GiNZA を用いた読点自動挿入システムの精度向上

情報科学科　黒田 真由　　指導教員：山村 毅

1　はじめに

読点は日本語の文章を構成する上でとても重要な役割を果たす．文の終わりに挿入すればよい句点と異なり，読点の挿入位置については明確な基準が存在しないため，留学生など日本語を母国語としない人々にとって，適切な位置に読点を挿入することは難しい．そのような人々の文章作成を支援するために，読点の自動挿入技術が重要となる [2].

坂ら [1] は，簡略化したモデルを考え，読点を挿入するシステムの開発を行った．具体的には，文の長さに依存して決まる読点の個数の確率と、形態素と形態素の間に読点が入る確率とを用いて読点を挿入する方法を提案した. また，文の長さの測り方として形態素数，文字数，文節数の 3 つを比較した．形態素解析に MeCab を用いたところ読点挿入の精度は再現率 46．0% 適合率 73．0% となった．しかし，地名の入った学校名や人物名，カタカナや漢字の羅列に対して，固有名詞の形態素解析の精度が特によくないことが大きな問題であった．そこで，固有表現抽出の精度や形態素間に読点の入る確率の質を高めていくことが，より実用的なシステムとすることに最も重要だと考えた．

本研究では，坂らの研究における課題点であった『固有名詞に対しての形態素解析の精度の低さ』について，『固有表現の抽出』を用いることで，読点自動挿入の精度向上を目指すこととする．

2　提案手法

坂らの研究では，日本語形態素解析システム MeCab の形態素解析の精度の低さに加えて，固有名詞に対して『Person，Place，Organization，Others』の 4 つのラベル付け以外行われていなかった．

このため，固有名詞の周辺情報が不足することになり，結果として読点挿入の精度低下を招いたと考えられる．

したがって，もし固有名詞に対しラベルを新たに与えることができれば，読点挿入の精度を向上させることができると考えられる．そこで，日本語自然言語処理ツール GiNZA[3] の固有表現抽出機能を利用して，固有名詞のラベルが与えられていなかった単語に対し，『Person，Company，Country，Province，City，Island，Time，Date，Money，Percent，Position Vocation 』の 11 個の固有表現ラベルを与え坂らの方法で読点自動挿入を行う．

3　評価実験

評価に先立って，まずは毎日新聞の 2008 年を用いて，GiNZA で形態素解析を行い固有表現抽出を行った後，Person(姓)，Person（名）など，同じラベルとして認識された単語を結合した．

そのあと文の長さとして，形態素数，文字数，文節数の 3 つを考え，それぞれにおいて，挿入される読点の個数の確率を調べたこうして求めた確率を用いて，評価対象の文に対して，最適な読点挿入の結果を導き，読点の挿入を行なった．毎日新聞の 2010 年の記事を対象とし，精度の検証方法としては再現率と適合率を用いた．

4　結果

先行研究の結果を表 1 に，GiNZA で固有表現抽出を用いなかった結果を表 2 に，固有表現抽出を用いた結果を表 3 に示す．

GiNZA の固有表現抽出を用いた場合，坂らの研究より精度が上がった．

表 1　先行研究 (単位 %)

	再現率	適合率
形態素数	45．6	74．0
文字数	46．0	73．0
文節数	44．9	73．4

表 2　固有表現抽出無し (単位 %)

	再現率	適合率
形態素数	44．2	68．8
文字数	44．4	70．8
文節数	43．2	70．6

表 3　固有表現抽出有り (単位 %)

	再現率	適合率
形態素数	54．5	73．0
文字数	52．7	74．8
文節数	52．5	75．1

5　おわりに

本研究では，固有表現ラベルを用いることで，坂らの研究における読点自動挿入の精度向上を試みた．表 2，3 に示すように，固有表現抽出を用いた場合，固有表現抽出を用いない場合と比べて精度が上昇していることが分かる．このことから，固有表現抽出により高い精度を得られると考えられる．本研究では，先行研究に対して有意差はあったが，適合率は 75 %程度で，まだ十分とは言えない．これは，ラベルを与えることのできていない固有名詞が，まだ多くあるためだと考えられる．今後の課題として，固有表現抽出のモデルを自前の学習データで学習することで更なる精度の向上ができるのではないかと見込んでいる．また，文の長さに依存して読点挿入を行ったことによって，短い文に対して読点が挿入されないという問題があった．読点挿入する確率が一定以上であれば読点挿入を行うなどの工夫が必要である．本システムは，形態素解析の精度や形態素間に読点の入る確率の質を高めていくことによって，より実用的なシステムとすることができると考える．

参考文献

[1] 坂祥太郎，山村毅:“文の長さと読点生成確率を用いた読点挿入システム”，言語処理学会第 25 回年次大会発表論文集，P3-37，2019

[2] 村田 匡輝，大野 誠寛，松原 茂樹: “読点の用法分類に基づく自動読点挿入”，情報処理学会研究報告，Vol．2010-NL-196，No．8，pp．1-8，2010

[3] Megagon Labs:“GiNZA Japanese NLP Library”，https://megagonlabs.github.io/ginza/