氏　　名	能地 宏
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1835 号
学位授与の日付	平成28年3月24日
学位授与の要件	複合科学研究科 情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Left-corner Methods for Syntactic Modeling with Universal Structural Constraints
論 文 審 査 委 員	主　査　准教授 宮尾 祐介 准教授 金沢 誠 准教授 持橋 大地 教授 中川 裕志 東京大学 准教授 Edson T. Miyamoto 筑波大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

Explaining the syntactic variation and universals including the constraints on that variation across languages in the world is essential both from a theoretical and practical point of view. It is in fact one of the main goals in linguistics. In computational linguistics, these kinds of syntactic regularities and constraints could be utilized as prior knowledge about grammars, which would be valuable for improving the performance of various syntax-oriented systems such as parsers or grammar induction systems. This thesis is about such syntactic universals.

The primary goal in this thesis is to identify better syntactic constraint or bias, that is language independent but also efficiently exploitable during sentence processing. We focus on a particular syntactic construction called center-embedding, which is well studied in psycholinguistics and noted to cause particular difficulty for comprehension. Since people use language as a tool for communication, one expects such complex constructions to be avoided for communication efficiency. From a computational perspective, center-embedding is closely relevant to a *left-corner* parsing algorithm, which can capture the degree of center-embedding of a parse tree being constructed. This connection suggests left-corner methods can be a tool to exploit the universal syntactic constraint that people avoid generating center-embedded structures. We explore such utilities of center-embedding as well as left-corner methods extensively through several theoretical and empirical examinations.

We base our analysis on dependency syntax. This is because our focus in this thesis is the language universality. Now the number of available dependency treebanks are growing rapidly compared to the treebanks of phrase-structure grammars thanks to the recent standardization efforts of dependency treebanks across languages, such as the Universal Dependencies project. We use these resources, consisting of more than 20 treebanks, which enable us to examine the universality of particular language phenomena, as we pursue in this thesis.

First, we quantitatively examine the universality of center-embedding avoidance using a collection of dependency treebanks. Previous studies on center-embedding in psycholinguistics have been limited to behavioral studies focusing on particular languages or sentences. Our study contrasts with these previous studies, and provides the first quantitative results on center-embedding avoidance. Along with these experiments, we provide a parser that can capture the degree of center-embedding of a *dependency* tree being built, by extending a left-corner parsing algorithm for dependency grammars. The main empirical finding in this study is that center-embedding is in fact a rare phenomenon across languages. This result also suggests a left-corner parser could be utilized as a tool exploiting the universal syntactic constraints in languages.

We then explore such utility of a left-corner parser in the application of unsupervised grammar induction. In this task, the input to the algorithm is a collection of sentences, from which the model tries to extract the salient patterns on them as a grammar. This is a particularly hard problem although we expect the universal constraint may help in improving the performance since it can effectively restrict the possible search space for the model. We build the model by extending the left-corner parsing algorithm for efficiently tabulating the search space except

those involving center-embedding up to a specific degree. Again, we examine the effectiveness of our approach on many treebanks, and demonstrate that often our constraint leads to better parsing performance. We thus conclude that left-corner methods are particularly useful for syntax-oriented systems, as it can exploit efficiently the inherent universal constraints in languages.

(別紙様式 3)
(Separate Form 3)

博士論文の審査結果の要旨

Summary of the results of the doctoral thesis screening

本博士論文は、全 6 章から構成される。第 1 章では、自然言語の構文構造に対する言語普遍的な制約に関する研究の科学的・工学的意義について議論し、本論文の中心テーマである中央埋め込み構造を避けるという言語普遍的制約を導入している。そして、博士論文の貢献として、中央埋め込みの数を定量化できる依存構造解析アルゴリズムの提案、中央埋め込みを避ける性質の言語普遍性を実際のデータで検証した実験、および本アルゴリズムを応用した教師なし文法獲得の実験について主張している。

第 2 章では、研究の背景として、句構造や依存構造などの構文構造、提案アルゴリズムの基盤となる左隅型構文解析、教師なし文法獲得のベースライン手法と関連研究について導入している。

第 3 章では、第 4 章、第 5 章の実験で用いるデータとして、CoNLL データセット、Universal Dependencies、Google Universal Treebanks を紹介し、これらのデータの特質について議論するとともに、これらのデータが仮定している文法構造の違いが教師なし文法解析の評価に問題を引き起こすこと、またその解決案について述べている。

第 4 章では、中央埋め込みの数を定量化することができる依存構造解析アルゴリズムを提案している。本アルゴリズムは、句構造に対して提案された左隅型構文解析アルゴリズムをベースとし、これを依存構造解析のために拡張・修正したものである。左隅型構文解析は、構文解析の途中結果をスタックに保存しながら解析を進めるアルゴリズムであるが、スタックに保存された要素の数と中央埋め込みの数が対応するという性質がある。この性質を満たすように、依存構造解析の遷移型構文解析アルゴリズムを構成した。さらに、上記のデータセットに対してこのアルゴリズムを適用し、中央埋め込みを避けるという性質が言語普遍的な制約であるということを実際のテキストデータで実証した。実データでこのような結果を示したのは本研究が初めてである。

第 5 章では、提案アルゴリズムを応用した教師なし文法獲得の実験を行っている。教師なし文法獲得は、全ての木構造について文法規則の期待値を計算する必要があり、動的計画法が必要となる。したがって、第 4 章のアルゴリズムを改良し、動的計画法が適用できるアルゴリズムを構成した。これを用いて、中央埋め込みが一定数以上含まれる構文木を学習対象から外す文法学習手法を提案した。上記のデータセットを用いた実験によると、中央埋め込みの制約を単独で用いたときは既存手法（依存構造の長さを小さくするバイアスをかける手法）より精度が低いが、文全体の主辞に対する制約と組み合わせると、既存手法より高い精度を達成し、またより多くの事前知識を必要とする既存手法と同等精度を達成することが示された。これにより、中央埋め込みを避けるという制約が、文法獲得に有効であることが実証された。

第 6 章では、以上の結果にもとづき本論文の貢献をまとめ、将来の研究課題について議論している。

博士論文の内容については、提案手法、定式化、評価実験など、博士論文として十分なオリジナリティとクオリティがあるとの評価がなされた。本論文の内容は査読付きジャー

(別紙様式 3)
(Separate Form 3)

ナル「自然言語処理」および査読付き国際会議 International Conference on Computational Linguistics に採録されている。以上のことから、全審査委員一致で本論文は学位授与に値するとの判断に至った。