主観素性を有する言語辞書を用いたビックデータ解析システムの研究開発（17944469）

Research and development of large amount of data analysis system using dictionary with modality feature.

研究代表者
足立　顕　アーカイブ技術研究所株式会社
Akira Adachi　Institute of archive technology co., Ltd.

研究分担者
山本　竜伸 †
Tatsunobu Yamamoto†
†株式会社ザイナス
†Zynas co., Ltd.

研究期間　平成 29 年度

概要

日本語を構成する要素は大きく自立語と付属語に分類できる。本研究は付属語部分に着目し可能な付属語の連続（付属語列）を１つの形態素として取り扱う場合に必要な語数の導出および付属語列を用いた形態素解析器の開発を目指す。付属語列は自立語間の関係を示す構文マーカであると同時に肯定否定や推量可能などの発話者の主観情報を持つ。主観素性を付与した付属語列辞書の構築と簡易な文法で動作する形態素解析器の開発を行う。

１．まえがき

計算機の能力向上および技術の進歩によって大規模な情報から特定のキーワードを含む情報や特定のパターンを持つ情報を高速かつ大量に抽出することができるようになり、傾向を捉えることは可能となったが、最終的に情報の意味を理解･解釈するためには人手による場合が少なくない。発信される情報は膨大であり監視することも限界があり、今後さらなる効率化が必要となる。

詳細な言語解析を行うためには、入力文をその構成要素に分割し、各々の要素が文中においてどのような役割（単語や品詞）を果たしているか解析する必要がある。

日本語文を構成する要素は大きく、「自立語」と自立語に付属して文法的な役割を果たす「付属語」に分類できる。「付属語」は文中において単独では意味をなさず直前の語に付属することによって語の係り受け関係を示す表層格マーカ（格助詞など）や「ある」「ない」など肯定否定、「～だろう」など推量を示す発話者の意思を表す重要な役割を持つ。

本研究は、可能な付属語の連続を１つの付属語（以下、付属語列と呼ぶ）として扱い、その付属語列に「否定の推量」などの素性を与える。直接、形態素解析結果から自立語に対する付属語の機能を得ることができる枠組みの実現を目指す。

付属語列として扱う場合、日本語の形態素解析器として必要な付属語の語数および付属語に与えるべき素性（否定、推量などの機能素性）の体系化を行う必要がある。本研究では抽出した付属語列を分類し形態素解析用の辞書を構築する過程で付属語列として扱う形態素の範囲等を決定し、適正な必要語数および素性について検討する。

２．研究開発内容及び成果

本研究開発では、単語 n-gram および文字 n-gram を用いて付属語列の出現傾向および付属語列抽出のためのシステムの開発を行った。文字 n-gram を用いた抽出実験によって得た文字列を最長一致または先頭からの最長一致で形態素解析するシステムを開発。画面上で分割結果を確認することができるシステムを構築した。また、この形態素解析器を利用して文字 n-gram から過剰な分割によって得た文字列を除去するための実験を行った。

(1). 単語 n-gram による実験

単語 n-gram による実験では、抽出条件を複数設定し付属語列の表記の異なりの変化を測定した。

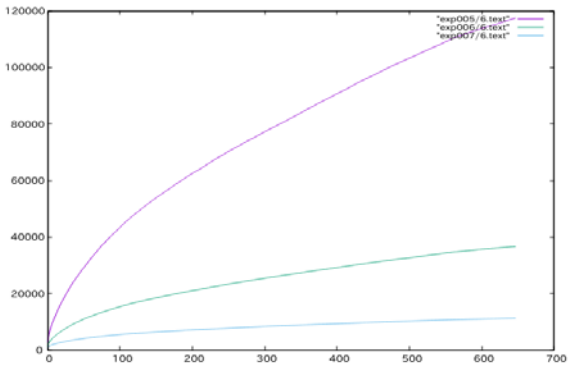

図１　付属語列の表記数の変化

付属語列の抽出には、新聞コーパスを用いた。形態素解析結果に未知語を含まない 647 万文から付属語列を得た。図１は、横軸に抽出対象文数（単位１万文）、縦軸に付属語列の表記数を示した。

抽出条件は、図１の図中の上から、【条件１】助詞、助動詞、名詞および動詞の非自立語，動詞語尾の連続。【条件２】助詞，助動詞，動詞の非自立語，動詞語尾の連続。【条件３】助詞，助動詞の連続である。

抽出対象コーパスを増加することによる付属語列の再現率を調査した結果を図２に示す。

全実験コーパス（647 万文）のうち、646 万文から取得した付属語列をこれとは異なる１万文に適用した場合、条件１で 99.8838%、条件２で 99.9677%、条件３で 99.9939%となった。（参考：1 万文から得る付属語列の生起頻度の平均は、条件１で 65,107 回、条件 2 で 66,695 回、条件３で 67,422 回）

条件３においては全コーパスで 50 回以上の頻度で生起す

る付属語列は、最初の 10 万文のコーパスにほぼ含まれていた。

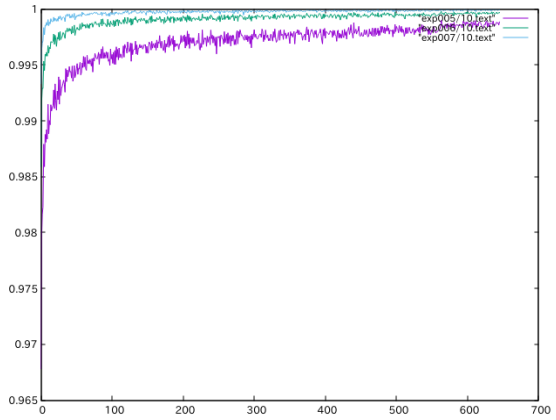

図２　付属語列の再現率

手始めに条件３で抽出した付属語列 11,208 表記について抽出対象としたコーパスの当該表記の記述部分を参照することで解析ミスによって得た付属語列を人手により除去した。この人手での作業で除去した付属語列は８４５表記あり、取得した付属語列のおよそ７．５％となる。解析ミスにより得た付属語列は、ひらがな表記した固有名詞等が影響した解析ミスによるものが主体であった。

(2). 文字 n-gram による実験

任意の n 文字が連続する文字列を取得することで付属語列を抽出する取り組みを行った。すべての入力文を任意の n 文字に分割し、各々の生起頻度を集計する。この文字列のから特定の文字列の部分列としてのみ存在する表記を除去した。

【例】	（表記）	（頻度）
	成田空港	10
	成田空港か	2
	成田空港から	2
	成田空港からバ	1

上記のような値を得た場合、「成田空港か」と「成田空港から」は同一頻度であるため「成田空港から」が生起したことによって得た部分列と考えることができるため除去する。同様にして後方から並べ替え同様の方法で部分列を除去した。付属語列に記号などを含まないと仮定し、新聞コーパス１年分より記号などを分割点として得た 4,345,086 文字列から上記手法により、7,741,275 文字列を得た。この文字列には自立語に関するものを含み、同時に付属語列の内部構造も含む。連続する最長の文字列を得るため、最長一致優先および文頭からの最長一致文字列を得ることによって分割する形態素解析器（図３）を作成し、解析結果として使用した文字列を測定した。

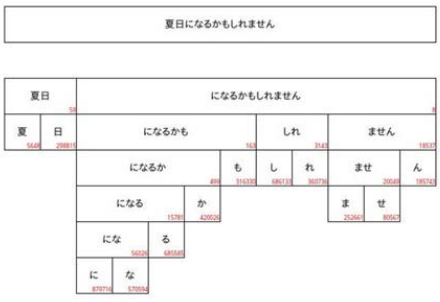

図３　形態素解析結果ビューア

その結果、最長一致優先では、2,791,242 文字列、先頭からの最長一致では、2,899,975 文字列使用した。両方の手法では共通して使用した文字列は、2,00135 文字列であった。

３．今後の研究開発成果の展開及び波及効果創出への取り組み

今後は、人手によってノイズ除去した付属語列を用いて、異なる付属語抽出条件（本文では条件１と条件２）で得たデータからノイズを除去するフィルタリングシステムを構築、フィルタリングした結果を人手で検証し、付属語列の辞書を構築することを目指す。

また、単語 n-gram で取得した文字列と文字 n-gram で取得した文字列の比較を行う実験を行うとともに単語 n-gram で得た品詞情報を元に文字 n-gram で得た文字列に適応するための実験を行うことを予定している。同時に主観に関する素性を文字列に与えるための仕組みについて研究を行い、最終的には形態素解析結果として主観素性を得ることができる枠組みの構築を目指す。

例えば、「太郎は次郎が犯人だと思うだろう」という文は「～と思う」と「～だろう」の２つのモダリディになる語を含む。「～と思う」の主体は「太郎」であり、「～だろう」の主体は話者である。文には複数の主体とモダリティを含む場合があり静的に決まるものではない。主観素性を語に与えることで、モダリティの構造の把握に貢献し主体と述部のスコープに対して制約を与えることができる期待がある。

主体と述部との関係の解析精度（構文解析精度）が向上することは、機械翻訳、要約、情報検索といった文書処理の質的向上に貢献する。モダリティの構造の把握は、日本語学習者の拠り所となるだけでなく、コンピュータによる、より質の高い言語解析を行う基盤を与えるものである。例えば、「雨に降られた」という文での日本語の迷惑の受身「～られた」は、それを表す表現が英語にはないとされる。しかし、この文が伝達する意図である「困惑」という感情は、例えば、助動詞や態の変化だけではなく、感嘆詞などで表わすことができる。たとえ形態上の表現法は異なっても、その伝達意図を正しく把握した文として表現することが可能になる。

４．むすび

本研究開発期間では、実験用システムの開発にとどまった。実験システムはパソコン程度のリソースで比較的大きな文字 n-gram のデータを取得することができるものを構築した。

まずは本実験によって得た付属語列について素性を付与した辞書を構築し付属語列を用いたシステムの有用性を評価するとともに、新聞コーパス以外のデータを用いて実験を行い付属語列のコーパスの種別による違いについて検討する。

【本研究開発課題を掲載したホームページ】

http://www.tech-a.co.jp/SCOPE/