文字言語モデルからの単語言語モデルの教師なし合成

村脇 有吾[1,a)]

概要：Transformer に基づく事前訓練文字言語モデルから単語言語モデルへの教師なし合成の実現可能性を示す．教師なし単語分割における興味の中心は学習を実現する帰納バイアスを明らかにすることだが，ニューラル言語モデルに基づく場合，アーキテクチャ上の制約から有限語彙を前提とするにもかかわらず，教師なし単語分割においては語彙を事前に決定できないという問題への取り組みを中心に据えざるを得ない．本稿では，この問題を解決するために多段階の訓練手続きを提案する．提案手法は単語境界における確率的不確実性を取っ掛かりとして利用しており，幼児の初期学習との関連が示唆される．

Unsupervised Synthesis of Word Language Models from Pretrained Character Language Models

Abstract: We demonstrate the feasibility of transforming a Transformer-based pretrained character language model into a word language model without explicit supervision on word segmentation. While the main interest in unsupervised word segmentation lies in identifying the inductive biases that facilitate language acquisition, neural language models face technical challenges due to their architectural constraints requiring a fixed vocabulary, despite the inability to predefine this vocabulary in unsupervised settings. To address this issue, we propose a multi-stage training procedure. Our method leverages the stochastic uncertainty pertaining to word boundaries to bootstrap the process, suggesting a connection to early learning in infants.

1. はじめに

教師なし単語分割は明示的な教師信号なしにテキストを単語列に分割するタスクであり，主に 2 つの観点から取り組まれてきた．1 つは幼児の言語獲得への計算論的取り組みという位置づけであり，言語獲得を可能にする帰納バイアスを明らかにすることに興味がある [11], [22]．もう 1 つは単語境界を明示しない書記言語に対する最初の言語処理タスクという位置づけである [42], [67]．実際のところ，そのような書記言語は長い歴史を持ち，豊富な言語資源をともなう傾向があることから，後者の需要は乏しい．ただし，テキストから連続音声への拡張は，書き言葉を持たない世界中の言語という潜在的に広い応用先を持っている [50]．

本稿では，近年の大規模言語モデル（LLM）の成功 [12], [44], [46] に刺激を受け，Transformer に基づく言語モデルを対象とした教師なし単語分割に取り組む．LLM の高い言語処理能力は，サブワード列を処理しているに過ぎないにもかかわらず，構文，語用論，談話関係等を暗黙的に理解していることを示唆する [13], [52], [64]．LLM がこのように低位の構造から高位の構造への写像を暗黙的に獲得できているのであれば，同様にして，Transformer に基づく事前訓練**文字**言語モデルは単語や単語同士の相互作用について相応の知識を暗黙的に得ている可能性が高い．したがって，モデルに適切な帰納バイアスを与えれば，事前訓練文字言語モデルから単語言語モデルを引き出すことが可能ではないかと期待できる．その一方で，LLM がしばしば意味をなさないサブワードの列を難なく扱っていることを考慮すると，そのような帰納バイアスは強力なものでなければならないと推測できる．

このように帰納バイアスが教師なし単語分割における興味の中心であるものの，ニューラル言語モデルに基づく場合，教師なし単語分割への適用を困難にするアーキテクチャ上の制約への取り組みを中心に据えざるを得ない．具体的には，入出力の語彙サイズを固定しなければならないという技術的制約は，語彙を事前に決定できないという教師なし単語分割の性質と相性が悪く，その解決策は自明ではない．入力語彙を不定サイズとすることは比較的容易だが，出力側は困難である．このような事情から，ニューラル

[1] 京都大学
Yoshida-honmachi, Sakyo-ku, Kyoto 606–8501, Japan
[a)] murawaki@i.kyoto-u.ac.jp

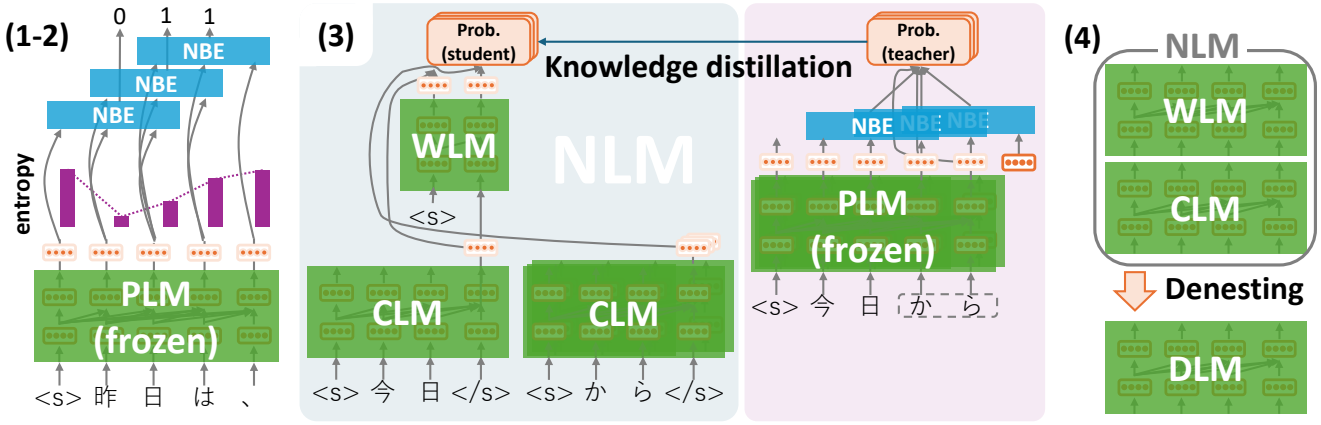

図 1 提案手法の概要（日本語の例に基づく）
Fig. 1 Overview of the proposed method using Japanese examples.

ネットを用いて教師なし単語分割に取り組む既存研究 [51] は，通常の単語言語モデルとは似ても似つかぬアーキテクチャを採用してきた．

本稿では，この問題を解決して通常のアーキテクチャの言語モデルを得るために，図 1 に示す多段階の訓練手続きを提案する．この複雑な手続きは，最終成果物として固定の語彙サイズの通常の単語言語モデル（DLM）を得るために必要となっている．このことを明らかにするために，訓練手続きを最後からさかのぼって説明したい．

DLM を直接訓練することは困難なため，代理のモデルとして文字言語モデル（CLM）と単語言語モデル（WLM）を入れ子にしたモデル（NLM）を用意する．DLM は，NLM の訓練後に入れ子構造を解消することで得る（4）．

NLM は指数的に大きな語彙を持っており，ソフトマックス交差エントロピー損失に基づく通常の言語モデル訓練を適用できない．確率の計算における分母が計算困難だからである．計算困難な分母を持つ確率モデルの推定の定跡は分母を打ち消すというものである．単純には確率変数の 2 つの値（単語候補）を用い，それらの確率の比をとれば良いが，2 個以上の単語候補を考え，局所的に正規化された擬似確率を用いることも可能である．

この擬似確率に適切な教師信号を与える必要があるが，どのようにすれば良いだろうか？本稿のアイデアは，事前訓練文字言語モデル（PLM）の生成確率を利用するというものである．ただし，PLM は与えられた文字列が 1 つの単語を構成するかを考慮していないため，事後的な確率の補正が必要となる．すなわち，PLM がある文字列に与える確率質量のうち，その前後に単語境界が来て，内部には単語境界が来ないという条件を満たす一部のみが真に単語の生成確率に比例する．このアイデアを具体化するために，各文字間に単語境界が来る確率を推定するモデル（NBE）を用意する．PLM と NBE の組合せによって NLM を訓練する（3）．

NBE はニューラルネットによって構成できるが，どのように（事前）訓練すれば良いだろうか？この問題への解法として，単語境界における確率的不確実性という，教師なし単語分割の研究の初期において注目されてきた手がかりを利用する（1–2）．この手がかりは幼児が初期段階において利用しているものの，発達とともに別の手がかりに移行することが示唆されている [53]．技術的制約を克服するために提案された本手続きが，結果的に幼児の段階的な発達との関連を示すことは興味深い．

実験の結果，提案する多段階の訓練手続きによって段階的に単語分割が改善されることが示された．提案モデルが長さバイアス [34] として知られる問題に直面することが確認されたが，モデルを大きくすることで自然に解消される可能性が示唆される．

2. 関連研究

2.1 教師なし単語分割

教師なし単語分割の初期の研究は単語境界の特定に取り組んでいた．核となるアイデアは，単語境界は統計的な系列モデルが不確実性を見出す位置に来やすいというものである [21], [25], [47]．このアイデアは言語処理の応用研究 [5], [19] だけでなく，ニューラルネットに基づく基礎研究 [15], [18] でも採用されていた．

その後，研究の焦点は単語そのものの特定に移行した．ここでのアイデアは，よく出現する文字列は単語である可能性が高いというものである [6], [10], [11]．このアイデアの素朴な実現は文脈非依存の unigram モデルだが，すぐに文脈を考慮した n-gram モデルへの拡張が提案されている [59]．ただし，n-gram モデルが普及したのは，推論の困難さをノンパラメトリックベイズが見事に解決してからであった [14], [22], [42], [56]．ベイズ言語モデルは少量のテキストであっても頑健に動き，教師なし単語分割研究の最盛期を築いたと言っても過言ではない．しかし，記号に基

づく言語モデルであるため，ニューラル言語モデルが持つ利点，すなわち長距離依存の処理能力や埋め込みに基づく汎化能力を欠いている [7].

教師なし単語分割の研究が低調化したニューラル時代の代表例は分節言語モデル [51] である．このモデルはいくつかの拡張 [31], [36], [62] を生んでいるが，いずれにしても文字列に基づく文脈符号化器から単語候補を文字列として条件付きで生成するという特徴を持つ．このように，分節言語モデルは通常のニューラル言語モデル（次のトークンは出力ベクトルと埋め込み行列とのドット積に基づいて選ばれる）とはまったく異なるアーキテクチャを採用していることから，通常の言語モデルを教師なしで合成したいという本稿の目的とは相容れない．また，単語は自律的な単位であり，その意味は文脈とはある程度独立に決まっていると考えられるが，通常のニューラル言語モデルはこの直感に沿っているのに対し，分節言語モデルはこの直感に合わない．また，精度面でも，ニューラルネットが記号に基づくモデルを大差で打ち破るという自然言語処理に一般的な傾向に反して，ベイズ言語モデルと同程度に過ぎない．したがって，本稿の実験ではベイズ言語モデルのみを比較対象として取り上げる．

Transformer アーキテクチャ [58] に基づく事前訓練符号化器の成功 [16] にともない，教師なし単語分割への応用 [17], [37], [60] が試みられていいたのに対し，その後の事前訓練復号器，つまり言語モデルの成功 [12] は同様の動きを引き起こしていない．事前訓練符号化器は教師**あり**単語分割においても支配的 [43], [57], [66] であり，言語モデルで置き換えられる兆候は見られない．言語モデルを単語分割に用いると，並列化が困難であったり，長距離依存の副作用として動的計画法が使えないなど欠点が理由として挙げられる．こうした欠点はあるものの，高いテキスト生成能力を持つニューラル言語モデルに単語分割を行う能力を持たせることは，人間のような知能を実現するうえで重要な一歩であると考えている．

2.2 ニューラル言語モデルにおける単語

現代のほぼすべてのニューラル言語モデルはサブワードを採用している [40]．サブワードはニューラル言語モデルが固定語彙を要求するという技術的制約への妥協であるだけでなく，高い性能を発揮することが実証されている [41]．とくに有名な ChatGPT の場合，トークナイザ[*1]が英語（ラテン文字）以外のテキストを容赦なく過分割し，しばしばバイト列にまで分割する [1], [30] にもかかわらず，多数の言語処理タスクにおいて圧倒的な性能を発揮している．このようなトークンの無意味さと全体的な性能との食い違いは，ニューラル言語モデルに単語単位で処理を行わせるには強力な帰納バイアスを課す必要があることを示唆する．

[*1] https://platform.openai.com/tokenizer

3. 提案モデル

3.1 入れ子言語モデル

本稿の目的は，Transformer に基づく事前訓練文字言語モデル（**PLM**, pretrained language model）から単語言語モデルへの教師なし合成である．テキストの先頭と終端を表す特殊記号をそれぞれ<s>と</s>で表すとする．また，単語の最大文字列長を L とする．L を導入することで単語語彙は閉集合となるものの，指数的に大きく陽に列挙できない．

最終成果物（**DLM**, denested language model）の代理モデルである入れ子言語モデル（**NLM**, nested language model）は，PLM の 2 個の複製を入れ子にすることで構成される．一方は文字系列に基づく単語モデル（**CLM**, character language model）である，もう一方は単語系列に基づくテキストモデル（**WLM**, word language model）である（図 1（3）の左側）．ニューラルモデルは入出力の端部を除いて固定長のベクトルを処理しているため，それまで文字埋め込みを処理していたモデルに単語埋め込みを与えることは造作もない．同様の手法は特にマルチモーダル設定 [3], [38] において多用されている．

PLM を複製する際，WLM に対しては，PLM の入出力の埋め込み行列を破棄し，<s>と</s>に対応する入力埋め込みのみを保持する．入出力の埋め込み行列の代わりに，両者を同一の行列で扱う [45]．ただし，そのような行列は計算困難であり，明示的に表現することはなく，各単語埋め込みを CLM を使って動的に生成する．CLM は<s>と</s>で囲まれた文字列を単語候補として受け取り，最終トークンである</s>に対応する隠れベクトルを出力する[*2]．

NLM をこのように構成する背景には，PLM が事前訓練を通じて単語やその相互作用について相当程度の知識を暗黙的に得ているという予想がある．比較的小規模な追加訓練だけで，CLM は単語ベクトル生成に適応し，WLM は単語同士の相互作用に注力することを期待している．

3.2 入れ子構造の解消

4 節で詳述する訓練手続きのあと，NLM の入れ子構造を解消することで最終成果物である DLM を得る．まず，付録 A.1 で述べる適当な手順に従って，NLM の巨大な語彙のごく一部を採用する．CLM は文脈非依存であり，訓練終了後は語彙内の単語ベクトルを一度計算して結果を保存しておけば良いため，CLM 本体は不要になる．こうして得られた単語ベクトルを並べて行列を形成し，WLM の入力および出力の埋め込み行列とすれば良い．厳密には，WLM 自身が持つ<s>と</s>の埋め込みと CLM 由来の埋め込み

[*2] <s>と</s>で囲むのは，大半の単語候補は文字言語モデルの事前訓練時に単独で観測し得るという期待を反映している．

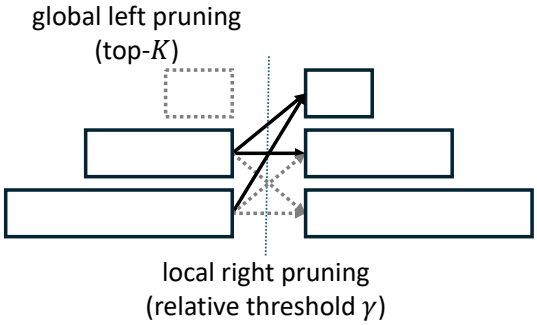

図 2 ビーム探索における 2 種類の枝刈り [54]（枝刈りされた候補を破線で示す．注目する境界で終わる左ノードについては，<s>からの累積対数確率に基づいて上位 K 候補のみを残す．残った各左ノードについて，その境界から始まる右ノードを単語生成確率によってソートし，確率が最大確率の γ 倍を下回る候補を枝刈りする．）

Fig. 2 Two types of pruning for beam search [54]. Pruned candidates are represented by dashed lines. For the left (ending at the boundary) nodes, only the top-K nodes are retained based on the cumulative log probability starting from <s>. Subsequently, for each remaining left node, the right (starting at the boundary) nodes are sorted by their word generation probabilities, and those with probabilities lower than γ times the maximum probability are pruned.

を結合して行列を作成する必要がある．最後に CLM を破棄することで入れ子構造を解消し，残ったモデルを通常の単語言語モデルとして扱う．

DLM をテキスト生成に用いるのは容易である一方で，DLM にテキストを入力する場合には注意が必要である．現代のニューラルモデルは，それとは独立したトークナイザを用いた前処理によってテキストをトークン列に分割するのに対して，DLM は自身を用いることで適切な単語列を得なければならない．この目的のために，伝統的な教師あり単語分割に似たアルゴリズムを採用する．すなわち，入力テキストに対して，まずは辞書引き [2] によって単語候補で構成されるラティスを構築する．次に，ビーム探索によってラティス中でもっとも確からしい経路を近似的に選択する．ビーム探索における枝刈りには複数の戦略が考えられる [20], [54] が，本稿では大域左枝刈りと局所右枝刈りの組合せ（図 2）を採用する．

4. 訓練手続き

NLM は DLM を直接教師なしで訓練できないことを理由に代理モデルとして導入したものだが，NLM の訓練方法もまた自明ではない．本稿で提案する多段階の訓練手続きの動機については 1 節で後ろからたどることで既に述べたので，本節では訓練手続きを前から順番に説明したい．

4.1 エントロピーに基づく分割

訓練手続きの最初の手がかりとして単語境界における確率的不確実性に着目する．この手がかりは教師なし単語分割の研究の初期に注目されていただけでなく，幼児の言語獲得の初期段階で利用されている可能性が示唆されている [29], [53]．不確実性の定式化には複数の方法が考えられるが，本稿ではエントロピーに基づく定式化 [28], [32], [55] を採用する．局所エントロピー値

$$s_n = -\sum_c P^{\mathrm{PLM}}(c|\boldsymbol{c}_{1:n}) \log P^{\mathrm{PLM}}(c|\boldsymbol{c}_{1:n})$$

は，文字列 $\boldsymbol{c}_{1:n} = (c_1, \cdots, c_n)$ に後続する文字についての不確実性の程度を表す．この条件付確率は PLM により求められる．

本稿で採用するエントロピーに基づく分割（**ES**, entropy-based segmentation）は，c_n のあとに単語境界を認定するか否かを局所エントロピー値に基づいて決める．具体的には，$s_n/s_{n-1} > \alpha$ $(\alpha \leq 1)$ かつ $s_n > \beta$ であれば，単語境界を認定するというヒューリスティックな規則を採用する．なお，s_1 はテキスト冒頭におけるエントロピー値を表す．2 つの条件のうち，最初の条件は，文字系列に対応するエントロピー値の系列を考えたとき，単語内部では減少し，単語境界で上昇する傾向があることを反映している．2 番目の条件は，この単調減少条件に対する例外への対処として設けられている．特に日本語の場合，形態変化によって単語末尾周辺でエントロピー値の上昇が起きやすい．

残る問題はハイパーパラメータ α および β をどのように求めるかである．本稿では訓練データを用いたハイパーパラメータ探索 [4] によって求めることとする．分割なしの生テキストのみからこれらの値を求める手法の確立は将来研究に委ねる．

4.2 NBE の事前訓練

ES は NLM の訓練に直接用いられるのではなく，ニューラル境界推定器（**NBE**, neural boundary estimator）の事前訓練に用いる．事前訓練と述べた通り，NBE は，人間の場合 [53] と同様に，訓練手続きの後の段階で調整される．

NBE は PLM の出力を受け取る小さな多層パーセプトロンによって構成し，文字 c_n の直後に単語境界が来る確率を以下のように求める:

$$\sigma\left(\frac{\mathrm{Linear}(\mathrm{ReLU}(\mathrm{Linear}(h^{\mathrm{PLM}}_{n-1} \oplus h^{\mathrm{PLM}}_{n} \oplus h^{\mathrm{PLM}}_{n+1})))}{\tau_{\mathrm{NBE}}}\right) \quad (1)$$

ここで σ はシグモイド関数，$\oplus$ はベクトル連結，h^{PLM}_n は PLM が c_n に対して出力する隠れベクトル，τ_{NBE} は温度パラメータである．事前訓練中は $\tau_{\mathrm{NBE}} = 1$ とする．

事前訓練中は，PLM がテキストをランダムに生成し，ES が各文字位置に対して擬似ラベルを付与し，NBE はその擬似ラベルを用いて二値交差エントロピー損失を最小化する

ようにパラメータを更新する．なお，PLM のパラメータは固定のままとする．c_{n+1} は事前訓練中は常に観測される（</s>となる場合がある）のに対して，後で述べるように，後段では観測されない場合がある．そのような場合には，h_{n+1}^{PLM} を調整可能なパラメータ h_*^{NBE} で代替する．h_*^{NBE} を事前訓練するために，c_{n+1} が</s>であるとき，0.2 の確率で h_{n+1}^{PLM} を h_*^{NBE} で置き換える．

4.3 NLM の訓練

提案する多段階の訓練手続きの山場は NLM の訓練である．再帰的な手続きを考えるために，ひとまず文字列 $\boldsymbol{c}_{1:n}$ が単語列 $\boldsymbol{w}_{1:m}$ に分割されているとする．NLM が次の単語 $w_{m+1} = \boldsymbol{c}_{n+1:n+l}$ （$1 \leq l \leq L$ であり，ひとまず w_{m+1} が</s>でないとする）を生成する確率は以下で与えられる:

$$P^{\mathrm{NLM}}(w_{m+1}|\boldsymbol{w}_{1:m}) \propto \exp(h_m^{\mathrm{WLM}} \cdot h_{m+1}^{\mathrm{CLM}}), \tag{2}$$
$$h_m^{\mathrm{WLM}} = \mathrm{Transformer}^{\mathrm{WLM}}(e_{\texttt{<s>}}^{\mathrm{WLM}}, \boldsymbol{h}_{1:m}^{\mathrm{CLM}}),$$
$$h_{m+1}^{\mathrm{CLM}} = \mathrm{Transformer}^{\mathrm{CLM}}(\texttt{<s>}, \boldsymbol{c}_{n+1:n+l}, \texttt{</s>})$$

ここで Transformer* は最後の入力トークンに対応する隠れベクトルを返す Transformer とする．入力については，CLM は通常の埋め込み層を使って入力の離散トークンを連続ベクトルに変換するのに対して，WLM は CLM が出力するベクトルを直接受け取ることに留意すべきである．ただし，WLM はテキスト冒頭<s>に対しては自身が持つ埋め込み $e_{\texttt{<s>}}^{\mathrm{WLM}}$ を用いる．また，式（2）で例外扱いしていた w_{m+1} が</s>の場合は CLM を呼び出さず，h_{m+1}^{CLM} の代わりに $e_{\texttt{</s>}}^{\mathrm{WLM}}$（WLM 自身が持つ</s>用の入出力兼用埋め込み）を用いる．テキスト冒頭・末尾の例外処理のために少々込み入っているが，要するに，図 1（3）左側に示すように，CLM が文字列で表現される単語候補，WLM が単語列を扱う入れ子構造を素直に実現している．

式（2）の右辺値の正規化は計算困難であり，したがって通常のソフトマックス交差エントロピー損失は計算できない．そのため，ES や NBE に基づく（誤りを含んだ）単語分割を単純に訓練に用いることはできない．計算困難性を回避するために，定跡に従い，単語候補群の確率を正規化しないまま訓練に用いる．

正規化されていない確率分布に対する教師信号を用意するために，PLE と NBE の組合せを用いる．

$$\begin{aligned} P(w_{m+1}|\boldsymbol{c_{1:n}}) \propto\ & P^{\mathrm{PLM}}(\boldsymbol{c}_{n+1:n+l}|\boldsymbol{c}_{1:n}) \\ & \times P(b_n = 1|\boldsymbol{c}_{1:n+l}) \\ & \times \prod_{i=n+1}^{n+l-1} P(b_i = 0|\boldsymbol{c}_{1:n+l}, \boldsymbol{b}_{n:i-1}) \\ & \times P(b_{n+l} = 1|\boldsymbol{c}_{1:n+l}, \boldsymbol{b}_{n:n+l-1}) \end{aligned} \tag{3}$$

ここで $b_i \in \{0, 1\}$ は文字位置 i の直後に単語境界を置く（1），もしくは置かない（0）ことを表す．最初の項は文字列 $\boldsymbol{c}_{n+1:n+l}$ の生成確率を表す．残りの項は，その文字列が 1 つの単語であるという条件を満たすように事後的な補正を行っている．すなわち，単語境界がその文字列の前後の来ると同時に，文字列内部はすべて非境界でなければならない．これらの項は NBE（$\tau_{\mathrm{NBE}} = 1$）によって近似される．なお，c_{n+l+1} は未観測であるため，前節で導入した h_*^{NBE} が必要となる．

Algorithm 1: NLM の訓練

```
PLM から文字列を生成;
τ_NBE = 0.2 で NBE によって確率的に文字列を単語列に分割;
foreach 単語列中の各単語について do
    PLM を用いて S − 1 個の単語候補を生成し，元の単語候補と合わせる;
    L_KD + L_WLM + L_NBE を計算し，系列レベルの損失に足す;
逆誤差伝播を実行;
CLM のパラメータを更新;
if 勾配累積が閾値に達した then
    WLM のパラメータを更新;
    （オプション）NBE のパラメータを更新;
```

式（2）と式（3）は知識蒸留 [26], [33], [48] によって接続する．知識蒸留は教師（PLM/NBE）の確率分布と生徒（NLM）の確率分布のカルバック・ライブラー（KL）情報量の最小化を行う（図 1 (3)）．ただし，上述の通り，計算困難性を回避するために，単語候補すべてを考慮した真の確率分布ではなく，単語候補の小さな部分集合 $\{w_1, \cdots, w_S\}$ のみを考慮して局所的に正規化した擬似確率を用いる．単語候補 w_s に対して，$\mathrm{logit}_s^{\mathrm{NLM}}$ を式（2）における指数関数の入力引数とする．また，$\mathrm{logit}_s^{\mathrm{NBE}}$ を式（3）右辺に指数関数を適用した結果とする．温度によってスケーリングされた生徒の擬似確率は以下で与えられる:

$$\tilde{p}_s^{\mathrm{NLM}} = \frac{\exp(\mathrm{logit}_s^{\mathrm{NLM}}/\tau_{\mathrm{KD}})}{\sum_{s'=1}^{S} \exp(\mathrm{logit}_{s'}^{\mathrm{NLM}}/\tau_{\mathrm{KD}})} \tag{4}$$

ここで τ_{KD} は温度パラメータである．同じ τ_{KD} を用いることで，教師の擬似確率 $\tilde{p}_s^{\mathrm{NBE}}$ も同様に定義できる．これらを用いると，KL 情報量損失は

$$\mathcal{L}_{\mathrm{KD}} = \tau_{\mathrm{KD}}^2 \sum_{s=1}^{S} \tilde{p}_s^{\mathrm{NBE}} \log \frac{\tilde{p}_s^{\mathrm{NBE}}}{\tilde{p}_s^{\mathrm{NLM}}} \tag{5}$$

と表される．温度パラメータが大きいほど，確率が低い単語候補同士の一致が相対的に重視されるようになる [26].

式（5）をもとに手続きとして書き下すと，アルゴリズム 1 となる．ここではモデル更新の 1 反復の概略を示している．$\mathcal{L}_{\mathrm{WLM}}$ および $\mathcal{L}_{\mathrm{NBE}}$ は 4.4 節で導入する．このアルゴリズムは並列化が困難であることと，メモリ上の制約から，

ミニバッチサイズは 1 とする．メモリ上の制約は CLM の順計算を大量に呼び出すことに由来する．そこで，CLM は反復のたびに更新するのに対して，WLM に対しては勾配累積を採用する．この勾配累積を単位としてステップと呼称することにする．4.4 節で述べるように，NBE のパラメータ更新も行う．

アルゴリズム 1 では PLM によるテキスト生成を用いているが，生コーパスを利用することも可能である．PLM によるランダム生成に頼る実用上の理由は，昨今の巨大事前訓練モデルが訓練に用いたコーパスと一緒に公開されることが少なく，コーパスの確保が手間である一方で，事前訓練モデルから生成されたテキストが十分に自然であることである．各文字列は NBE により確率的に分割される．$\tau_{\mathrm{NBE}} = 0.2$ という値は，ほぼ決定的な振る舞いをしつつ，単語境界確率が 0.5 に近く真に曖昧な場合には確率的な振る舞いを見せることを考慮して決めている．

残る問題は $S-1$ 個の単語候補の生成である．様々な方法が考えられるが，本稿が採用した手法は以下の通りである．まず PLM に $\boldsymbol{c}_{1:n}$ に後続する最大で長さ L の文字列を生成（</s>が選択されると生成を打ち切る）させ，次にその接頭辞群を抽出する．この処理を繰り返し，重複を取り除いて $S-1$ 個の候補が集まるまで続ける．こうして得られた単語候補群は真の単語と偽の単語を含むと同時に，真の単語間でも文脈適合性に差がつく．

4.4 最小エントロピー正則化

単語候補の大半が偽の単語であることを考慮すると，（擬似）確率質量が少数の単語候補に集中するのが望ましい．この目的を促進するために最小エントロピー正則化 [9], [23] を導入する:

$$\mathcal{L}_{\mathrm{WLM}} = \lambda_{\mathrm{WLM}} \sum_{s=1}^{S} \hat{p}_s^{\mathrm{NLM}} \log \hat{p}_s^{\mathrm{NLM}}$$

ここで λ_{WLM} はハイパーパラメータ，$\hat{p}_s^{\mathrm{NLM}}$ は $\tau_{\mathrm{KD}} = 1$ である点を除いて式（4）と同一である．

知識蒸留の一般的な想定では教師のパラメータは固定されるが，生徒が訓練初期段階の無秩序状態を脱した段階で NBE も更新するのが望ましい．しかし，NBE に完全な自由を与えると，訓練が予期せぬ方向に向かう可能性がある．

NBE を現実的は範囲に押しとどめるために，また最小エントロピー正則化を導入する．この正則化は式（3）の第 2，第 3 項に（つまり最後の項を除いて）適用する．こうするのは，次の文字を含む広い文脈を観測している際には，単語境界確率は 0 または 1 に近づくことが好ましいと考えられるからである．

$$\mathcal{L}_{\mathrm{NBE}} = \lambda_{\mathrm{NBE}} \frac{1}{l} \sum_{i=n}^{n+l-1} \sigma(\texttt{---}) \log \sigma(\texttt{---}),$$

表 1 実験で用いたデータセット（各マスは文数を示す）
Table 1 Datasets used in the experiments. Each cell indicates the number of sentences.

コーパス	訓練	テスト
All	50,479	3,978
KC	36,623	1,783

ここで λ_{NBE} はハイパーパラメータ，$\sigma(\texttt{---})$ は式（1）の略記（ただし n を i で置換）である．

5. 実験

5.1 設定

5.1.1 データセット

表 1 に示す通り，以下の 2 種類の日本語コーパスを統合した: （1）京都大学テキストコーパス（**KC**）[35]，および（2）京都大学ウェブ文書リードコーパス（**KWDLC**）[24]．両者は同一の単語分割基準に従っている．KC が教師なし単語分割の先行研究 [42], [56] でも用いられていたことを考慮して，KC 単独での精度についても報告する．ただ，1995 年の古い新聞記事からなる KC と，言語モデルの事前訓練に用いられた比較的新しいウェブテキストの間には埋めがたい差があるように思われることから，内容的に後者に近いタグ付きウェブコーパスも用いることにした．いずれのコーパスについても公式の訓練・テスト分割に従った．ただし，提案手法は訓練データを 2 個のハイパーパラメータの最適化にしか用いていない．訓練手続きの大半は事前訓練文字言語モデル自身がランダム生成したテキストによって行われている．

5.1.2 モデル

PLM としては筆者自身が訓練した日本語文字レベル GPT-2 Large[*3]を用いた．このモデルの諸設定は OpenAI の GPT-2 Large に合わせてあるが，語彙サイズが約 6 千である点が異なる．文字レベルの語彙はバイトへのフォールバック [61] をサポートしており，原理的に未知語は発生しない．モデルと訓練の設定の詳細は付録 A.1 に示す．

5.1.3 ベースライン

提案モデルである **DLM** をいくつかのベースラインと比較した．**ES** の特殊版は訓練データにおける単語 F1 スコアを最大化するようにハイパーパラメータを最適化したものである．また，事前訓練済みの **NBE** に決定的分割を行わせた結果も用いた．精度上界の参考値として，**NBE** の教師あり学習も行った．さらに従来研究で KC に対して報告されているスコアも転記する．

5.1.4 評価尺度

単語分割タスクにおいて標準的な単語レベルの適合率，再現率，F1 スコアを用いる．

[*3] https://huggingface.co/ku-nlp/gpt2-large-japanese-char

表 2　**All** テストデータにおける結果
Table 2　Results for the **All** test data.

モデル	適合率	再現率	F1
ES（単語 F1 に基づく）	62.8	65.9	64.3
ES（NBE 向け）	63.8	62.1	63.0
NBE（事前訓練）	**65.1**	66.4	65.7
DLM	65.0	**68.2**	**66.6**
NBE（教師あり）	92.6	97.7	95.1

表 3　**KC** テストセットにおける結果の既存研究との比較（既存研究は無作為に選択した千文を用いているのに対し，本実験では公式の訓練・テスト分割を用いているため，厳密には比較可能ではない）
Table 3　Comparison with reported scores using the **KC** test set. Note that the scores may not be strictly comparable because others randomly selected 1,000 sentences while we followed the official train/test split.

モデル	F1
DLM (提案手法)	68.1
NPY [42]	62.1
PYHSMM [56]	71.5

5.2　結果

表 2 に主要な結果を示す．提案する多段階の訓練手続きが段階的に精度を改善していることが確認できる．古典的な n-gram 言語モデルと異なり，Transformer に基づく事前訓練ニューラル言語モデルは強力であり，文字系列を進むにつれて劇的にエントロピー値を下げる．そのため，ES はテキストの前半を過分割，後半を過小分割する傾向が見られた．一方，ES の結果に基づいて訓練したにもかかわらず，NBE による分割はテキストの位置による偏りが緩和されていた．こうなった原因は，NBE の表現力が限定的であり，ES を完全再現するようにできていないからかもしれない．その傍証として，NBE の教師あり設定における精度が挙げられる．教師なし設定を大幅に上回っているものの，F1 スコアが 99 を上回ることが普通となっている通常の教師あり単語分割モデルには遠く及ばなかった．

NBE は ES よりも緩和されているとはいえ，その単語分割が生み出す単語は文脈依存性を示しており，文脈に特有の一度きりの単語候補にある程度の確率質量を割く傾向が見られた．このようなエラーの目立つ教師を模倣したにもかかわらず，DLM（のもとになった NLM）はより高い精度を示した．NLM を構成する CLM は文脈非依存な形で単語候補を評価するため，頻繁に登場する単語候補を好む帰納バイアスを有していると推測できる．

表 3 では提案モデルを先行研究の結果と比較している．提案モデルは先行研究が採用していた強力なノンパラメトリックベイズモデルと同等の精度を示していることが確認できる．日本語においては，モデルが返す単語分割の主観的な良さがベンチマーク精度に反映されないという見解が

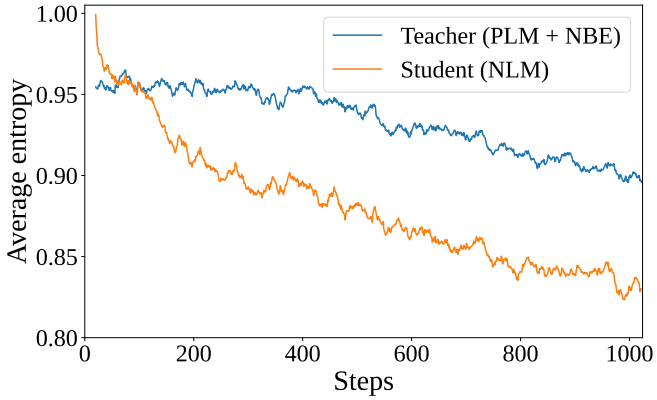

図 3　教師と生徒のエントロピー値の訓練時の軌跡（各文字列上で平均したうえで，ウインドウサイズ 20 で移動平均をとっている）
Fig. 3　Training trajectories of entropy values for both the teacher and student, averaged over each character sequence, with transitions smoothed by a moving average (window size: 20).

繰り返し報告されてきた [42] が，本稿でもその傾向を再確認した．例えば，KC，KWDLC においては動詞の完了を表す「た」は屈折語尾，つまり単語の一部とみなされているが，形と意味の対応が比較的規則的であるためか，統計的モデルは「た」を動詞語幹から切り離す傾向が見られる．

6.　議論

6.1　長さバイアス

提案する訓練手続きの大きな特徴は，次の単語候補を長さにかかわらず比較するという局所化された訓練を行っていることである．しかし，教師なし単語分割における一般的見解に基づくと，比較は与えられた文字列の分割によってなされなければならない．例えば，ある文字列が 1 単語であるか 2 単語であるかを比較したり [22]，1 文を構成するラティスからどの経路が最適かを探索したりする [42], [51] などの方法が必要と考えられてきた．そうでなければ，長いが不自然な単語候補が，単語生成回数が減るために，短い自然な単語よりも優先されがちだからである．同じ問題は教師あり設定も生じており，そこでは**長さバイアス**とよばれている [34]．

本実験でも長さバイアスに遭遇した．知識蒸留の教師は真の単語に対して，より長い偽の単語よりも大きな確率質量を与える傾向が見られたものの，NBE の予測は長さバイアスを克服できるほど決定的ではなく，体系的な過小分割をもたらしていた．

本稿における長さバイアスへの解決策は 4.4 節で導入した最小エントロピー正則化である．図 3 に訓練途中での教師と生徒のエントロピー値に変遷を示す．実験で採用した $\lambda_{\mathrm{WLM}} = 0.05$ のとき，生徒が先にエントロピー値を下げ，続いて教師を引き下げるという結果が得られた．ただしこの振る舞いはハイパーパラメータ λ_{WLM} の値に敏感であっ

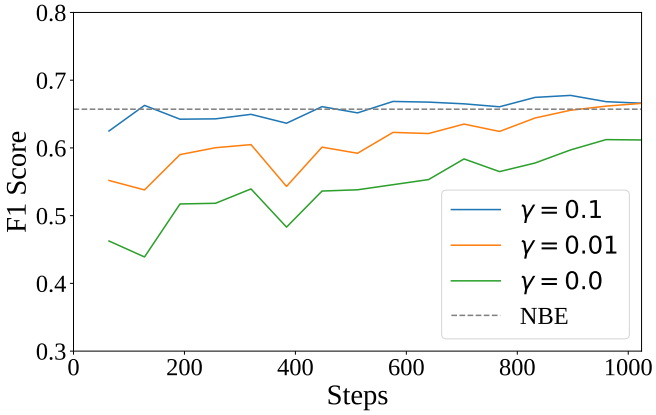

図 4 ビーム探索の局所右枝刈りが F1 スコアに与える影響（相対閾値 γ をデフォルトの 0.01 から変更した場合．事前訓練済み NBE の精度を長さバイアス克服度合いの参考値として示す）

Fig. 4 Effect of local right pruning of beam search on the F1 score with varying relative threshold γ. The default choice was $\gamma = 0.01$. The pretrained NBE serves as a reference point for assessing the extent to which length bias is addressed.

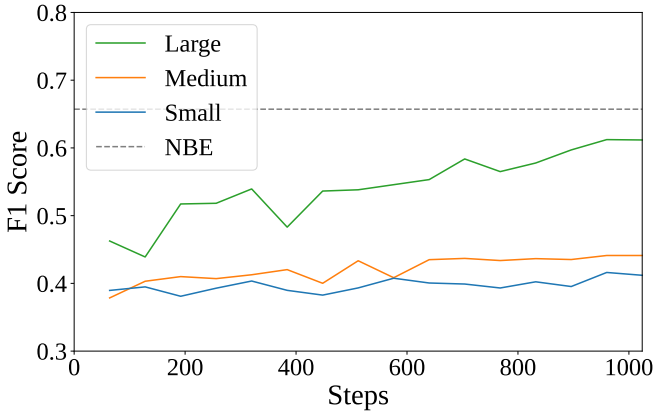

図 5 モデルサイズの F1 スコアへの効果（折れ線はモデルサイズごとのスコア．局所右枝刈りの相対閾値 $\gamma = 0$．事前訓練済み NBE は Large モデルのもの．）

Fig. 5 Effect of model size on the F1 score. The first three plots represent different model sizes. The relative threshold γ of local right pruning was set to zero. The pretrained NBE was based on the Large model.

た．$\lambda_{\mathrm{WLM}} = 0$，つまり正則化を適用しない場合，生徒のエントロピー値は常に教師を上回り，教師を引き上げるさえするという結果に終わった．一方で，$\lambda_{\mathrm{WLM}} = 0.1$ と大きくした場合，生徒は過度に少数の単語候補に確率質量を割り当て，訓練効率が低下した．なお，NBE に対しても別の最小エントロピー正則化を適用しているが，こちらは教師のエントロピー値への影響は小さかった．ただし，教師の擬似確率分布を有意味な範囲に押しとどめるには欠かせないという感触を得た．

本稿の実験の範囲では，残念ながら最小エントロピー正則化だけでは長さバイアスへの対策としては十分ではなく，ビーム探索における局所右枝刈り（図 2）に補完的な役割を委ねることになった．この手法は，次の単語候補集合のうち，最大確率の γ 倍を下回る確率を持つ候補を枝刈りする．したがって，長いが不自然な単語候補は，最終的に累積確率において短い対抗馬を上回ることになる場合でも効果的に除外されることになる．図 4 に示すように，訓練は長さバイアスを克服する方向に向かっていたが，1,024 ステップ時点では枝刈りしない（$\gamma = 0$）設定のモデルは依然として NBE ベースラインを下回ってしまった．

興味深いことに，長さバイアスを克服する能力は創発的能力 [49], [65] のように見える．図 5 は，デフォルトの Large（717M）モデルを Medium（310M）[*4] および Small（90M）[*5] モデルと比較している．こうした比較的小規模なモデルについても訓練にともない精度が上昇する傾向が確認できるが，NBE ベースラインを依然として大幅に下回っている．Medium と Large の間の大きな落差を考慮すると，より大きい XL モデルがあれば，枝刈りなしで長さバイアスを克服するのではないかと期待できる．ただし，そのようなモデルは未発表であり，また，提案する訓練手続きを適用するのは計算量的に難しい．

[*4] https://huggingface.co/ku-nlp/gpt2-medium-japanese-char
[*5] https://huggingface.co/ku-nlp/gpt2-small-japanese-char

6.2 言語獲得への示唆

実験心理学では幼児が発達段階に応じて単語分割に使う手がかりを変えていることが示唆されている [53]．しかし，そのような遷移を再現するような計算モデルがどのようなものかは不明なままである．本稿での提案は，単語境界認定に基づく教師から単語系列に基づく生徒への知識蒸留という定式化でこれを実現するものであり，その潜在的可能性を実験的に検証した．脳内の言語処理モジュールが一夜にして 3 個に複製されるとは考え難いが，提案する訓練手続きが脳内の再配線機構を大雑把に，かつ非効率的に近似できているとしたら興味深い．また，提案手法は結果的に教師なし単語分割の研究史 [22], [25] をなぞる形にもなっている．

提案手法ではモデル自身が内部生成する空想データによって訓練される．この仕組みは，wake–sleep アルゴリズム [27] における sleep 段階を思い起こさせる．sleep 段階では，生成器が生成した空想データによって認識器が訓練される．ただし，wake–sleep アルゴリズムでは生成器と認識器が表層構造と潜在構造との間の循環を形成するのに対して，提案手法で生成器と言える PLM は潜在構造を扱わず，認識器と言える NLM と NBE の協調によって両者が訓練される．さらに，単語候補の生成に関しては，空想データに対する代替案を検討するという点で二重に反実仮想的なデータを生成している．このように提案手法と wake–sleep

アルゴリズムには様々な違いがあり，後者にかかわる議論をそのまま転用はできないが，ニューラルネットが自身の作る夢を通じて自己組織化するという点が興味深い．

本稿の取り組みは主に幼児の言語獲得に関する研究に着想を得たものだが，実験では日本語という書記言語を対象とせざるを得なかった．同様の指向を持つ従来研究は CHILDES コーパス [8], [39] を使う傾向にあった．CHILDES コーパスは幼児に向けられた発話を収集したものだが，幼児が言語獲得にあたって受ける刺激の総量と比較してもあまりに小さい．言語獲得の研究は刺激の貧困の議論の強い影響下にあるが，近年の LLM の成功を考慮すると，巨大モデルにおける実証を優先し，そのあとで小規模化 [63] を試みるという方向性が有望と思える．

7. 結論

本稿では事前訓練文字言語モデルから単語言語モデルへの教師なし合成の実現可能性を示した．管見の限り，本稿は Transformer に基づく（因果）言語モデルを教師なし単語分割に適用した最初の事例である．提案モデルと訓練手続きの組合せは複雑だが，結論としては以下の 3 種類の帰納バイアスを盛り込んでいることになる．(1) 取っ掛かりとしての単語境界における確率的不確実性，(2) 単語の意味が文脈非依存な形である程度決まるという単語の自律性，および (3) 真の単語候補と偽の単語候補を含む確率分布における疎性．本稿では，十分とは言えないものの，提案する多段階の訓練手続きが段階的に精度を改善することを示した．

謝辞 本研究は一部科研費 24K15068 の支援を受けた．

参考文献

[1] Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D. R., Smith, N. A. and Tsvetkov, Y.: Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models, arXiv:2305.13707 (2023).

[2] Aho, A. V. and Corasick, M. J.: Efficient string matching: an aid to bibliographic search, *Communications of the ACM*, Vol. 18, No. 6, pp. 333—340 (online), DOI: 10.1145/360825.360855 (1975).

[3] Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y. and Gong, B.: VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text, *Advances in Neural Information Processing Systems* (Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. and Vaughan, J. W., eds.), Vol. 34, Curran Associates, Inc., pp. 24206–24221 (2021).

[4] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (online), DOI: 10.1145/3292500.3330701 (2019).

[5] Ando, R. and Lee, L.: Unsupervised Statistical Segmentation of Japanese Kanji Strings, Technical report, Cornell University (1999).

[6] Batchelder, E. O.: Bootstrapping the lexicon: A computational model of infant speech segmentation, *Cognition*, Vol. 83, No. 2, pp. 167–206 (online), DOI: 10.1016/S0010-0277(02)00002-1 (2002).

[7] Bengio, Y., Ducharme, R. and Vincent, P.: A Neural Probabilistic Language Model, *Advances in Neural Information Processing Systems* (Leen, T., Dietterich, T. and Tresp, V., eds.), Vol. 13, MIT Press (2000).

[8] Bernstein-Ratner, N.: The phonology of Parent-Child Speech, *Children's Language*, Psychology Press, pp. 159–174 (1987).

[9] Brand, M.: Structure Learning in Conditional Probability Models via an Entropic Prior and Parameter Extinction, *Neural Computation*, Vol. 11, No. 5, pp. 1155–1182 (online), DOI: 10.1162/089976699300016395 (1999).

[10] Brent, M. R.: An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery, *Machine Learning*, Vol. 34, pp. 71–105 (online), DOI: 10.1023/A:1007541817488 (1999).

[11] Brent, M. R. and Cartwright, T. A.: Distributional regularity and phonotactic constraints are useful for segmentation, *Cognition*, Vol. 61, No. 1, pp. 93–125 (online), DOI: 10.1016/S0010-0277(96)00719-6 (1996).

[12] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D.: Language Models are Few-Shot Learners, arXiv:2005.14165 (2020).

[13] Cai, Z. G., Haslett, D. A., Duan, X., Wang, S. and Pickering, M. J.: Does ChatGPT resemble humans in language use?, arXiv:2303.08014 (2023).

[14] Chen, M., Chang, B. and Pei, W.: A Joint Model for Unsupervised Chinese Word Segmentation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Moschitti, A., Pang, B. and Daelemans, W., eds.), Doha, Qatar, Association for Computational Linguistics, pp. 854–863 (online), DOI: 10.3115/v1/D14-1092 (2014).

[15] Christiansen, M. H., Allen, J. and Seidenberg, M. S.: Learning to Segment Speech Using Multiple Cues: A Connectionist Model, *Language and Cognitive Processes*, Vol. 13, No. 2-3, pp. 221–268 (online), DOI: 10.1080/016909698386528 (1998).

[16] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *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)* (Burstein, J., Doran, C. and Solorio, T., eds.), Minneapolis, Minnesota, Association for Computational Linguistics, pp. 4171–4186 (online), DOI: 10.18653/v1/N19-1423 (2019).

[17] Downey, C., Xia, F., Levow, G.-A. and Steinert-Threlkeld, S.: A Masked Segmental Language Model for Unsupervised Natural Language Segmentation, *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (Nicolai, G. and Chodroff, E., eds.), Seattle, Washington, Association for Computational Linguistics, pp. 39–50 (online), DOI: 10.18653/v1/2022.sigmorphon-1.5 (2022).

[18] Elman, J. L.: Finding Structure in Time, *Cognitive Science*, Vol. 14, No. 2, pp. 179–211 (online), DOI: https://doi.org/10.1207/s15516709cog1402_1 (1990).

[19] Feng, H., Chen, K., Deng, X. and Zheng, W.: Accessor Variety Criteria for Chinese Word Extraction, *Computational Linguistics*, Vol. 30, No. 1, pp. 75–93 (online), DOI: 10.1162/089120104773633394 (2004).

[20] Freitag, M. and Al-Onaizan, Y.: Beam Search Strategies for Neural Machine Translation, *Proceedings of the First Workshop on Neural Machine Translation* (Luong, T., Birch, A., Neubig, G. and Finch, A., eds.), Vancouver, Association for Computational Linguistics, pp. 56–60 (online), DOI: 10.18653/v1/W17-3207 (2017).

[21] Gammon, E.: Quantitative Approximations to the Word, *International Conference on Computational Linguistics COLING 1969: Preprint No. 12*, Sånga Säby, Sweden, (online), available from ⟨https://aclanthology.org/C69-1201⟩ (1969).

[22] Goldwater, S., Griffiths, T. L. and Johnson, M.: A Bayesian framework for word segmentation: Exploring the effects of context, *Cognition*, Vol. 112, No. 1, pp. 21–54 (online), DOI: 10.1016/j.cognition.2009.03.008 (2009).

[23] Grandvalet, Y. and Bengio, Y.: Semi-supervised Learning by Entropy Minimization, *Advances in Neural Information Processing Systems* (Saul, L., Weiss, Y. and Bottou, L., eds.), Vol. 17, MIT Press (2004).

[24] Hangyo, M., Kawahara, D. and Kurohashi, S.: Building a Diverse Document Leads Corpus Annotated with Semantic Relations, *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation* (Manurung, R. and Bond, F., eds.), Bali, Indonesia, Faculty of Computer Science, Universitas Indonesia, pp. 535–544 (online), available from ⟨https://aclanthology.org/Y12-1058⟩ (2012).

[25] Harris, Z. S.: From Phoneme to Morpheme, *Language*, Vol. 31, No. 2, pp. 190–222 (online), DOI: 10.2307/411036 (1955).

[26] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *NIPS Deep Learning and Representation Learning Workshop*, (online), available from ⟨http://arxiv.org/abs/1503.02531⟩ (2015).

[27] Hinton, G. E., Dayan, P., Frey, B. J. and Neal, R. M.: The "Wake-Sleep" Algorithm for Unsupervised Neural Networks, *Science*, Vol. 268, No. 5214, pp. 1158–1161 (online), DOI: 10.1126/science.7761831 (1995).

[28] Hutchens, J. L. and Alder, M. D.: Finding Structure via Compression, *New Methods in Language Processing and Computational Natural Language Learning*, (online), available from ⟨https://aclanthology.org/W98-1210⟩ (1998).

[29] Johnson, E. K. and Jusczyk, P. W.: Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics, *Journal of Memory and Language*, Vol. 44, No. 4, pp. 548–567 (online), DOI: 10.1006/jmla.2000.2755 (2001).

[30] Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y. and Radev, D.: Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations, arXiv:2303.18027 (2023).

[31] Kawakami, K., Dyer, C. and Blunsom, P.: Learning to Discover, Ground and Use Words with Segmental Neural Language Models, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Korhonen, A., Traum, D. and Màrquez, L., eds.), Florence, Italy, Association for Computational Linguistics, pp. 6429–6441 (online), DOI: 10.18653/v1/P19-1645 (2019).

[32] Kempe, A.: Experiments in Unsupervised Entropy-Based Corpus Segmentation, *EACL 1999: CoNLL-99 Computational Natural Language Learning*, (online), available from ⟨https://aclanthology.org/W99-0702⟩ (1999).

[33] Kim, T., Oh, J., Kim, N. Y., Cho, S. and Yun, S.-Y.: Comparing Kullback-Leibler Divergence and Mean Squared Error Loss in Knowledge Distillation, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (Zhou, Z.-H., ed.), International Joint Conferences on Artificial Intelligence Organization, pp. 2628–2635 (online), DOI: 10.24963/ijcai.2021/362 (2021).

[34] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (Lin, D. and Wu, D., eds.), Barcelona, Spain, Association for Computational Linguistics, pp. 230–237 (online), available from ⟨https://aclanthology.org/W04-3230⟩ (2004).

[35] Kurohashi, S. and Nagao, M.: Building a Japanese Parsed Corpus while Improving the Parsing System, *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC-98)*, pp. 719–724 (1998).

[36] Li, H.-W., Lin, Y.-J., Li, Y.-T., Lin, C. and Kao, H.-Y.: Improved Unsupervised Chinese Word Segmentation Using Pre-trained Knowledge and Pseudo-labeling Transfer, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Bouamor, H., Pino, J. and Bali, K., eds.), Singapore, Association for Computational Linguistics, pp. 9109–9118 (online), DOI: 10.18653/v1/2023.emnlp-main.564 (2023).

[37] Li, W., Song, Y., Su, Q. and Shao, Y.: Unsupervised Chinese Word Segmentation with BERT Oriented Probing and Transformation, *Findings of the Association for Computational Linguistics: ACL 2022* (Muresan, S., Nakov, P. and Villavicencio, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 3935–3940 (online), DOI: 10.18653/v1/2022.findings-acl.310 (2022).

[38] Lu, J., Batra, D., Parikh, D. and Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, *Advances in Neural Information Processing Systems* (Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R., eds.), Vol. 32, Curran Associates, Inc. (2019).

[39] MacWhinney, B. and Snow, C.: The child language data exchange system, *Journal of Child Language*, Vol. 12, No. 2, pp. 271–295 (online), DOI: 10.1017/S0305000900006449 (1985).

[40] Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B. and Tan, S.: Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP, arXiv:2112.10508 (2021).

[41] Mielke, S. J. and Eisner, J.: Spell Once, Summon Anywhere: A Two-Level Open-Vocabulary Language Model, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pp. 6843–6850 (online), DOI: 10.1609/aaai.v33i01.33016843 (2019).

[42] Mochihashi, D., Yamada, T. and Ueda, N.: Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor

Language Modeling, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Su, K.-Y., Su, J., Wiebe, J. and Li, H., eds.), Suntec, Singapore, Association for Computational Linguistics, pp. 100–108 (online), available from ⟨https://aclanthology.org/P09-1012⟩ (2009).

[43] Nguyen, M. V., Lai, V. D., Pouran Ben Veyseh, A. and Nguyen, T. H.: Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (Gkatzia, D. and Seddah, D., eds.), Online, Association for Computational Linguistics, pp. 80–90 (online), DOI: 10.18653/v1/2021.eacl-demos.10 (2021).

[44] OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W. and Zoph, B.: GPT-4 Technical Report, arXiv:2303.08774 (2023).

[45] Press, O. and Wolf, L.: Using the Output Embedding to Improve Language Models, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Lapata, M., Blunsom, P. and Koller, A., eds.), Valencia, Spain, Association for Computational Linguistics, pp. 157–163 (online), available from ⟨https://aclanthology.org/E17-2025⟩ (2017).

[46] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I.: Language Models are Unsupervised Multitask Learners (2019).

[47] Saffran, J. R., Newport, E. L. and Aslin, R. N.: Word Segmentation: The Role of Distributional Cues, *Journal of Memory and Language*, Vol. 35, No. 4, pp. 606–621 (online), DOI: 10.1006/jmla.1996.0032 (1996).

[48] Sanh, V., Debut, L., Chaumond, J. and Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv:1910.01108 (2020).

[49] Schaeffer, R., Miranda, B. and Koyejo, S.: Are Emergent Abilities of Large Language Models a Mirage?, arXiv:2304.15004 (2023).

[50] Scharenborg, O., Besacier, L., Black, A., Hasegawa-Johnson, M., Metze, F., Neubig, G., Stüker, S., Godard, P., Müller, M., Ondel, L., Palaskar, S., Arthur, P., Ciannella, F., Du, M., Larsen, E., Merkx, D., Riad, R., Wang, L. and Dupoux, E.: Linguistic Unit Discovery from Multi-Modal Inputs in Unwritten Languages: Summary of the "Speaking Rosetta" JSALT 2017 Workshop, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4979–4983 (online), DOI: 10.1109/ICASSP.2018.8461761 (2018).

[51] Sun, Z. and Deng, Z.-H.: Unsupervised Neural Word Segmentation for Chinese via Segmental Language Modeling, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Riloff, E., Chiang, D., Hockenmaier, J. and Tsujii, J., eds.), Brussels, Belgium, Association for Computational Linguistics, pp. 4915–4920 (online), DOI: 10.18653/v1/D18-1531 (2018).

[52] Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S., Das, D. and Pavlick, E.: What do you learn from context? Probing for sentence structure in contextualized word representations, *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=SJzSgnRcKX⟩ (2019).

[53] Thiessen, E. D. and Saffran, J. R.: When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants, *Developmental Psychology*, Vol. 39, No. 4, pp. 706–716 (online), DOI: 10.1037/0012-1649.39.4.706 (2003).

[54] Tolmachev, A., Kawahara, D. and Kurohashi, S.: De-

sign and Structure of The Juman++ Morphological Analyzer Toolkit, *Journal of Natural Language Processing*, Vol. 27, No. 1, pp. 89–132 (online), DOI: 10.5715/jnlp.27.89 (2020).
[55] Tung, C.-H. and Lee, H.-J.: Identification of Unknown Words from a Corpus, *Computer Processing of Chinese & Oriental Languages*, Vol. 8, pp. 131–145 (1994).
[56] Uchiumi, K., Tsukahara, H. and Mochihashi, D.: Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Zong, C. and Strube, M., eds.), Beijing, China, Association for Computational Linguistics, pp. 1774–1782 (online), DOI: 10.3115/v1/P15-1171 (2015).
[57] Ueda, N., Omura, K., Kodama, T., Kiyomaru, H., Murawaki, Y., Kawahara, D. and Kurohashi, S.: KWJA: A Unified Japanese Analyzer Based on Foundation Models, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (Bollegala, D., Huang, R. and Ritter, A., eds.), Toronto, Canada, Association for Computational Linguistics, pp. 538–548 (online), DOI: 10.18653/v1/2023.acl-demo.52 (2023).
[58] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Vol. 30, Curran Associates, Inc. (2017).
[59] Venkataraman, A.: A Statistical Model for Word Discovery in Transcribed Speech, *Computational Linguistics*, Vol. 27, No. 3, pp. 351–372 (online), DOI: 10.1162/089120101317066113 (2001).
[60] 和田有輝也，村脇有吾，黒橋禎夫：セミマルコフ CRF 自己符号化器による教師なし単語分割，言語処理学会第 28 回年次大会 発表論文集，pp. 806–810 (2022).
[61] Wang, C., Cho, K. and Gu, J.: Neural Machine Translation with Byte-Level Subwords, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 9154–9160 (online), DOI: 10.1609/aaai.v34i05.6451 (2020).
[62] Wang, L. and Zheng, X.: Unsupervised Word Segmentation with Bi-directional Neural Language Model, *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 22, No. 1 (online), DOI: 10.1145/3529387 (2022).
[63] Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T. and Cotterell, R.: Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T. and Cotterell, R., eds.), Singapore, Association for Computational Linguistics, pp. 1–34 (online), DOI: 10.18653/v1/2023.conll-babylm.1 (2023).
[64] Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F. and Bowman, S. R.: BLiMP: The Benchmark of Linguistic Minimal Pairs for English, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 377–392 (online), available from ⟨https://aclanthology.org/2020.tacl-1.25⟩ (2020).
[65] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J. and Fedus, W.: Emergent Abilities of Large Language Models, *Transactions on Machine Learning Research*, (online), available from ⟨https://openreview.net/forum?id=yzkSU5zdwD⟩ (2022).
[66] Yang, H.: BERT Meets Chinese Word Segmentation, arXiv:1909.09292 (2019).
[67] Zhao, H. and Kit, C.: An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework, *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, (online), available from ⟨https://aclanthology.org/I08-1002⟩ (2008).

付　　録

表 A·1 日本語文字レベル GPT-2 Large のモデル仕様
Table A·1 Model specifications of Japanese character-level GPT-2 Large.

パラメータ数	717M
アーキテクチャ	GPT-2
埋め込みサイズ	1,280
層の数	36
ヘッドの数	20
語彙サイズ	6K

A.1　モデル訓練の詳細

表 A·1 に日本語文字レベル GPT-2 Large のモデル仕様の概要を示す．比較のために Medium（310M）および Small（90M）モデルも用いたが，いずれも語彙サイズを除いて OpenAI の同名のモデルと仕様を揃えてある．

表 A·2 に一連のモデルのハイパーパラメータを示す．ES のハイパーパラメータ最適化には Optuna [4] を用いた．最適化の目的関数は境界認識の F スコアとした．ただし，$\beta = 0.7$ とし，適合率を優先することで，過分割よりも過小分割が起きるようにした．

単語の最大文字数は 7 とした．この値で訓練データにおける単語トークンの 99.92%が被覆される．

NLM の訓練に用いたハイパーパラメータは訓練を安定させるという観点から試行錯誤により選んだ．単語の最大長は短いが，CLM を膨大な回数呼び出すため，提案手法はメモリ消費が激しい．そこで，A100 80GB GPU カードを 2 枚用い，PyTorch の分散データ並列機能を用いてメモリ消費を分散させた．さらに，メモリあふれを防ぐために，CLM の呼び出しが規定回数を超えた時点で文字系列に対するループを打ち切った．

DLM のための語彙選択は以下の手順で行った．まず，事前訓練文字言語モデルの文字語彙（バイトを含む）をす

べて採用することで未知語が生じないようにした．次に，訓練中に勘定した単語候補の頻度をもとに，規定の語彙サイズに達するまで上位の単語候補を採用した．ここで用いた頻度は，$\hat{p}_s^{\mathrm{NLM}} > 0.001$ の累積頻度である．

計算資源上の制約から，各モデル設定について，1 回の試行のみを行った結果を報告している．Large モデルについては，1 ステップに約 15 分を要した．

A.2 本稿の限界

教師なし単語分割の研究の多くは科学的興味によって駆動されるものであり，短期的な実応用先を求めるのは難しい．ただし，1 節でも述べた通り，テキストから連続音声への拡張は膨大な潜在的応用先を持つ．連続音声からの語彙獲得の現在の研究は，音素獲得と同時に行われるなど，より低次の言語構造への対応に追われている．このことを考慮すると，現時点ではテキストを起点とした研究を進めておくのが有望と考えている．

提案手法は Transformer に基づく事前訓練文字言語モデルの存在に依存しており，しかも，実験結果はより大きなモデルの必要性を示唆している．しかしそのような公開モデルは稀であり，自分で訓練する場合に要するコストは馬鹿にできない．本稿の実験が日本語に限られる理由はこれが大きい．

付録 A.1 に示すように，提案手法は大量のハイパーパラメータを持っている．教師なしタスクにおいてはそもそもハイパーパラメータを求める適正な手続きを見出すのが難しい．本稿の実験で採用したハイパーパラメータは主に訓練を安定させるという観点から，損失関数やその他の指標を観測することで選ばれたものであり，（2 個の例外を除いて）教師データに基づくものではない．とはいえ，計算資源上の制約からハイパーパラメータ探索を十分に行えたとは言えず，より良いハイパーパラメータ設定が存在する可能性は高い．

表 A·2 一連のモデルのハイパーパラメータ（各ブロックは 4 段階の訓練の各段階に対応）
Table A·2 Hyperparameters of the series of models. Each block corresponds to one of the four stages of training.

ES のハイパーパラメータ最適化の目的関数 n	境界推定の $F_{0.7}$
α および β の推定結果	0.732 および 0.299
NBE の事前訓練のステップ数	50
ミニバッチサイズ	512
最適化器	Adam（学習率は 0.001）
文字系列の最大長	120
NLM の訓練のステップ数	1024
NBE のパラメータの初期凍結ステップ数	256
勾配累積のステップ数	32
単語の最大文字数 L	7
単語候補数 S	32
知識蒸留の温度パラメータ τ_{KD}	4
λ_{WLM}	0.05
λ_{NBE}	1.0
CLM 用の最適化器	Adam（学習率は 2^{-4}）
WLM 用の最適化器	Adam（学習率は 2^{-4}）
NBE 用の最適化器	Adam（学習率は 1^{-5}）
文字系列の最大長	120
CLM 呼び出し回数の限度	$1,920$
DLM の語彙サイズ	100K
大域左枝刈り	上位 $K = 1$
局所右枝刈り	相対閾値 $\gamma = 0.01$