E-032

Web 日本語 N グラムを用いた高頻度連鎖語表現の選定
Selection of Multi-Word Expressions from Google N-gram Corpus for Speech Recognition

高橋 伸弥† 森元 逞†
Shinya Takahashi Tsuyoshi Morimoto

1. はじめに

音声認識に用いられている言語モデルは，一般に形態素を単位とすることが多い．しかし助詞・助動詞のような単語長の短い付属語は誤認識を起こしやすいことが知られている．また熟語や慣用表現などは短い単位で認識するよりも長い単位で認識するほうがよい．これらの問題に対して，高頻度形態素連鎖語を辞書登録して言語モデルを改善する手法が提案されている[1],[2],[3]．文献[1]ではコーパス内の高頻度な形態素連鎖語を扱っているのに対し，文献[2]では慣用表現などの定型表現を対象としている点で相違があるが，いずれも長単位での連鎖語を言語モデルに組み込む手法について検討したものである．

これらの手法においては言語モデルに組み込む連鎖語をどのように選定するかが重要となる．文献[1]では，(1)出現頻度の高い連鎖語を選定する方法，(2)エントロピーの減少に貢献する連鎖語を選定する方法，(3)述べ単語数の減少に貢献する連鎖語を選定する方法の 3 種を比較しているが，実験としては語彙数 5000，最大追加連鎖語数 1000 とやや小規模なサイズにとどまっており，実験結果としては選定方法の違いで大きな差は得られなかったとしている．また貢献度の高い選定候補を絞りこむためには繰り返し計算が必要なこと，さらには信頼できる連鎖語の出現頻度情報を得るには膨大なコーパスが必要であることといった問題があった．

これらの問題に対し，本研究では，膨大な Web 上のドキュメントを対象として 7grams までの単語Nグラムの頻度を集計した Web 日本語 N グラム[4]を利用して，形態素連鎖語の単語連接確率を計算することにより，高頻度かつ定型的な表現を選定する手法を検討する．更にこれらの高頻度連鎖語を組み込んだ言語モデルを用いて音声認識実験を行い，その有効性を検証した．なお文献[2][3]で連語という呼称を用いているが，本稿では，形態素連鎖語もしくは単に連鎖語と呼ぶこととする．

2. 高頻度かつ定型的な連鎖語表現の選定方法

2.1 単語連接確率の計算

表 1 に，日本語話し言葉コーパス（以下、CSJ と呼ぶ）内の講演データに現れた連鎖語表現の例と Web 日本語 N グラムにおける出現頻度を示す．ここでは，3～7 形態素からなる連鎖語表現（約 60 万）を対象とした．表から分かるように，出現頻度は形態素数が少ないほど大きくなっている．出現頻度上位の語は，形態素数が 3 または 4 の短い表現が占める結果となった．さらに「て/い/ま/す」のような文末表現（"/"は形態素境界を示す）は，異なる形態素数の高頻度連鎖語表現に共通して現れていることが示された．以上のことから，出現頻度の上位のものを単純に選定する方法では，定型的な長単位の表現が組み込まれにくいことが予想される．そこで連鎖語表現の単語連接確率を計算し，高確率のものを選定することを考える．

表 1　高頻度な連鎖語表現の例

出現頻度	連鎖語表現	順位	形態素数
369M	て/い/ます	1	3
146M	し/て/い/ます	6	4
26M	で/は/あり/ませ/ん	129	5
16M	を/使用/し/て/い/ます	268	6
5M	の/で/は/ない/でしょ/う/か	1037	7

表 2　定型的な連鎖語表現の例

$\hat{H}(W_1^n)$	連鎖語表現	順位	形態素数
0.0061	見当たり/ませ/ん	1	3
0.0098	構い/ませ/ん	2	3
0.0471	細心/の/注意	3	3
0.1672	清聴/ありがとう/ござい/まし/た	20	5
0.2647	申し訳/あり/ませ/ん	61	4
0.3428	よろしく/お願い/し/ます	89	4

単語列$W_1^{n-1} = w_1 w_2 \cdots w_{n-1}$に単語$w_1$が後続する単語連接確率は，以下のように計算できる．ここで$C(\cdot)$は Web 日本語 N グラムにおける出現数を表している．

$$P(w_n|W_1^{n-1}) = \frac{C(W_1^n)}{C(W_1^{n-1})}$$

定型的な連鎖語表現では，この連接確率は高くなると予想されるが，W_1^nとW_1^{n-1}の出現数が共に小さいときでも$P(w_n|W_1^{n-1})$は 1 に近くなってしまうため，この確率だけでは高頻度かつ定型的な連鎖語を選定することはできない．ここで，連鎖語W_1^nは，単語列W_1^{n-k}に単語列W_{n-k+1}^nが後続したものと考えることができることから，以下に示すような平均単語連接確率及び（近似）クロスエントロピーを計算する．

$$\hat{P}(W_1^n) = \sqrt[n]{\prod_{k=1}^{n-1} P\left(W_{n-k+1}^n | W_1^{n-k}\right)}$$

$$\hat{H}(W_1^n) = \log_2 \hat{P}(W_1^n) = -\frac{1}{n}\sum_{k=1}^{n-1} \log_2 P\left(W_{n-k+1}^n | W_1^{n-k}\right)$$

表 2 に，$\hat{H}(W_1^n)$を用いて選定した連鎖語表現の例を示す．対象とした連鎖語表現の形態素数は表 1 と同じである．出現頻度のみに基づくよりも，連鎖語表現内の単語連接確率を考慮したほうが，より定型的な表現をうまく選定できることが示されている．また携帯素数が多くても定型的な表現は上位に来ていることが分かる．

† 福岡大学工学部

2.2 連鎖語を組み込んだ言語モデルの学習

言語モデル構築のための学習用コーパスには，CSJ に含まれる講義音声のうちテストセットを除いた 967 講演を使用した．表 2 に示したような連接確率の高い連鎖語のうち上位 500～10000 の語句を選定し，学習用コーパス内のテキストを形態素解析したのち，選定された連鎖語で形態素列を置き換えたテキストを学習テキストとしてトライグラム言語モデルを構築した．言語モデルの構築には，統計的言語モデルツールキット palmkit を使用し，語彙数制限は 2 万とした．ここで，この語彙数制限により頻度上位の語句として選定された連鎖語であっても学習コーパス内での出現頻度が低い場合は認識辞書から除外される点に注意が必要である．

表 3 に頻度上位の選定連鎖語数と実際に認識辞書に登録された連鎖語数および言語モデルのエントリ数を示す．なお認識辞書に登録された未知語数はいずれも約 1000 語程度である．表から分かるように選定した連鎖語のうち 4 割程度は言語モデルで使用されておらず，特に上位 10000 とした場合は半数以上が使用されない結果となった．これは，連鎖語選定に用いた Web 日本語 N グラムの対象文書集合と，学習用コーパスとして用いた CSJ の講演データの対象文書集合とが異なる性質のものであるためと考えられる．また N グラムのエントリ数は，使用する連鎖語が増加するにつれて増大している．

表 3　認識辞書に登録された連鎖語数

選定連鎖語	連鎖語認識辞書 登録数	2grams エントリ数	3grams エントリ数
連鎖語不使用	0	356,196	1,134,238
上位 500	372	369,935	1,154,487
上位 1000	714	373,918	1,162,528
上位 2000	1,369	382,098	1,178,546
上位 3000	1,917	388,982	1,191,937
上位 5000	2,885	401,986	1,216,358
上位 10000	4,687	423,808	1,249,792

表 4　音声認識実験結果（単語正解率(%)）

選定連鎖語	A01F0001	A01F0034	A01M0056	A01M0141
連鎖語不使用	72.6	70.7	73.3	59.9
上位 500	74.8	71.9	76.7	63.6
上位 1000	75.4	72.0	76.7	64.0
上位 2000	74.8	72.0	76.3	63.6
上位 3000	74.7	71.7	76.5	63.2
上位 5000	74.2	71.6	76.0	63.8
上位 10000	73.9	70.6	75.3	64.4

表 5　音声認識実験結果（単語認識精度(%)）

選定連鎖語	A01F0001	A01F0034	A01M0056	A01M0141
連鎖語不使用	66.0	61.9	65.9	50.9
上位 500	68.4	63.1	69.0	54.6
上位 1000	69.1	63.4	69.0	54.9
上位 2000	68.5	63.4	68.5	54.6
上位 3000	68.3	63.2	68.8	53.9
上位 5000	67.9	62.8	68.4	53.4
上位 10000	67.6	61.7	67.4	53.6

3. 音声認識実験

前節で示した連鎖語 Ngram を使用して音声認識実験を行った．音声認識には Julius Ver.4.1.5 を用いた[5]．音響モデルには CSJ 付属の性別非依存モデルを使用した．実験データには CSJ 内の test-set1 に含まれる 4 種類の講演音声データ（男性 2 講演，女性 2 講演）を使用した．

(正解文)
に/思わ/れる/かも/しれ/ませ/ん/が/、/えっと/、/我々/は

(連鎖語言語モデルによる認識結果)
に/思わ/れる/かも/しれ/ませ/ん/が/声/と/我々/は

(連鎖語を使用しない場合の認識結果)
に/生まれる/か/も/し/ませ/ん/か/って/と/我々/の

図 1 連鎖語使用による効果

音声認識実験の結果を表 4，5 に示す．表には，連鎖語の選定条件を上位 500～10000 とした際の結果を示している．また比較のため，連鎖語を使用せずに言語モデルを作成した場合の結果も併せて示した．図 1 は連鎖語使用により単語正解率が向上した文の例である．図中の"/"は形態素区切りを示し，アンダーラインは 1 単位の連鎖語として登録された語句を示す．なお，ここで示した「思われるかもしれません」という連鎖語は 5199 位である．

単語正解率では 1 つを除いて，また単語正解精度では全ての講演データに対して，上位 1000 の連語を選定したケースで最大値を示し，選定数が増えると値が低下する結果となった．上位 1000 の連語を使用したケースでは，最大で 4%，平均でも 2.9%の改善を得ることが出来た．また 1 つのケース（A01F0034 で上位 10000 連語を用いたケース）を除いて連鎖語を使用することにより単語正解率，単語正解精度のいずれも改善が見られた．上記の結果より，語彙数制限に対して適切なサイズの連鎖語追加数があることが予想される．今回の実験では語彙数制限 2 万に対して連鎖語追加の割合が高いケースにおいて悪影響が現れていると思われるため，より大語彙な言語モデルでも検証する必要がある．

4. まとめ

本稿では，Web 日本語 N グラムを用いて高頻度かつ定型的な連鎖語表現を選定する方法について検討した．小規模なテストセットに対する音声認識実験を行い，提案手法により選定した連鎖語表現を組み込んだ言語モデルを用いることで認識精度が向上することを確認した．今後は，さらにテストセットの規模を大きくした音声認識実験を行い，その性能評価を行う予定である．また，今回は頻度および接続確率という観点から連鎖語を選定することを試みたが，文法的観点からの連鎖語の種別による選定方法も併せて検討したいと考えている．

参考文献

[1] 和田 陽介 他,“大語彙連続音声認識における連鎖語の追加による語彙拡大の効果”, 情処論, Vol. 40, No. 4, pp. 1413-1420 (1999)
[2] 岩瀬 修, 森元 逞, 首藤 公昭, “連語を組み込んだ統計言語モデル”, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 100, No.521, pp. 109-114 (2000)
[3] 高橋他, “日本語話し言葉コーパスを用いた連語 Ngram 音声認識の検討”, 電気関連学会九州支部連合大会講演論文集（2012）
[4] 工藤他, “Web 日本語 N グラム第 1 版”, 言語資源協会発行
[5] T. Kawahara et.al., “Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository -- Software of Continuous Speech Recognition Consortium,” Proc. of ICSLP-2004 (2004)