文字の並びの統計的規則性が引き起こす単語らしさの心理的印象

鳥居 拓馬
東京電機大学 理工学部

概要

私たちは自分が未知の単語でもそれがその言語体系の中でありそうか・なさそうか（文法性）を，既知の単語の知識から概ね判断できる．本研究では，Wikipedia コーパスで訓練した高次 n-gram モデルから生成した英語の単語もどき（文字列）に対して人間が「単語らしい」と感じるかを調査した．その結果，人間の単語らしさの主観評定は対数尤度÷文字列長と強い正の相関を示した．形態論（字の並び）は統語論（語の並び）と比べると「語彙爆発」が抑えられ，生成系（言語知識）と生成物（言語資料）とを双方向に結ぶ理論の探究や実証に適する．

1 単語もどきの文法性判断

私たちは自分が未知の単語でもそれがその言語体系の中でありそうか・なさそうかを，既知の単語の知識から概ね判断できる．英語の例では tittle はあるが tittje はない．自らの言語知識（文法規則）に照らして，ある語の並びがその言語の正しい文か非文か（文法規則を充たすか否か）を判定することを言語学では文法性判断という．多くの現存のテキストが正しい文（正例）だと考えると，非文（負例）かを教示する文法性判断は言語学では特別な役割をもつ．離散無限性とも称されるように，ヒトの言語は膨大な語彙をもち，また一文を成す語の個数には原理的に上限がない．この爆発のため，理論的な高次 n-gram モデルは実践的には推定困難に陥る．

本研究では，ヒトの言語知識（文法規則）を解明したい動機のもと，統語論（一文を成す語の並び）ではなく，形態論（一語を成す字の並び）の文法性判断を研究する．なぜなら，英語の字母は 26 個で，一語の字数は大半が短く，語の出現頻度はテキストから推定しやすいため，高次 n-gram が愚直に推定可能である．文⇔語の対応関係の模型として，語⇔字の対応関係を解明することは，言語知識の本質に迫る一歩となろう．本研究では，形態論の文法性判断を「単語らしさ」の主観評価とも呼ぶ．

本研究の目的は，質問紙調査を通して，人間がどのような文字列に対して「単語らしい」と感じるかを明らかにすることである．

人間は新奇の文字列に対しても単語らしさを判断できることから，何かしらの汎化能力を有する．その汎化能力が言語知識を行使した結果だとしたら，論点は形態論の言語知識に関する仮説となろう．本稿ではふたつの仮説を検討する．

第一の仮説は，人間は（高次 n-gram のような）近隣の字の並びを統べる条件付確率の構造を獲得している．この仮説が導く予測としては，ある人の単語らしさの主観評価はその人の確率モデルがその文字列を生起する確率（尤度）と（正）相関する．

第二の仮説は，人間は単語の文字列を計量空間に配置して記憶している．この仮説が導く予測としては，ある人の単語らしさの主観評価はその文字列がその人の既知の単語とどれくらい字面が近いか（距離）と（負）相関する．

ふたつの仮説を対比すると，第一の仮説はマルコフ過程（**生成系**）の構造が，第二の仮説は字面（**生成物**）の共起的類似性が強調されている．しかしどちらも実在する短い文字列との一致・乖離を測ることから，実在する語の文字列に対してこれらの指標は最大・最小となり，対立仮説の関係ではない．むしろ形式言語理論で見られる同じ対象の異なる特徴づけかもしれない．

既存研究[2,3,4,5]では音韻的な近さやパタンの影響が研究されている．本研究のひとつの独自性は，語の統計や高次 n-gram モデルを用い，単語もどきを系統的に生成して調査した点だと考えている．

2 質問紙設計

単語もどきを自然的・系統的に生成するため，確率モデルを構築した．English Wikipedia 20171001 コーパスに含まれる記事の本文を通してすべての単語の出現頻度を数えたのち，4 文字以上 20 文字以下の語の中で出現頻度上位 100,000 語を抽出した．本稿では，この集合に属する語（文字列）を正しい語とした．ある語の長さ ℓ，第 t 番目の字 a_t，終端記号 \$ と書く．それぞれの語 $a_1 a_2 \ldots a_\ell \$$ の出現

確率から p 次の同時確率 $P(t, a_{t-p+1}, \ldots, a_{t-1}, a_t)$ を推定した．遷移確率 $P(a_t | t, a_{t-p+1}, \ldots, a_{t-1})$ を求めた．事前確率 $P(a_1, \ldots, a_{p-1})$ は $q \geq p$ を充たす q 次の同時確率から求めた（一通りでない）．

文字列の生成では事前分布から初期状態を決め，終端記号が選ばれるまで状態遷移を反復する．この確率モデルは次数を制限するため，語彙集合に属さない文字列が生成されうる．ただし，この確率モデルから無作為に生成される文字列は必然的に文字列長の短いものが多くなる．そこで質問紙に含める文字列の選定においては，その質問紙に表れる文字列の文字列長の頻度分布が，語彙集合に表れる語の文字列長の頻度分布と概形が近づくように作為的な抽出を行った．質問紙の文字列を集めると，文字列長の意味でも自然な頻度を反映している．

質問紙調査では，確率モデル (p, q) が生成した文字列のうち語彙集合に属さない文字列（非語＝単語もどき）を採用した．これに加え，日本の高校英単語から無作為に抽出して加え，英語語彙力と指示操作チェックとした．

確率モデルは，文字列を左から右へ処理する順向き，右から左へ処理する逆向きを，それぞれ $1 \leq p \leq q \leq 4$ を充たす 10 種ずつを用意した．高校英単語を加えると，21 種の文字列の集合となる．21 種×40 個＝840 個の文字列を用意した．これを 21 種×5 個＝105 個の文字列を含む質問紙 8 枚に分けた．回答者は最大で 4 枚の質問紙に回答した．

質問文は「以下の文字の並びが英単語として存在しそうだと，どれくらい思いますか」，選択肢は「まるでなさそう」「なさそう」「あまりなさそう」「どちらでもない」「すこしありそう」「ありそう」「とてもありそう」の 7 段階評価とした．

某大学 1 年生を対象とし，2023 年 7 月 4 日～11 日を調査期間とした．回答者数は 244 名だった．多くが日本人のため，高校英単語の語彙力で篩に掛けた．ひとつの質問紙でも，高校英単語をほぼ検出できている，$p = 1$の確率モデルが生成した文字列に対して異常なほど「ありそう」と評価していない，これらの条件を充たさなかった回答者を除外した．最終的に 85 名分の回答をえた（高校英単語条件で半数近く脱落した）．

3 尤度仮説：分析と結果

本研究では前述の通り，高校英単語を除いて，質問紙調査で用いた文字列はすべて English Wikipedia に出現する上位 100,000 語に含まれないという意味で，正しい語ではないとした．もとの言語データの中でその生起確率は 0 である．しかしながら，どの文字列もある確率モデルから抽出したものでそのモデルの上では尤度が定まる．第一の仮説からの予測では，人間の単語らしさの評定はこの尤度と正の相関をもつ．

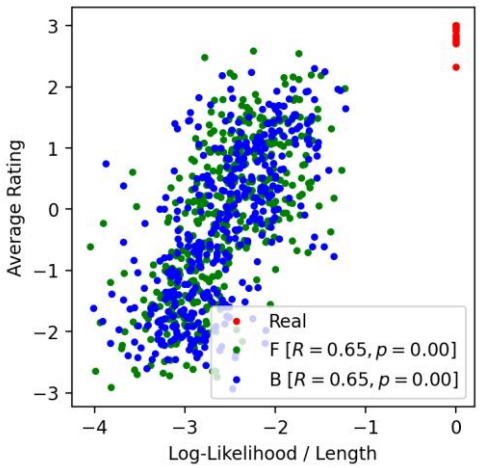

図 1．対数尤度と平均評定
高校英単語 Real，左から右 F，右から左 B
-3 なさそう～ありそう +3

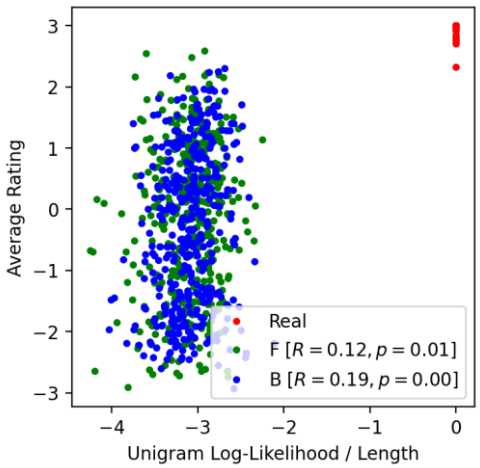

図 2．0 次モデル対数尤度と平均評定

対数尤度は文字列長が長くなるほど小さくなる傾向にあるため文字列長で割引した．割引前は文字列長と対数尤度の相関係数 $R = -0.79$ だったが，割引後は $R = 0.09$ となった．

図 1 は，文字列 840 個について（全被験者の）平均評定と対数尤度÷文字列長の関係を示す．相関係数 $R = 0.65$，$p < 0.01$ だった．

図 2 は，文字列 840 個についてそれを生成した真

の確率モデルではなく $(p,q)=(1,1)$ モデルで計算した対数尤度÷文字列長との関係を示す．これは字の並びを無視して字の出現頻度のみで評価した場合に相当する．相関係数 $R=0.15, p<0.01$ だった．このことから，人間は近隣の字の並びを考慮していると考えられる．

4 距離仮説：分析と結果

次に，第二の仮説の予測を検証する．第二の仮説はその人の既知の語彙との知覚的類似性で新奇な文字列を評価するものだが，本研究では回答者の多くが大学 1 年生だったことから，中学英単語と高校英単語を既知の語彙とし，ある文字列と既知の語彙の間の距離を，その文字列とその語彙の語の間の編集距離の最小値と定義した．

編集距離もまた文字列長が長くなるほど大きくなる傾向にあるため文字列長で割引した．割引前は文字列長と編集距離の相関係数 $R=0.87$ だったが，割引後は $R=0.38$ となった．対数尤度に比べると文字列長の影響を相殺できていない．

図 3 は，文字列 840 個について（全被験者の）平均評定と編集距離÷文字列長の関係を示す．相関係数 $R=-0.57, p<0.01$ だった．

既知の語彙と知覚的類似性が編集距離の意味で高いときほど，単語らしさの主観評価は高くなる傾向が示唆される．以上の結果から，対数尤度と編集距離は強く相関すると予想される．対数尤度÷文字列長と編集距離÷文字列長の相関係数 $R=-0.47, p<0.01$ だった．

5 議論

本研究では，統語論（一文を成す語の並び）ではなく，形態論（一語を成す字の並び）の文法性判断を研究した．質問紙調査の結果，n-gram モデルが生成した単語もどき（新奇な文字列）で，その対数尤度と単語らしさの主観評定は強い正の相関を示した．人間は文字列の中の文字の並びに関する文法規則を構築している可能性が示唆される．

このことはまた，この高次 n-gram モデルが人間の文法性判断に適合しうる形態論（字の並び）に関する文法規則を獲得していることを示唆する．本研究で 3 次か 4 次のマルコフ連鎖（テンソル）でも形態論の文法規則が捉えれていることが示唆された．これは統語論（語の並び）の語彙の数が数百万個に

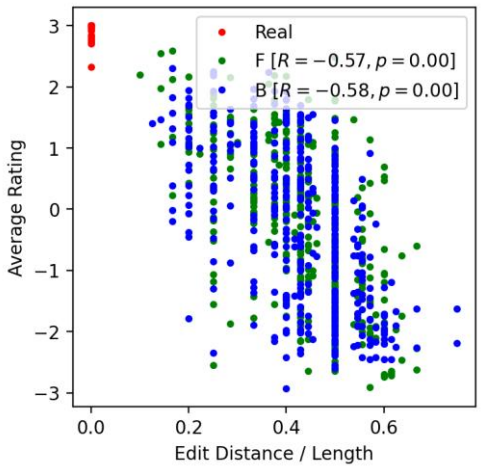

図 3．編集距離と平均評定

も上る言語モデルと比較すると遥かに小さい．高次 n-gram が容易に推定可能と見込まれ，生成系（マルコフ連鎖）と生成物（文字列）の間の対応を具体的に探究・実証する「ミニチュア」となりえる．本研究で行った単語らしさの印象調査によれば，生成系と生成物どちらの側面の指標からも単語らしさの文法性判断を特徴づけている．いま分布仮説[1]を提唱した Zellig S. Harris が探究した形態論に立ち返ることで見えてくる言語構造の本質もあるのではないだろうか．

謝辞

本研究は JSPS 科研費 22K17966 の助成を受けたものです．編集距離の分析は学部生の玉置涼祐くんに協力していただきました．

参考文献

[1] Zellig S. Harris. Distributional hypothesis. Word, 10(2-3), 146-162, 1954.

[2] Todd M. Bailey, Ulrike Hahn. Determinants of wordlikeness: phonotactic or lexical neighborhoods? Journal of Memory and Language, 44, 568-591, 2001

[3] Stefan A. Frisch, Nathan R. Large, David B. Pisoni. Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. Journal of Memory and Language. 42, 481-496, 2000

[4] 深澤はるか・北原真冬．日本語の語彙層と単語らしさの関係について．文法と音声 4

[5] 川上正浩．単語及び非単語の単語らしさ評定に neighbor 数が及ぼす効果．大阪樟蔭女子大学紀要, 2017