幅優先探索を行ってデコードする手法である．これにより，貪欲法で選ばれるような局所的な最適解にとらわれず，より文脈を考慮した出力が可能になる．三つ目は *n*-gram 言語モデルを利用したビーム探索である．これは，ビーム探索に語単位の *n*-gram 言語モデルの確率を加えることで，さらに広い文脈を考慮したデコーディングを行う．

実際に，貪欲法よりも，ビーム探索や言語モデルを組み合わせたデコーディングがより正確に書き起こしを予測できることは経験的に知られている [4, 5]．近年の研究では，*n*-gram 言語モデルの統合が低資源言語にもある程度有効であることが示されている [6] が，その有効性は言語によって異なる上に，どのような状況下で言語モデルが有効であるのかは示されていない．そこで，本研究では，21 の低資源言語を対象として，それぞれに音声認識モデルを訓練し，貪欲法，ビーム探索，*n*-gram 言語モデルを用いたデコーディングの精度を比較する[1)]．

1) 訓練・推論・分析のコードは，https://github.com/ctaguchi/multi-kakiokoshi および https://github.com/ctaguchi/multi-kakiokoshi-inference で公開している．

2 関連研究

BERT[7] による文脈表現がテキストの言語処理タスクに大きな影響を与えたように，音声のモダリティにおいても自己教師あり学習を用いたエンコーディングモデルが発展してきた．そのうち，特に Wav2Vec2[3] と HuBERT[2] は，ラベルなし訓練音声データを元に音声特徴を事前学習することで，音声認識や音声分類といった下流タスクに有効であることを示した．さらに，Wav2Vec2 では，多言語のデータを事前学習に用いることで，多様な言語の音声認識に効率的にファインチューニングすることが可能である [8, 9, 10, 11]．音声認識タスクでのファインチューニングでは，エンコーダを通して得られた文脈音声表現を入力として，文字素（grapheme）数の次元のベクトルに写像する CTC 層を導入し，20 ミリ秒のフレームごとに文字素を予測する．この CTC デコーダ自体は，文字間の依存関係などを直接は考慮しないため，そのままでは，文脈的に不自然な出力をすることがある．

これに対して，エンコーダ・デコーダモデルを用いて教師あり学習を行なった Whisper[12] のようなモデルでは，音声を入力として自己回帰的にトークンを予測するため，言語モデルのように文脈を適切に考慮した出力が可能である．しかし，非音声区間で発話されていないはずのトークン列を出力するハルシネーションを起こしたり，音声終了時に同じフレーズを何度も繰り返すループに陥るといった，小規模の言語モデルで散見される事象が同様に報告されている [12, 13]．そこで，上述の Wav2Vec2 の CTC デコーディングの弱みを克服するために，デコーダに小規模の言語モデルを組み合わせることで，文脈やドメインを考慮したより高精度な出力が可能であることが報告されている [3]．

3 手法と実験設定

まず，どのようなデコーディング手法が低資源音声認識に有効であるのかを検証するために，21 の低資源言語について類似の条件下でファインチューニングを行い，精度を比較する．

3.1 データ

訓練・評価データはともに Common Voice Spontaneous Speech 2.0[2] から表 1 にまとめる 21 言語を用いる．このデータは，事前に用意された質問に対して，話者が台本なしで自由に回答する形式の発話で構成されている．合計データは約 3 万秒の音声からなり，そのうち約 7 割が訓練データとして，約 3 割が検証データとして割り当てられている．なお，本研究の実験時点でテストデータは公開されていないため，本実験では検証データを評価データとして用いる．ラベルの前処理として，重複した空白，括弧・コンマ・疑問符などの記号，メタタグを除去するが，小文字化や句点の除去は行わない．

言語名	コード	グループ
アルバニア語ゲグ方言	aln	印欧語族
ブタウィ語	bew	オーストロネシア語族
ブクス語	bxk	バントゥー語群
チガ語	cgg	バントゥー語群
ギリシャ語キプロス方言	el-CY	印欧語族
ウィチョル語	hch	ユト・アステカ語族
ヌビ語	kcn	アラビア語系クレオール
コンゾ語	koo	バントゥー語群
レンドゥ語	led	ナイル・サハラ語族
ケニ語	lke	バントゥー語群
レブ・トゥル語	lth	ナイル・サハラ語族
ミステク語南西トラヒアコ方言	meh	オト・マンゲ語族
マサワ語ミチョアカン方言	mmc	オト・マンゲ語族
西ペナン語	pne	オーストロネシア語族
ルーリ語	ruc	バントゥー語群
アンバ語	rwm	バントゥー語群
スコッツ語	sco	印欧語族
トバ・コム語	tob	ワイクル語族
トトナク語パパントラ方言	top	トトナク語族
トーロ語	ttj	バントゥー語群
ク語	ukv	ナイル・サハラ語族

表 1 実験対象言語．

3.2 モデル訓練

ファインチューニングでは，MMS-1B-all[3] を用いる．このモデルは 1600 言語以上の音声で事前訓練した後，1000 言語のラベル付き音声で追加学習を行なったものである．ファインチューニング時には，新たに付加される CTC デコーダ層以外のすべてのパラメータを固定し，CTC デコーダ層のみのパラメータ更新を行う．訓練は 10 エポック繰り返し，学習率は 0.0003 とし，最初の 100 ステップをウォームアップステップとした線形学習率スケジューラを用いる．

3.3 デコーディング手法

x_t をフレーム t の文脈化埋め込みベクトル，π_t をフレーム t の出力文字，π を出力文字列，$\mathcal{B}$ を文字列を縮約する関数とする．ここで，原像 $\mathcal{B}^{-1}(y)$ は，ラベル系列 y に縮約し得る全ての文字列 π の集合を指す．このとき，各デコーディング手法は以下のよ

2） https://datacollective.mozillafoundation.org

3） https://huggingface.co/facebook/mms-1b-all

うに定式化できる．

• 貪欲法:

$$\hat{y}_{\text{greedy}} = \mathcal{B}\left(\arg\max_{\pi} \prod_{t} P(\pi_t|x_t)\right)$$

• ビーム探索:

$$\begin{aligned}\hat{y}_{\text{beam}} &= \arg\max_{y} P_{\text{CTC}}(y|x) \\ &= \arg\max_{y} \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t} P(\pi_t|x_t)\end{aligned}$$

• n-gram モデルを用いたビーム探索:

$$\hat{y}_n = \arg\max_{y} \left(\log P_{\text{CTC}}(y|x) + \alpha \log P_{\text{LM}}(y) + \beta|y|\right)$$

ここで，ハイパーパラメータ α は言語モデルをどれだけ信頼するかの重みで，β は単語数を制御するペナルティである．

貪欲法では常に確率が最大となる文字素そ選択するため，隣接する文字素同士の文脈を直接は考慮しないが，ビーム探索ではより柔軟に文字素の共起関係等に基づいた予測を行う．さらに n-gram 言語モデルを組み込むことで，文字素動詞のみならず，語同士の共起関係も考慮し，文脈上および正書法上正しい語形の予測を促す．

実験では，貪欲法，ビーム探索，2-gram 言語モデルを用いたビーム探索，5-gram 言語モデルを用いたビーム探索の四つのデコーディング手法を用いて比較する．なお，ビーム幅はどの場合も 50 で固定し，ハイパーパラメータとして $\alpha = 0.2$，$\beta = 0.0$ を用いる．

4 結果と分析

4.1 実験結果

実験の結果を表 2 に示す．検証した 21 言語のうち，大多数の言語（18 言語）において，5-gram 言語モデルを用いたデコーディングは貪欲法よりも誤り率を平均 18.8%低下させ，音声認識精度の向上に有効であった．しかし，ウィチョル語（hch），コンゾ語（koo），ケニ語（lke）の三言語では，5-gram 言語モデルはむしろ認識精度を大きく（平均 19.3%）悪化させた．以下の節では，この性能悪化の原因の仮説を立て，検証する．

4.2 なぜ特定の言語で性能が低下したか

データはどの言語においても自発的な発話であり，訓練データの合計音声長は一貫して約 6 時間弱であるため，ドメインや訓練データ量の差異といった非言語的な要因に起因する結果とは考え難い．また，言語の音韻的複雑性が認識精度に影響を与えることも考え難い [14] ことから，言語の形態統語論および正書法上の特徴がこのような差を引き起こしているという仮説を立てることができる．実際に，5-gram 言語モデルの使用によって認識精度が悪化した言語 (1) とそうでない言語 (2) を比較すると，精度が悪化した言語は一語[4]がより多くの文字から成っていることがわかる．ウィチョル語（ユト・アステカ語族）の形態論は複統合的（polysynthetic）な性格を持ち，名詞抱合（noun incorporation）や動詞の屈折を通して，一語が多くの形態素から成立しうる [15]．また，コンゾ語とケニ語（バントゥー諸語）は膠着語であり，特に複数の形態素が語基の前に付加されうる [16]．そのため，これらの言語では語のタイプが多様になりやすく，n-gram 言語モデルの確率分布がまばらになり，誤った語の組み合わせが予測されていると考えられる．

(1) a. **hch**: Ne nunuutsiyari nepɨtiukwaiximekai mɨtiumawekaitsie ’unakɨ, kuukurikɨ (...)
b. **koo**: Abalhwere abakabonde isithwasolholhabo nakutsibutsibu omwa mibiri eyathukakolha, (...)
c. **lke**: Insobola ekyokwembesia nadala nga okubba emgoma eyolukenye injisobola ninjikuba (...)

(2) a. **aln**: kur shoku em mw pershwndet pwr zotin dal e pi njw kafe, pwr shembull, edhe pi njw (...)
b. **bxk**: Yani byakhulya nibyo olya byakhuwelesya bulamu bulayi, elibyakhulya byebatekhile (...)
c. **led**: Ma dho vùgá nyù ndaní ma rónga déy lǎní ma dho zz ddí nyǔ ndaní ma rónga déy lǎní (...)

4.3 仮説検定

上記の議論に基づき，本節では「語の分布のまばらさ」を定量化する様々な指標を用いて，スパースな語の分布が n-gram 言語モデルを組み合わせた CTC 音声認識の性能と相関するかを検証する．語の分布のまばらさは多様な定義ができるため，ここでは以下に説明する複数の指標を用いて，回帰分析を行う．ここで，目的変数は，ベースラインの貪欲法の WER と 5-gram 言語モデルデコーディングの WER の差とする．以下では，V を語彙の集合，$p(x)$ を語 x の確率とする．

4) ここでは，スペースで区切られるトークンを語とする．

	aln	bew	bxk	cgg	el-CY	hch	kcn	koo	led	lke	lth	meh	mmc	pne	ruc	rwm	sco	tob	top	ttj	ukv
貪欲法	47.39 19.07	49.69 16.84	47.54 13.12	40.56 9.53	38.41 12.38	54.74 11.81	48.40 20.15	66.99 19.51	28.30 10.31	54.58 15.03	34.69 15.17	39.10 15.21	60.70 26.07	32.38 11.72	32.02 11.97	56.80 13.75	56.08 20.32	54.51 16.61	56.36 12.87	23.84 4.30	38.98 12.78
ビーム探索	46.73 18.90	48.90 16.64	47.75 13.09	40.23 9.47	38.13 12.19	54.46 11.71	47.43 19.97	**66.95** **19.40**	28.30 10.29	**53.98** **14.89**	34.54 15.13	38.29 15.04	60.32 25.82	32.16 11.69	31.76 11.85	56.39 13.63	55.69 20.18	54.51 16.58	55.25 12.71	23.48 4.24	39.14 12.85
2-gram	42.96 18.58	45.48 16.47	48.62 13.33	39.83 9.48	32.74 11.73	**52.91** **11.38**	43.00 19.38	72.66 20.17	26.39 9.83	55.63 15.46	31.00 14.63	35.58 14.86	58.47 25.66	28.63 11.15	27.70 11.58	57.06 14.06	56.70 20.79	53.34 16.43	59.51 13.35	21.98 4.13	36.36 12.25
5-gram	**37.95** **17.72**	**40.17** **15.61**	**39.34** **12.02**	**31.12** **8.29**	**28.71** **10.94**	62.89 12.90	**38.58** **18.43**	74.59 21.32	**21.56** **8.59**	66.62 17.75	**28.25** **14.04**	**31.56** **13.82**	**55.28** **24.81**	**24.29** **10.14**	**23.36** **10.52**	**45.71** **12.42**	**50.65** **19.85**	**46.07** **15.09**	**45.84** **11.62**	**14.12** **3.11**	**32.27** **11.48**

表 2　実験結果。上段の値は WER、下段の値は CER。

1-gram エントロピー　1-gram エントロピーは各語形の確率をもとに求められる情報量であり，以下のように定義される．

$$H_{\text{unigram}} = \frac{-\sum_{x \in V} p(x) \log_2 p(x)}{\log_2 |V|}$$

2-gram エントロピー　2-gram エントロピーは語 x の次に語 y が観測される確率をもとに求められる情報量であり，以下のように定義される．

$$H_{\text{bigram}} = \frac{-\sum_{x,y \in V} p(x,y) \log_2 p(y|x)}{\log_2 |V|}$$

Zipf 傾斜　Zipf の法則によれば，自然言語の語彙は $f(r) \propto r^{-s}$（r は頻度の順位）のように，少数の高頻度語彙と多数の低頻度語彙から成る指数関数的な分布を示す [17]．しかし，形態論的に複雑な言語では高頻度の形態素も複数の語形で出現しうるため，分布がより緩やかな傾斜を示し，$-s$ がより大きくなると考えられる．そこで，頻度と順位の対数近似を行い，求められた $-s$ の値を比較する．

タイプ・トークン比　タイプ・トークン比（type-token ratio）はコーパスの合計語数 N を語彙数で割ったものであり，高いほど多様な語形が出現していると言える．

孤語比　孤語（hapax legomena）とは，コーパス中で一度しか出現しない語形のことであり，形態論的に複雑な言語ほど孤語比が高いと考えられる．

テイル確率質量　孤語比と関連して，形態論的に複雑な言語では，コーパス中に数回しか出現しない語が多く存在すると考えられる．そこで，出現確率が 1%以下の語をテイルとして，その確率質量を指標として用いる．

ジニ係数　ジニ係数 $\text{Gini} \in [0,1]$ は主に経済学で用いられる指標であるが，データの不均等さを表すため，本分析の指標としても用いることができる．長さ n の数列のジニ係数は以下の式から算出される（p_i は第 i 位の語の確率）．

$$\text{Gini} = \sum_{i=1}^{n} (2i - n - 1) p_i$$

平均語長　複統合語や膠着語では多数の形態素が接辞として組み合わされるため，一語が長大になる傾向がある．そこで，語の平均文字数 $\frac{|x| \cdot f_x}{N}$（f_x は語 x の出現回数）を指標の一つとして用いる．

4.4　検証結果

すべての指標の数値を説明変数として用いて回帰分析を行った結果，回帰モデルは説明変数の分散のおよそ半分を説明しており（調整済み $R^2 = 0.48$），説明変数と目的変数の間に統計的に有意な関連が認められた（$p = 0.032 < 0.05$）．しかし，これら 8 つの説明変数はどれも形態論的複雑性や語の分布のばらつきを測ることを目的とした指標に基づくものであり，説明変数同士が非常に強い相関関係を持つ多重共線性（multicollinearity）を有していると考えられ，今回のデータに過剰適合したモデルである可能性がある．

そこで，主成分分析を用いて多重共線性を除去し，次元数は保持して主成分回帰を行ったところ，やはり回帰モデルは説明変数の分散を概ね説明し（調整済み $R^2 = 0.61$），説明変数と目的変数の間に統計的に有意な関連が認められた（$p = 0.0073 < 0.05$）．したがって，形態論的に複雑で，スパースな語の分布を持つ言語のデータを用いた音声認識モデルは，n-gram 言語モデルを組み合わせてもむしろ性能が悪化する傾向が強いと言える．

5　結論

実験から，一般的に n-gram 言語モデルを用いたビーム探索デコーディングは，低資源言語の音声認識においても有効であることが示された．しかしながら，一部の低資源言語ではむしろ精度が大幅に悪化した．検証の結果，これらは形態論的に複雑な複統合性や膠着性を示す言語であり，CTC 音声認識における n-gram 言語モデルの無効性は形態論的複雑性や語の分布のスパース性と有意に関連していることが示された．

謝辞

本研究の着想のきっかけとなった Mozilla Foundations の Robert Pugh 氏に感謝の意を表します．本研究はアメリカ国立科学財団 IIS-2137396 の助成を受けたものです．また，本研究の実験に際して，限られた計算資源の使用の順番を譲って下さった研究室の Ken Sible 氏，Aarohi Srivastava 氏，Katsumi Ibaraki 氏に深く感謝いたします．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[2] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[4] Andrew Maas, Ziang Xie, Dan Jurafsky, and Andrew Ng. Lexicon-free conversational speech recognition with neural networks. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 345–354, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[5] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In Eric P. Xing and Tony Jebara, editors, **Proceedings of the 31st International Conference on Machine Learning**, Vol. 32 of **Proceedings of Machine Learning Research**, pp. 1764–1772, Bejing, China, 22–24 Jun 2014. PMLR.

[6] Zhaolin Li and Jan Niehues. Enhance contextual learning in ASR for endangered low-resource languages. In Sang Truong, Rifki Afina Putri, Duc Nguyen, Angelina Wang, Daniel Ho, Alice Oh, and Sanmi Koyejo, editors, **Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)**, pp. 1–7, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020.

[9] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale, 2021.

[10] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages, 2023.

[11] Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages, 2025.

[12] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[13] Yingzhi Wang, Anas Alhmoud, Saad Alsahly, Muhammad Alqurishi, and Mirco Ravanelli. Calm-whisper: Reduce whisper hallucination on non-speech by calming crazy heads down, 2025.

[14] Chihiro Taguchi and David Chiang. Language complexity and speech recognition accuracy: Orthographic complexity hurts, phonological complexity doesn't. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15493–15503, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[15] José Luis Iturrioz and Paula Gómez López. **Gramática Wixarika I**. No. 3 in Lincom Studies in Native American Linguistics. Lincom Europa, Munich, 2006.

[16] Derek Nurse. **Tense and Aspect in Bantu**. Oxford University Press, 2008.

[17] George K Zipf. **Human behavior and the principle of least effort**. Addison-Wesley, 1949.

A　検証指標の数値の詳細

表 3 に実験結果と，各言語に用いたコーパスにおける検証に用いた指標の数値を示す．

言語	貪欲法	beam	2-gram	5-gram	H_{unigram}	H_{bigram}	$-s_{\text{Zipf}}$	TTR	HLR	TPM	Gini	AWL
aln	47.39	46.73	42.96	37.95	0.73	0.21	-1.12	0.11	0.06	0.69	1.83	3.61
bew	49.69	48.90	45.48	40.17	0.78	0.20	-1.07	0.14	0.08	0.84	1.79	5.11
bxk	47.54	47.75	48.62	39.34	0.85	0.18	-0.80	0.30	0.20	0.93	1.64	6.68
cgg	40.56	40.23	39.83	31.12	0.84	0.19	-0.79	0.30	0.21	0.91	1.64	7.09
el-CY	38.41	38.13	32.74	28.71	0.75	0.21	-0.98	0.16	0.09	0.77	1.78	5.20
hch	54.74	54.46	52.91	62.89	0.80	0.21	-0.65	0.33	0.26	0.83	1.64	7.69
kcn	48.40	47.43	43.00	38.58	0.73	0.20	-1.31	0.07	0.04	0.71	1.85	3.73
koo	66.99	66.95	72.66	74.59	0.88	0.19	-0.69	0.39	0.28	0.95	1.56	7.71
led	28.30	28.30	26.39	21.56	0.68	0.17	-1.43	0.05	0.02	0.60	1.89	3.11
lke	54.58	53.98	55.63	66.62	0.86	0.20	-0.75	0.33	0.24	0.96	1.62	6.52
lth	34.69	34.54	31.00	28.25	0.75	0.20	-1.28	0.08	0.04	0.77	1.83	3.92
meh	39.10	38.29	35.58	31.56	0.69	0.17	-1.47	0.04	0.02	0.62	1.89	3.54
mmc	60.70	60.32	58.47	55.28	0.71	0.20	-0.80	0.19	0.14	0.66	1.77	4.12
pne	32.38	32.16	28.63	24.29	0.71	0.18	-1.35	0.06	0.03	0.67	1.87	4.41
sco	32.02	31.76	27.70	23.36	0.72	0.21	-1.20	0.08	0.04	0.71	1.85	3.93
ruc	56.80	56.39	57.06	45.71	0.84	0.19	-0.79	0.29	0.21	0.94	1.65	6.43
rwm	56.08	55.69	56.70	50.65	0.78	0.18	-0.84	0.23	0.16	0.81	1.73	5.72
tob	54.51	54.51	53.34	46.07	0.77	0.20	-0.86	0.22	0.15	0.74	1.72	5.61
top	56.36	55.25	59.51	45.84	0.81	0.20	-0.74	0.30	0.22	0.85	1.66	6.04
ttj	23.84	23.48	21.98	14.12	0.83	0.19	-0.85	0.25	0.17	0.90	1.69	6.76
ukv	38.98	39.14	36.36	32.27	0.69	0.19	-1.12	0.10	0.06	0.64	1.85	3.97

表 3　実験結果。beam はビーム探索デコーディング、2-gram はビーム探索+2-gram 言語モデルデコーディング、5-gram はビーム探索+5-gram 言語モデルデコーディングの WER を指す。H_{unigram} は 1-gram エントロピー、H_{bigram} は 2-gram エントロピー、$-s_{\text{Zipf}}$ は Zipf 傾斜（Zipf slope）、TTR はタイプ・トークン比（type–token ratio）、HLR は孤語比（hapax legomena ratio）、TPM はテイル確率質量（tail probability mass）、Gini はジニ係数、AWL は平均語長（average word length）を指す。