RNN による日本語音声認識のための文字レベル言語モデル

Character-level RNN Language Model for Japanese Speech Recognition

1W130538-3 源　啓希　　指導教員　尾形　哲也　教授
MINAMOTO Hiroki　　Prof.　OGATA Tetsuya

概要：　本研究は，Neural Network(NN)を用いて音声認識の為の文字レベル言語モデル作成を目指したものである．従来の単語レベルの言語モデルは，予測単位が単語であるため膨大な語彙を考慮せねばならないが，文字単位での学習であれば予測に必要なニューロン数を減らすことができる．文字レベルの日本語言語モデルの研究は過去にも存在するが，NN を用いた作成は行われておらず，また漢字やひらがなが混在したデータセットを用いている．そこで本研究では NN の一種である Recurrent Neural Network(RNN)を用いて，ひらがなのみの日本語文章による言語モデルを学習し，評価を行う．具体的には「青空文庫」より収集したデータセットを用いて，従来手法である n-gram 言語モデルとの比較を行った．その結果，提案手法単体では従来手法に迫る性能を示し，両者を併用することによって，両者をそれぞれ単体で利用した場合に比べて大幅に性能が改善されることが確認できた．

キーワード：音声認識，言語モデル，ニューラルネットワーク
Key words: speech recognition,　language model,　Neural Network

1.　はじめに

高精度な音声認識の実現には，音声から音素を識別する音響モデルに加えて，言語モデルを用いて生成文を評価することが効果的である．

近年，機械学習の分野で注目される Recurrent Neural Network(RNN)は単語を単位とする時系列データとして扱うことで従来の統計的手法を上回る汎化性能を持つことが知られている[1]．しかし，単語数が莫大であるため大量のコーパスから統計的に単語の特徴量を介して RNN に予測させる必要がある[2]．日本語は英語のように単語ごとに分かち書きされていないため形態素解析を前もって行う必要があり，同じく RNN によって精度向上が計られているものの[3]，評価時に形態素解析システムに依存せざるを得ない．
対して文字を予測単位とする言語モデル[4]では形態素解析を行うことなく，かつ出力のためのニューロン数を抑えることができる．また，日本語のひらがなのような表音文字では音響モデルとの対応が容易になる．

2.　関連研究

日本語の文字を予測単位とした研究[5]では統計的手法によるものがあるが，単語に比べ言語制約が弱くなるため認識率の改善はならず，語彙数の低減は実現したが，漢字とかなの混ざった状態であるなどの課題がある．

3.　提案手法

そこで本研究では音声認識用の文字レベルの日本語言語モデルを提案する．　日本語はかな/カナ，漢字などが混在することが問題だが，今回は漢字の読みを取得してすべてひらがなに変換することによって語彙サイズを大幅に削減した．

3．1.　データセットの作成

著作権切れ文学作品公開サイト「青空文庫」[6]より 31 作家の収録された全作品を網羅して加工し，700000 文超のデータセットを構築した．はじめに句読点以外の記号を削除し，漢字を読みがなに変換した．

その後，カタカナなど変換が容易な部分をひら

がなに直し，解析不能や修正困難な個所については一文ごと削除する．

3．2．　ネットワークの構造

学習は RNN および Gated Recurrent Unit (GRU)の二種類で行った．GRU は RNN の一種で，時系列情報を引き継ぐか忘却するかを選択できる’update gate’を備えたネットワークである．この機能の効果を図るため，tanh を活性化関数とした RNN と性能比較する．

入力は文字，出力及び教師データは次の文字とし，隠れ層ユニット数を 125，250，500 の 3 種類の場合で学習を行った．

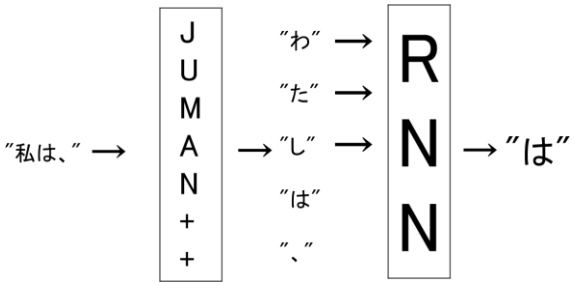

図 1 提案手法と学習過程

4．結果と考察

以下に本研究で作成した言語モデルと，同じデータセットに対し従来手法 n-gram で作成したモデル(n=3，5，7，10)の perplexity を図示する．Perplexity は以下で表される指標で，値が小さいほどモデルの性能が高いとされる．

$$\mathrm{PPL} = 2^{\mathrm{entropy}}$$

図より，RNN 単体での性能では従来手法に僅かに及ばなかったが，両者を併用することで性能の大幅な改善が確認できた．前の数単語という局所に強い n-gram と，長い文脈に強い RNN が相互に補完したと考えられる．

また，隠れ層のユニット数の増加につれてモデル精度が高くなり，同一ユニット数では GRU の方が高い性能を示した．

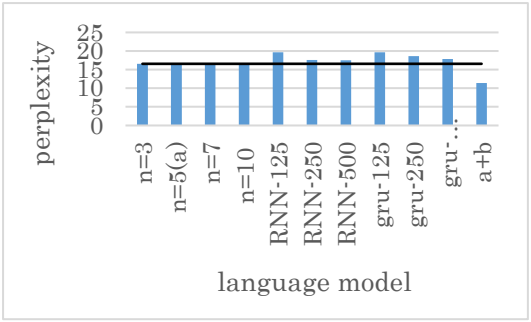

図 2　各モデルの perplexity

5．今後の展望

今後は実際の音声認識で，単語レベルの RNNLM と文章の認識率による比較を目指す．

データセットについては，〇人，～年といった数に関わる文字列がコーパスに収録できなかったという問題や，音声認識に対応させる上で‘きょ’‘ぴゅっ’といった音とどう対応させるかといった課題が残っている．

また，RNN のパラメータ設定についても，ビームサーチなどの手法を用いて最適な設定が行えるようにしたい．

参考文献

[1] T Mikolov，M Karafiát，L Burget，J Cernocký，S Khudanpur Recurrent neural network based language model.　Interspeech，2010

[2] T Mikolov，Kai Chen，Greg Corrado，Jeffrey Dean，Efficient estimation of word representations in vector space.　ICLR 2013

[3]森田 一，黒橋 禎夫，RNN 言語モデルを用いた日本語形態素解析の実用化 ，IPSJ，2016

[4] Yoon Kim，Yacine Jernite，David Sontag，Alexander M．Rush，Character-Aware Neural Language Models，AAAI，2016

[5]金野 弘明，加藤 正治，かな・漢字文字列を単位とした言語モデルの検討 ，IPSJ，2002

[6]青空文庫，2017，2，1 (http://www．aozora．gr．jp/)