出願研究科: 情報科学研究科
希望希望研究室: 自然言語処理研究室研究室
氏名: 岩田　晟

ニューラルネットを用いた教師なし単語分割の発展

1　はじめに

私は，学部時代の卒業論文ではトピックモデルと点過程の一種である Hawkes 過程を Yahoo Japan の検索データに対して用いることで，検索トピック内の流行を予測する研究を行った．その過程で，MeCab[*1]を用いて単語分割を行ったが，「楽天市場」が「楽天」と「市場」というように分解して欲しくない単語まで分解してしまうことがあった．そのような問題を解決するために，単語データの処理をよりスムーズに行うことに興味を持った．そこで，私が奈良先端科学技術大学院大学 (NAIST) で取り組みたい研究テーマは「ニューラルネットを用いた教師なし単語分割の発展」である．本稿では，この研究テーマの研究背景，先行研究，提案手法について述べる．

2　研究背景・目的

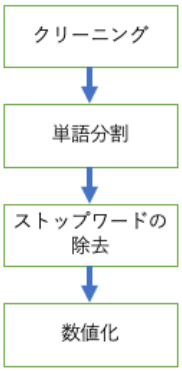

図 1　前処理の流れ

自然言語処理ではテキストデータを数値化する必要があるため，図 1 のような前処理をする必要がある．図 1 の 2 つめの工程のように，スペース区切りの英語と違い，日本語の文書は単語の区切りが明らかではないため，テキストデータに対して単語分割を行う必要がある．単語分割は主に MeCab や JUMAN++[*2]といった形態素解析器を用いて行われる．

森田ら [1] は形態素解析器を単語分割・品詞推定の精度を F 値を用いて比較した．そのなかで JUMAN++ が採用する Recurrent Neural Network Language Model(RNNLM)[2] に部分的アノテーションを加えたものが MeCab や JUMAN を抑えて最も良かった．その RNNLM の許容できない誤りとして，未知語や複合語に対する誤りがある．それらの誤りは学習時に用いる大量のアノテーション付きテキストが起因する．MeCab のように，固有語に強い辞書 mecab-ipadic-NEologd[*3] を用いることでその誤りを軽減することができる．しかし，SNS やインターネットの普及に伴い，新語も急増しており，辞書を更新するのも容易ではない．また，単語分割が付与されたコーパスが存在するドメインや言語は限られており，そのような辞書が用意できない場合もある．そのため，教師なし単語分割方法が求められる．

3　先行研究

3.1　ベイズ階層言語モデル

教師なし単語分割としては持橋ら [3] の Pitmon-Yor 過程を用いたノンパラメトリック階層ベイズ言語モデルがある．これは Pitman-Yor 過程に基づく n-gram モデルである Hierarchical Pitman-Yor Pitman-Yor Language Model(HPYLM) を拡張したものである．具体的には，図 2 のように単語 HPYLM に文字 HPYLM を埋め込んだ Nested Pitman-Yor Language Model (NPYLM) が提案している．

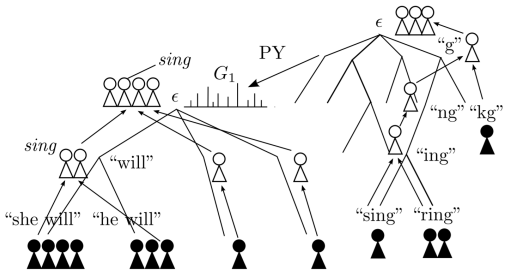

図 2　NPYLM の階層 CRP 表現 ([3] より引用)

3.2　ニューラルネットワークモデル

近年では，リカレントニューラルネットワーク (RNN) による言語モデルが提案されている．Zhiqing[4] らは

[*1] https://taku910.github.io/mecab/
[*2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++
[*3] https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md

RNN ベースの segmental language models (SLMs) を提案し，中国コーパスに対して実験を行なった．その結果，NPYLM と同程度の性能を誇っている．SLMs では単語区切りにどれだけの文字 (漢字) が続く数の候補のなかから尤度を最大化する候補を選ぶ．しかし，実際に全て候補を考えるのは不可能なため単語長の最大候補数を与えなければいけない．論文では中国語では 5 文字以上の単語は少ないため，2 ～ 4 文字としていた．また，SLMs の単語分割におけるエラーとして次のようなものが見られた．

- 分割するべきでない箇所に単語間の境界線を挿入するエラー (挿入エラー)
- 分割するべき箇所の単語間の境界線を削除するエラー (削除エラー)

4　取り組みたい研究内容

SLMs はモデルの最大候補数の決め方が難しいという問題があった．一見すると最大候補数が大きければ，多くの候補数が見られるので大きいほうが良いように思える．確かに，挿入エラーは最大候補数が大きくなるほど減ったが，その一方で．削除エラーが増えた．このようにエラーを解決するために，方法として Putman-Yor 過程のような階層的構造をニューラル言語モデルに用いることで改善することが考えられる．

HPYLM は次のような階層的構造を持っている．

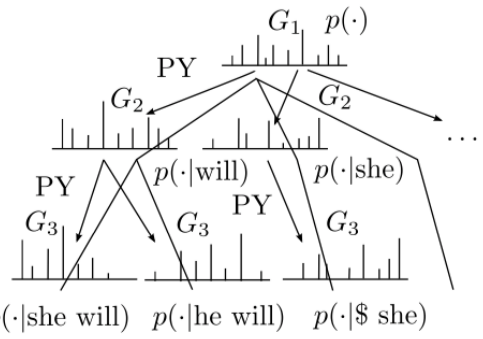

図 3　Pitman-Yor 過程による，n グラム分布 G_n の階層的な生成 ([3] より引用)

ここで n グラム分布を次のように一つ前の $n-1$ グラム分布を反映している．

$$G_n \sim \mathrm{PY}\left(G_{n-1}, d, \theta\right) \tag{1}$$

このように行うことで，n グラムでは出現しないが，$n-1$ グラムでは出現する事象を考慮することで，未知の単語にも確率が割り当てられる．そして，G_0 は文字 ∞ グラムすることで単語ごとでデータに出てこない単語にも等確率ではなくばらつきを持たせることが可能になる，そのため，文字ベースの言語モデルと考えることができる．

このように文字ベース，単語ベースと階層構造をニューラル言語モデルに適応し，スムーシングを行うことで頻出が少ない単語にも対応し，単語分割の性能の向上が期待できる．

5　おわりに

教師なし単語分割が未知語により強くなれば，よりスムーズに自然言語処理の前処理が期待できる．今回の話では議論しなかったが前処理の手段として，既存の手法と比べて挿入エラーに強い，押切らによる「単語分割を経由しない単語埋め込み」[5] が提案されており，単語分割にこだわらず，様々な手法を考慮して研究を行いたい．上記のような研究テーマに取り組むにあたって，なぜ NAIST を志望するのかというと，貴研究室では，Chasen などの自然言語処理ツールや日本語テキストコーパスが整備しており，研究を研究として終わらせるだけでなく，成果物として公表している．私も研究をして終わりではなく，最終的には他の人にも役に立つようにツールなどの利用できるように研究を形にしたいと考えている．

参考文献

[1] 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．情報処理学会 第 78 回全国大会，2016

[2] Hajime Morita，Daisuke Kawahara， and Sadao Kurohashi．Morphological analysis for unsegmented languages using recurrent neural network language model．EMNLP ，pp，2292 – 2297， 2015.

[3] 持橋大地， 山田武士， 上田修功．ベイズ階層言語モデルによる教師なし形態素解析．情報処理学会研究報告 2009-NL-190， 2009.

[4] Zhiqing Sun， Zhi-Hong Deng． Unsupervised Neural Word Segmentation for Chinese via Segmental Language Modeling． EMNLP， 2018.

[5] 押切孝将，下平英寿．単語分割を経由しない単語埋め込み．言語処理学会　第 23 回年始大会， pp. 258-261，2017.