# 平成30年度 情報工学コース卒業研究報告要旨

| 松原 研究室 | 氏　　名 | 宮 地　航 太 |
|---|---|---|
| 卒 業 研 究 題 目 | 読みにくい語順の文への読点の自動挿入 | |

日本語テキスト生成は，機械翻訳や自動要約，音声筆記などの性能を決める重要な技術である．生成されたテキストが高い品質を備えているためには，読点が適切な位置に挿入されている必要がある．というのも，読点は文中の区切りを明示する記号であり，その挿入位置は，文の読みやすさや読み手による文の解釈に影響を与えるためである．

これまでに，日本語テキストへの読点挿入手法が村田らにより開発されている．村田らは，文構造を明確にする，並列する語の区切りを示す，など，読点の用法ごとにその出現傾向を捉える特徴素を設定している．しかし，この手法では，新聞記事など，語順が整った文を対象としており，字幕など語順が整っていない文に対する効果は明らかでない．

本論文では，読みにくい語順の文に対応した読点挿入手法について述べる．日本語文は，同一文節から構成されていても，その語順が変われば読点の打ち方も変わる．例えば，

S1. 鈴木さんは都会に憧れ、家を飛び出した。

S2. 鈴木さんは家を、都会に憧れ、飛び出した。

の場合，読みやすい語順の文 S1 では「家を」の後に読点は挿入されないが，読みにくい語順の文 S2 では読点が入る．

本研究では，新聞記事文から，読みにくい語順の文 (546 文) を擬似的に作成し，3 名の作業者 (A,B,C) が読点を付与したデータから 273 文を用いて分析した．作業者間の読点の重なり具合を図 1 に示す．分析では，隣接文節の係り先の同一性に注目した．係り受け関係にあるものを除く隣接文節の組 (計 771 組) に対して，係り先の一致と両文節間の読点の有無との関係を調査したところ，係り先が異なる文節間には読点が挿入されやすく (表 1)，また，隣接文節の係り先の不一致率は，読みにくい語順の文の方が高いことが明らかになった．このため提案手法では，村田らの手法の特徴素に対し，隣接文節の係り先の一致を表す特徴素を新たに導入した．これにより，読みにくい語順の文における，係り先が同じ文節同士が離れて位置することによって打たれる読点を捉えることができる．

提案した特徴素の効果を評価するため，読みにくい語順の文データを用いて読点挿入実験を行った．学習には，京大テキストコーパスのうち，分析データの作成に用いた文を除く 37,854 文を用いた．評価は，2 名以上の作業者によって付与された読点位置を正解の読点位置としたときの，正解に対する再現率，及び，適合率により行った．実験の結果，本手法の有効性を確認した．

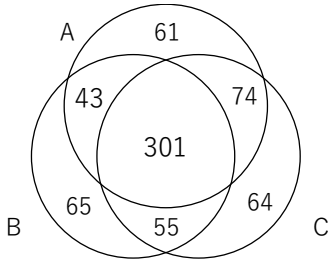


図 1　3 名の作業者による読点の集計

表 1　隣接文節の係り先の同一性と文節間の読点の有無の関係

| 読点 | | 有 | 無 |
|---|---|---|---|
| 係り先 | 一致 | 141 (38.01%) | 230 (61.99%) |
| | 不一致 | 246 (61.50%) | 154 (38.50%) |