審査の結果の要旨

氏名 靳 志輝（きん しき, Zhihui Jin）

本論文は、言語の分節メカニズムに関するものである。チョムスキーの師であるZellig S. Harris は、1955 年の「From phoneme to Morpheme」の論文の中で、音素列における「後続音素の種類数」の変化の極大点が形態素や単語の境界と一致するとの仮説を示している。この仮説は、言語学上は二重分節における音素列と形態素列の関係を示す興味深いものであると共に、言語工学上もunsupervised segmentation の一手法としての有用性を秘めている。Jin 氏は、この仮説がどの程度成立するのかを英語と中国語において大規模なデータ上で確かめ、また、Harris のモデルの限界を指摘し、改良する一方策をその博士論文において示した。

本研究は、二つの意味で重要である。第一に、Harris のモデルがどの程度成り立つのかを大規模に確かめた初めての研究である点である。言語学者Harris のモデルは、分節に関して言語データに内在する原理を記述するものである。しかし、Harris のモデルがどの程度成立するのかについては、未だ明らかではなかった。論文が発表された1955 年当時はコーパスや計算機がなかったため、数文について人を対象とした実験を通して、モデルは検証されているにとどまる。その後、1970 年代にHafer らが計算機を利用して実験を行っているが、用いられたデータも数十の文についてのみであり、小規模にとどまる。Harrisのモデルを厳密に踏襲して大規模実験によりその信憑性を確認したのは本研究が最初である。

第二は、教師なし学習に基づく文分割の一方法としてのHarris のモデルの可能性を明らかにした点である。昨今の言語工学における文分割方式は、HMMやCRF といったモデルを仮定し、人手で分割点を付与した正解付きコーパスから機械学習を用いてモデルの最適なパラメータを推定する方式が一般的であり、97 %を超える高い精度が現在報告されている。一方で、教師あり学習による方式は、正解付きコーパスの作成に莫大な人手コストがかかること、正解付きコーパスのない言語や分野には適用できないこと、また、高い精度で文分割はできてもそれがどのような原理に基づくものであるのかについては何も明らかにはならないこと、などの問題点が指摘されている。この問題を解決する方式として、正解が付与されてはいない大規模な言語データだけを仮定する教師なし学習方式に期待が寄せられている。

教師なし学習にもさまざまな方式があり、その最も一般的なものにおいては、教師あり学習時に用いる言語モデルをそのまま仮定し、必要なパラメータをEM アルゴリズムで推定する。しかし、この方式は、分節原理に関する知識が枠組みに組み込まれていないため、精度上の問題がある。ヒューリスティクスに基づく方法も提案されているが、その言語学的意味は明らかではない。一方で、Jin 氏が用いた方法は、Harris が提唱した言語の分節

点に関する原理に基づいて分割を行うものである。方法は極めて単純であるにもかかわらず、精度は他の教師なし手法に劣るものではなく、中国語の形態素解析において、現存する他の教師なし学習手法に対しわずかではあるが勝ることが、Jin 氏の論文では示されている。

以上をふまえ、論文は以下のように構成されている。まず以上のような言語学上、工学上の背景が第二章にまとめられている。つぎに第三章で、Harris のモデルを用い、要素列を要素まとまりの列に分割する工学的な方法として、後続要素のエントロピーの変化の増大点を分節の候補点とする手法の定式化がなされている。

第四章では、中国語の形態素解析に本方式を適用した際の精度が示されている。単純な方式にもかかわらず、精度は88.5%であることが明らかとなった。また、学習データ量の対数に対して再現率が線形で増大することが示された。この内容については、国際会議のポスター論文が一編出版されている。

第五章は、Harris の論文に忠実に、音素から形態素への分節が本方式でどの程度の精度で行われるのかを、中国語と英語で検証した結果を報告している。具体的には、百Mバイト程度のテキストデータを音素列に変換し、それを用いて、音素列が形態素列や単語列にどの程度正しく分割されるのかを調べている。結果、形態素への分割は8 割強であった。また、英語では、単語への分節が形態素への分割精度に勝る一方で、中国語では、単語への分割精度は低く、英語と中国語における表記システムの差異などの影響を窺うことができると報告されている。この内容については、英文学術論文誌論文が一編、国際会議論文が一編出版されている。

第六章では、Harris のモデルが局所的な文脈情報にのみ基づく方式となっていることが問題であることが論じられ、文書全体で分割が整合するように最適化をすることで、モデルを改良することができることを示している。具体的に中国語の形態素解析において実験が行われ、文書全体での最適化を行うと、再現率が向上し、また、精度もわずかではあるが向上する。結果として、2004 年に報告されている中国語の形態素解析の最高精度をわずかではあるが上回る結果を出している。この内容については、Jin 氏は現在国際会議論文を執筆中である。

最後に、Jin 氏は、以上の研究とは別に、漢字対応に基づく日中辞書引き方法を考案し、日中の架け橋となる有用なソフトウエアを創造した。本研究についても、一編の国際会議論文ならびに国内会議論文が執筆された。この論文は本論文の内容上のまとまりを保つため、付録として収録されているが、その有用性は高く評価できる。

以上、言語学的にも、言語工学的にも、本論文で述べられている成果には高い価値が認められる。よって、本論文は博士(情報理工学) の学位請求論文として合格と認められる。