文の長さの制約を緩和することによる読点自動挿入システムの性能向上

情報科学部 情報科学科　水藤 聖也　　指導教員：山村 毅

1　序論

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は異なり，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[1]．また近年では，SNS などの急速な普及により短文でのコミュニケーションが増加し，読点を使って文章を作成する機会が減っている．文章の意味を正しく伝え，読みやすくするために読点は重要であるが，読点の挿入支援システムはいまだ実用化されていないため，本研究では日本語文への読点の挿入支援システムの開発を行う．

2　提案手法

坂ら[2]は，文の長さに依存して決まる読点の個数の確率と，2 つの形態素の間に読点が入る確率を用いて読点を挿入する方法を提案している．文の長さの測り方(形態素数，文字数，分節数)の違いによって精度がどう変わるかを考察し，再現率 46.0%，適合率 73.0%という高い適合率を実現している．また，黒田[3]は，日本語自然言語処理ツール GiNZA[4]の固有表現抽出機能の利用により，再現率を 54.5%まで改善している．

しかし，これらの方法では，文の長さに基づいて挿入読点数を決定してから読点を挿入していたため，読点が複数入っている短い文や，逆に読点があまり多くない長い文に十分に対応できないという問題があった．

そこで，挿入する読点数に幅を持たせ，数パターンの文章を一度生成し，そこから適切な文章を選ぶという形で文章を生成することができれば，読点挿入の精度向上に繋がると考えた．

また，読点の有無によって形態素解析が間違ってしまうような場合，正しい結果が得られないため，文章を形態素解析してから読点の挿入位置を決めるのではなく，先に機械的に読点を入れてから形態素解析を行い，読点位置の評価を行うことにする．

3　評価実験

評価に先立って，まず毎日新聞の 2008 年の記事データを GiNZA で形態素解析し，隣り合った 2 つの形態素の間に読点が入るトライグラム確率を求めた．次に，文の長さとして文節数を採用し，それぞれに対して挿入される読点の個数の確率を求めた．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位置を求めた．評価には，毎日新聞の 2010 年の記事を用い，正解率と再現率，適合率を求めた．

4　結果

先行研究の結果と本研究の結果を表 1 に示す．ただし，GiNZA が固有表現の解析誤りを起こすことがあったため，生成した 1 万行の文のうち，実際に精度計算に使用したのは 7607 行である．表 1 からわかるように，提案した手法を用いた場合，先行研究[3]よりも精度が向上した．なお，先行研究の精度は，トライグラム確率の計算に問題があり，それを修正して改めて評価し直したため，2 で記載した精度とは異なっている．

	先行研究の精度	本研究の精度
正解率	97.73	98.78
適合率	63.18	63.42
再現率	47.34	50.94

表 1　研究の精度比較(単位%)

実験後，本研究の正解率に関する有意差を調べるためにマクネマー検定を行った．マクネマー検定は複数のシステムの出力を対応のある 2 群に分類し，システムの差異を検定する方法である．本研究と先行研究の実験結果をマクネマー検定で使用する式に当てはめると，p 値は 0.018 と計算でき，有意水準を 0.050 とした場合，本研究のシステムと先行研究のシステムの正解率の違いは有意なものであると判定できた．

5　おわりに

本研究では，読点挿入の際に文の長さの制約を緩和し，さらに形態素解析の手順を変更することで，読点自動挿入の精度向上を試みた．表 1 に示したように，このような方法を用いることで高い精度が得られた．しかし,本研究は先行研究に対しての有意差はあったが，精度にまだ改善の余地がある．これには様々な原因が考えられるが，学習データに利用した新聞記事にあまり出現しない形態素(特に名詞)に関して，読点が入る確率が実態以上に高くなる場合や低くなる場合があり，正しい文章を生成できなかった点は大きい．しかし，表層表現の情報を除外し，品詞の情報のみを用いた別の実験を行ったが,それだけでは精度の向上は認められなかったため，データ数の少ない形態素に関しては別のラベルを用意するなど,特別な対応が求められる．

また，「読点を入れてから形態素解析を行う」という方法に関しては，読点の位置によっては形態素解析が正しく行われない問題があった．そこで，形態素解析を行うツールを変更するというアプローチはもちろん，依存構造解析やニューラルネットワークを用いた解析など，別の新たなアプローチを追加することでさらなる精度の向上ができるのではないかと考える．

参考文献

[1] 村田 匡輝，大野 誠寛，松原 茂樹：“日本語テキストにおける読点位置の検出”, 言語処理学会第 16 回年次大会発表論文集，pp.812-815，2010.
[2] 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, pp.655-658，2019
[3] 黒田 真由： “GiNZA を用いた読点自動挿入システムの精度向上”，愛知県立大学情報科学部卒業論文，2022
[4] Megagon Labs：“GiNZA Japanese NLP Library”，https:/ /megagonlabs.github.io/ginza/