T1-A-6-05

ExhaustiveTree: 網羅的な二次元座標木を用いた表の表現

林 知司, 宮森 恒 （京都産業大学大学院）

目的 網羅的な二次元座標木を用いて表を表現することにより，表を有効的に活用する

背景 表形式のデータは様々な構造で記述され，その内容理解やテキストとの関連付けは依然として容易ではない

手法 表中の結合セルの大きさを網羅的に把握した上で，その大きさの降順に二次元座標木を構築することにより表を表現する

結論 提案手法は，セルタイプ分類のタスクにおいて，最も高い性能を示した

はじめに

表には，結合セル等によって作られる複雑な構造が存在することに着目し，これをいくつかの方法で，表の表現に取り入れる手法を開発した

TUTA

表に含まれる階層的な情報を二次元座標木と呼ばれるツリーベースの構造で表現することにより，5つのデータセットで最先端の結果を達成

TUTAの課題

表内に階層構造がある場合，表の上(左)から下(右)に向かって結合セルの大きさが徐々に小さくなることを前提としている

従来手法

表の上(左)から下(右)に向かって二次元座標木を構築

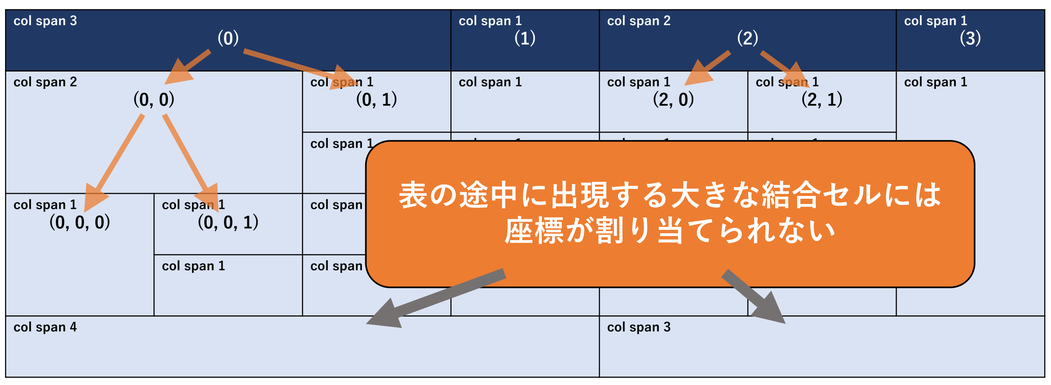

図1: default vertical tree

提案手法

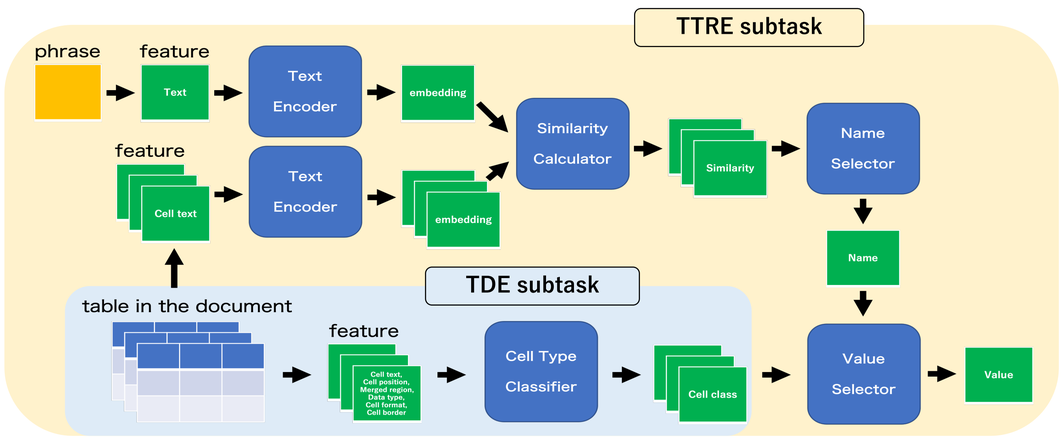

図2: 提案手法の処理概要

表中の結合セルの大きさを網羅的に把握した上で，
その大きさの降順に二次元座標木を構築

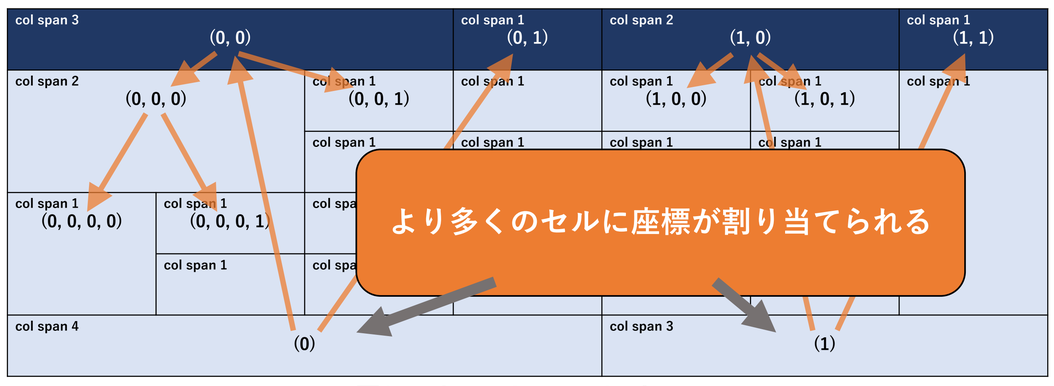

図3: exhaustive vertical tree

実験

- 表構造理解に有効であるかをTDE データセットとTTREデータセットを用いて検証
- TDEデータセット，TTREデータセットはHTML形式
- TDEサブタスクは，有価証券報告書の表の各セルを，4 つのクラスに分類する
- TTREサブタスクは，有価証券報告書中の与えられたテキストに対して，それに関連する表中の該当するセルを選択する

評価方法

TDE (Table Data Extraction)

表の木構造を考慮しない手法 (no tree)，従来手法 (default tree)，提案手法 (exhaustive tree)の3種類の手法を比較することにより有用性を検証

TTRE (Text-to-Table Relationship Extraction)

表の木構造を考慮しない手法 (no tree)，従来手法 (default tree)，提案手法 (exhaustive tree)，セルのクラスを考慮しない手法の4種類の手法を比較することにより有用性を検証

結果と考察

- マクロ平均の適合率，再現率，F1スコアにおいて，exhaustive treeによる手法，no treeによる手法，default treeによる手法の順で高い性能を示す結果となった
- 全ての手法において，Header，Attribute，Dataクラスに比べ，Metadataクラスに対しては低い性能を示す結果となった
- これは，Metadataクラスに分類される事例が極端に少なく，効率的に学習できなかったことが原因として挙げられる

表1: TDEデータセットでの評価結果

Method	Metadata	Header	Attribute	Data	マクロ平均		
	F1	F1	F1	F1	適合率	再現率	F1
TUTA no tree	0.6744	0.8620	**0.9886**	0.9046	**0.8863**	**0. 8365**	**0. 8574**
TUTA default tree	0.6654	**0.8652**	0.9878	0.9002	0. 8847	0. 8330	0. 8546
TUTA exhaustive tree	**0.6743**	0.8679	0.9901	**0.9017**	0. 8887	0. 8375	0. 8585

- no treeによる手法が全てのValue，TotalのF1スコアにおいて最も高い性能を示した
- これは，TDEサブタスクでの，Dataクラスを分類するためのモデルの性能が，TTREサブタスクに大きな影響を与えたことが原因として挙げられる

表2: TTREデータセットでの評価結果

Method	Name			Value			Total
	適合率	再現率	F1	適合率	再現率	F1	F1
multilingual-e5	0.3556	0.4574	0.3221	0.0857	0.5069	0.1186	0.2204
multilingual-e5 + **TUTA** no tree	0.3556	0.4574	0.3221	0.2733	**0.5048**	0.2742	0.2982
multilingual-e5 + **TUTA** default tree	0.3556	0.4574	0.3221	0.2514	0.5001	0.2596	0.2909
multilingual-e5 + **TUTA** exhaustive tree	0.3556	0.4574	0.3221	**0.2613**	0.5044	**0.2639**	**0.2930**

- 所与のテキストに対して上位1件のセルを取得した場合に，高い性能を示しており，複数セルの検索については，十分な性能を発揮できていない

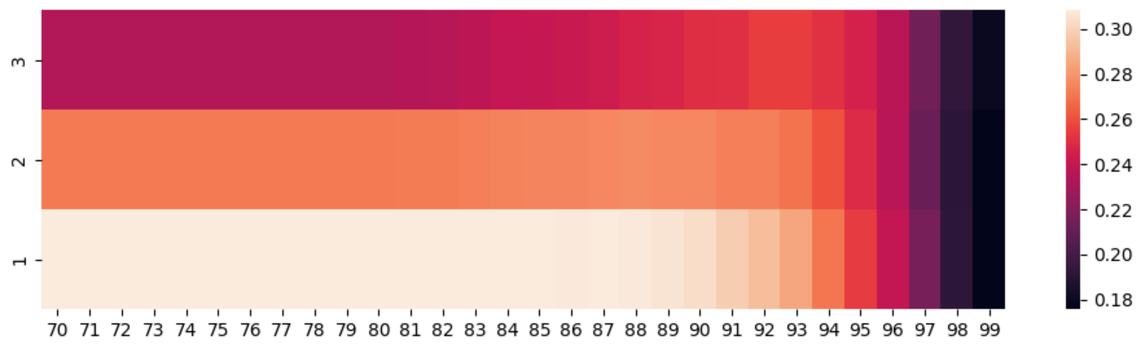

図4: Totalにおける類似度と上位K件による閾値ごとの評価結果のヒートマップ

むすんで、うみだす。
京都産業大学
KYOTO SANGYO UNIVERSITY