論文 / 著書情報
Article / Book Information

題目(和文)	修辞構造解析器の高度化に関する研究
Title(English)	
著者(和文)	小林尚輝
Author(English)	Naoki Kobayashi
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12389号, 授与年月日:2023年3月26日, 学位の種別:課程博士, 審査員:奥村 学,熊澤 逸夫,中山 実,篠崎 隆宏,船越 孝太郎
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12389号, Conferred date:2023/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信 系 情報通信 コース	申請学位（専攻分野）： Academic Degree Requested	博士 （ 工学 ） Doctor of Engineering
学生氏名： Student's Name	小林 尚輝	指導教員（主）： Academic Supervisor(main)	奥村 学
		指導教員（副）： Academic Supervisor(sub)	

要旨（和文 2000 字程度）
Thesis Summary（approx.2000 Japanese Characters ）

本論文では，修辞構造理論に基づいて文書構造の解析を行う研究課題である修辞構造解析の高度化に取り組む．修辞構造解析は文脈解析のひとつであり，単語の並びから文の構造を解析するように，文の並びから文書の構造を解析する．文脈解析を通して文書の構造を正しく理解することで，文書分類や文書要約などの応用課題の性能改善が期待される．修辞構造解析では，まず文書を構成する最小単位である談話単位（Elementally Discourse Unit: EDU）を定義し，文書を EDU の系列へと分割する．そして，それらの間の関係性から隣り合う EDU やその系列であるスパンが結合してより大きなスパンを構築することで，入れ子上のスパン，つまり木構造として文書構造を表現する．このとき隣り合うスパンの間には従属関係を表す核性と修辞関係ラベルが付与される．したがって，修辞構造木を構成する要素である入れ子上のスパンの構造および核性・修辞関係ラベルの推定性能を二つの研究を通して改善することで修辞構造解析の高度化に貢献する．

一つ目の研究では，スパン分割に基づく下向きの解析手法を修辞構造解析へと導入し，さらに文書の持つ階層構造のひとつである段落を用いて解析空間を階層化することで修辞構造解析の性能改善に貢献する．修辞構造木を入力として活用する応用タスクでは，修辞構造木から変換して得られる係り受け木が多く活用されるが，係り受け木における長期の依存関係を正しく獲得するためには，修辞構造木の上部における解析性能が重要である．しかし，従来の修辞構造解析器は Shift-Reduce 法による上向きの解析が一般的であり，上向きの解析手法では解析の誤りが伝播して木の上部の解析性能が低下する懸念がある．解析誤りの伝播を緩和する方法の一つとして，解析空間を分割する方法が考えられ，従来手法として文内と文間による分割がある．本研究では，木の上部の解析性能を改善するために，木の上部から下向きに解析を行う解析手法であるスパン分割に基づく解析手法を修辞構造解析に導入する．これは，文の句構造解析に向けて提案された解析手法であり，文に対応するスパンから解析を始め，スパンが単一の単語になるまで再帰的にスパンを分割することで句構造木を構築する．本研究では文書に対応するスパンから解析を始め，スパンが単一の EDU になるまで再帰的に分割することで修辞構造木を構築する．また，本研究では解析誤りの伝播を緩和するための解析空間の分割において，文書の持つ階層構造の一つである段落に着目し，段落と文を用いた 3 層の階層化を提案する．実験では，下向きの解析手法は特に構造の推定において性能改善を示し，段落を利用した階層化についても有効性が示された．

二つ目の研究では，複数の解析器の間で一致する部分木を活用したデータ拡張を提案し，大規模かつ信頼性のある疑似正解データセットの構築により修辞構造解析の性能改善に貢献する．解析器の学習は人手によって文書に修辞構造木が付与されたデータセットが必要となる．しかし，文書を対象とした修辞構造木の付与は専門性を必要とすることからアノテーションコストが大きく，人手によるデータセットの拡張は容易ではない．特に，ニューラルネットワークを利用した解析器の学習には大量の学習データを必要とするが，修辞構造解析における最大のコーパスである RSTDT でさえ 385 文書しかない．そこで，人手のアノテーションを必要とせず自動的に獲得した疑似正解データセットの構築を行う研究が行われている．疑似正解データセットの作成では，修辞構造の付与されていないテキストコーパスを対象に既存の解析器や従属関係を判別する分類器を教師として用いて修辞構造木を付与する．データ拡張手法のひとつである tri-training では教師となる分類器を複数用意し，それらの間で合意が取れたデータを疑似正解データとして選択することで疑似正解データセットの信頼性を高める．しかし，修辞構造解析は解析対象の文書が多くの EDU から構成されるため，複数の解析器の間で文書全体の一致をとることが難しい．そこで本研究では，複数の教師解析器の間で一致した部分木を疑似正解データセットとして活用することで大規模かつ信頼性のある疑似正解データセットを構築する．複数の木の間で一致する部分木の抽出には，木の捜査（tree-traversal）に基づく高速なアルゴリズムを提案し，木のノード数 n に対して O(n)の実行時間で部分木を抽出可能とした．実験では，疑似正解データの種類による性能の変化を比較し，部分木を活用した提案手法が性能および学習効率においてその他の手法を上回ることを示した．

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報通信 系 情報通信 コース	申請学位（専攻分野）： Academic Degree Requested	博士 （ 工学 ） Doctor of Engineering
学生氏名： Student's Name	小林 尚輝	指導教員（主）： Academic Supervisor(main)	奥村 学
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 300 語程度）
Thesis Summary（approx.300 English Words ）

In this paper, we work on the advancement of RST parsing through two pieces of research.
RST parsing is a research topic that analyzes document structure based on rhetorical structure theory (RST).

In RST parsing, document structure is represented by nested text spans, and nuclearity and rhetorical-relation labels are assigned between neighbor spans. The dependencies between these spans are important for understanding the documents and are expected to be used for application tasks such as document summarization.

In the first research, in order to improve the parsing performance at the top of the parsed RST tree, which is important for converting the RST tree to a dependency structure, we introduced a top-down parsing method based on span splitting. In order to reduce the propagation of parsing errors caused by the difficulty of treating large-sized trees, we also proposed to divide the search space by paragraph boundaries which are one of the hierarchical structure of a document. In the experiments, the proposed top-down parser achieved higher parsing performance than existing bottom-up parsers, and dividing search space by paragraph boundaries worked effectively.

In the second research, we augment the dataset and construct a silver dataset used for training the RST parser because the training of a neural network requires a large dataset however RSTDT which is the largest dataset in RST parsing has only 381 documents. To make the silver dataset large and reliable, we construct the dataset by using multiple RST parsers and extracting subtrees that agree among them as silver data. In the experiments, we compared multiple types of silver datasets (subtree-level, whole tree-level, and using a single teacher parser) and proposed subtree-level augmentation achieved the best parsing performance and showed better training efficiency.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).