

# 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

DEIM 2024 Day 2: 2月29日(木) 17:45~18:10 (JST)

Track 1: 自然言語処理・機械学習基礎 構造データ処理

# 目次



はじめに

関連研究

データセット

提案手法

実験

追加実験

まとめ

# 目次



はじめに

関連研究

データセット

提案手法

実験

追加実験

まとめ

# はじめに 研究の背景

**表形式のデータ**は，テキストとともに様々な文書中で一般的に使用されるデータ形式であり，多くのアプリケーションにとって重要で不可欠な要素



表. ～に関する表

	ABC	DEF
〇〇	123	456
× ×	234	567

# はじめに 表形式のデータを扱う際の難点

しかし、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

例えば、ファクトチェック支援の分野では・・・



**フェイク!**

テキスト)

〇〇株式会社の固定資産は、  
前事業年度から3倍以上増加した

表)

(単位：百万円)

	前事業年度 (2020年3月31日)	当事業年度 (2021年3月31日)
流動資産	343,236	989,612
固定資産	1,587,631	2,089,331
繰延資産	731	2,419
資産合計	1,931,598	3,081,362

# 目次



はじめに
関連研究
データセット
提案手法
実験
追加実験
まとめ

# 関連研究 表の表現に特化した事前学習モデル

## TaBERT [1]

- 自然言語と(半)構造化テーブルの表現を共同で学習する事前学習モデル
- WikipediaとWDCデータセットによる2,600万の表とその英語コンテキストからなる大規模なコーパスで学習
- 提案手法とは、表の階層構造を表現して学習した点において異なる

## TAPAS [2]

- 論理形式を生成せずにテーブル上で質問に応答する事前学習モデル
- BERTのアーキテクチャを拡張して表を入力としてエンコードし、テキストセグメントとWikipediaのテーブルの効果的な事前学習から初期化し、エンドツーエンドで学習
- 提案手法とは、表の階層構造を表現して学習した点において異なる

[1] Pengcheng Yin and Graham Neubig and Wen-tau Yih and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data.

[2] Herzig, Jonathan and Nowak, Pawel Krzysztof and Müller, Thomas and Piccinno, Francesco and Eisenschlos, Julian. 2020. TaPas: Weakly supervised table parsing via pre-training.

# 関連研究 表の表現に特化した事前学習モデル

## TUTA [3]

- 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
- 5つのデータセットで最先端の結果を達成

### TUTAの前提条件

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



# 関連研究 従来手法(TUTA)の課題

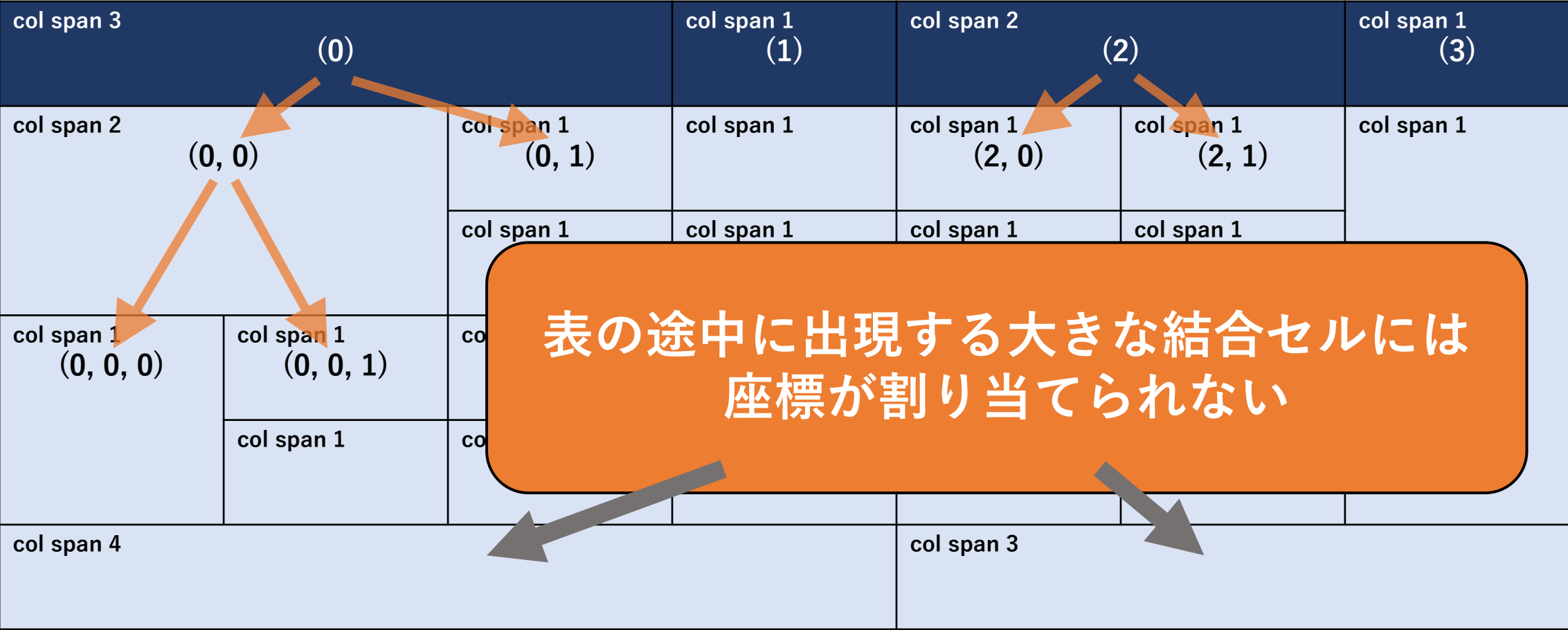


図1: default vertical tree

# 目次



はじめに
関連研究
データセット
提案手法
実験
追加実験
まとめ

# データセット 概要

## TDE (Table Data Extraction)

- HTML形式の有価証券報告書
- 有価証券報告書に含まれる表の各セルを,  
4つのクラスに分類するためのデータセット

## TTRE (Text-to-Table Relationship Extraction)

- HTML形式の有価証券報告書
- 有価証券報告書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ためのデータセット

# データセット TDEデータセット

有価証券報告書に含まれる表の各セルを、  
**Metadata**, **Header**, **Attribute**, **Data**の  
4つのクラスに分類

(単位：百万円)

	第20期	第21期	第22期
流動資産	34,323	98,961	57,158
固定資産	158,763	208,933	189,903
繰延資産	73	241	180
資産合計	193,159	308,135	247,241

	Metadata		
	Header	Header	Header
Attribute	Data	Data	Data
Attribute	Data	Data	Data
Attribute	Data	Data	Data
Attribute	Data	Data	Data

# データセット TTREデータセット

有価証券報告書の与えられたテキストに対して、  
それに関連する表中の**Name**, **Value**に  
該当するセルを選択

テキスト)

第21期及び第22期の固定資産

表)

(単位：百万円)

	第20期	第21期	第22期
流動資産	34,323	98,961	57,158
固定資産	158,763	208,933	189,903
繰延資産	73	241	180
資産合計	193,159	308,135	247,241

# データセット 表の構造の特徴

有価証券報告書に含まれる表には、  
表中の結合セル等によって作られる複雑な構造が存在する

(Unit : One million yen) (単位：百万円)

ヘッジ会計 Hedge accounting method の方法	取引の種類 Type of transaction	主なヘッジ Main hedged subject 対象	契約額等 Contract amount and so on	うち1年超 In over one years	時価 Current market price
原則的 処理方法 Processing method in general	商品スワップ取引 Swap transactions of Merchandise				
	受取変動・支払固定 Receive-floating・Pay-fixed 原油 Crude oil	営業未払金 Non-operating trade payables	69,132	27,452	△24,304
	商品オプション取引 Option transactions of Merchandise				
	売建 Going short プット 原油 Put Option Crude oil	Non-operating trade payables 営業未払金	33,120	15,468	△7,229
	買建 Going long コール 原油 Call Option Crude oil	Non-operating trade payables 営業未払金	42,798	20,103	△1,717
	合計 Sum		145,051	63,025	△33,250

図2: 有価証券報告書に含まれる表の例

# 目次



はじめに
関連研究
データセット
<b>提案手法</b>
実験
追加実験
まとめ

# 提案手法 処理概要

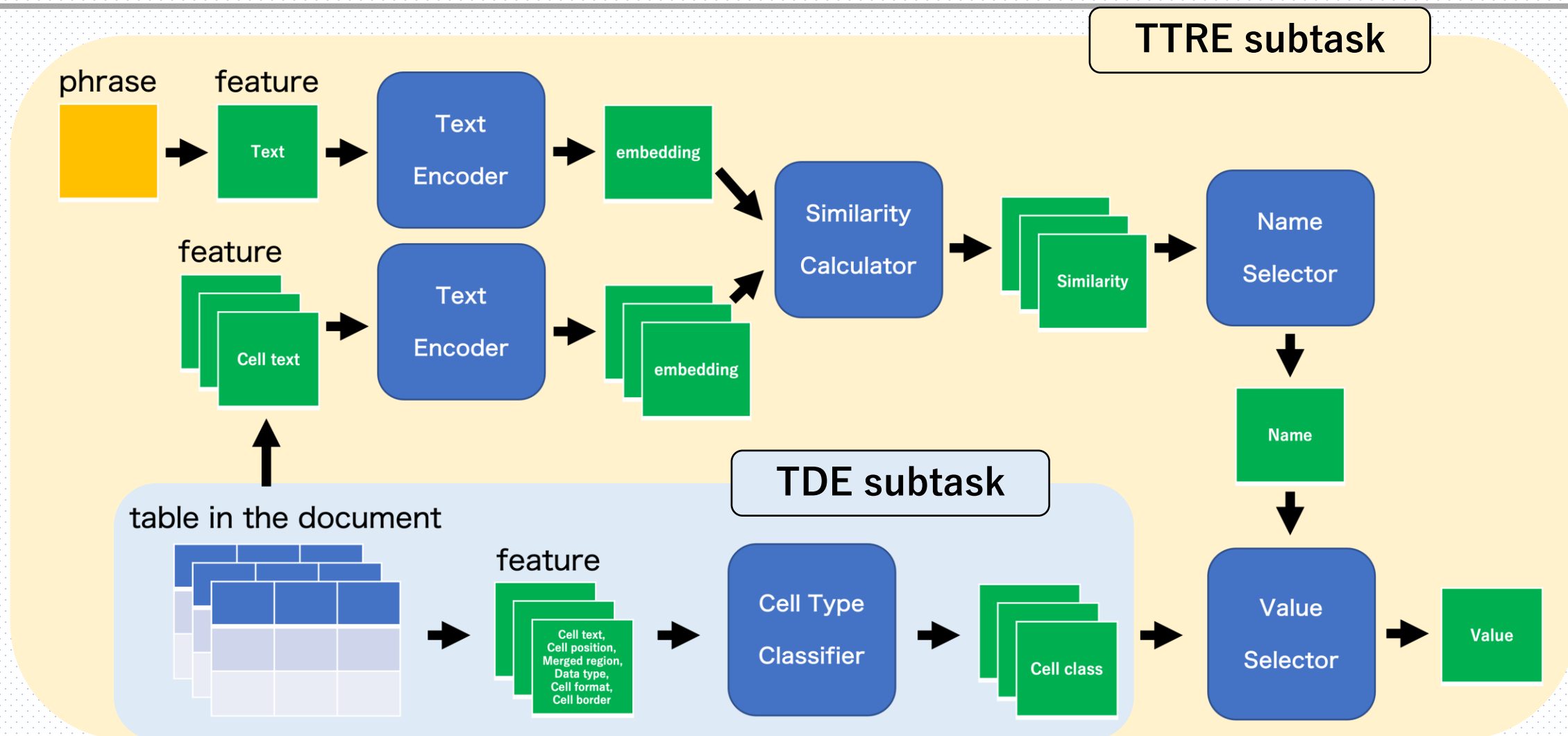


図3: 提案手法の処理概要



# 提案手法 表の特徴

表1: 使用した表の特徴

表の特徴	例
セルの文字	<ul style="list-style-type: none"><li>セルの文字</li></ul>
セルの位置	<ul style="list-style-type: none"><li>行 / 列 の座標</li><li>木に基づく座標</li></ul>
結合セル	<ul style="list-style-type: none"><li>行方向に結合されたセルの数</li><li>列方向に結合されたセルの数</li></ul>
データタイプ	<ul style="list-style-type: none"><li>セルの中に日時を表す文字があるか</li><li>セルの中に数式があるか</li></ul>
セルの形式	<ul style="list-style-type: none"><li>太字であるか</li><li>背景色が白か</li><li>文字色が黒か</li></ul>
セルの罫線	<ul style="list-style-type: none"><li>上部に罫線があるか</li><li>下部に罫線があるか</li><li>左部に罫線があるか</li><li>右部に罫線があるか</li></ul>

# 提案手法 表の表現

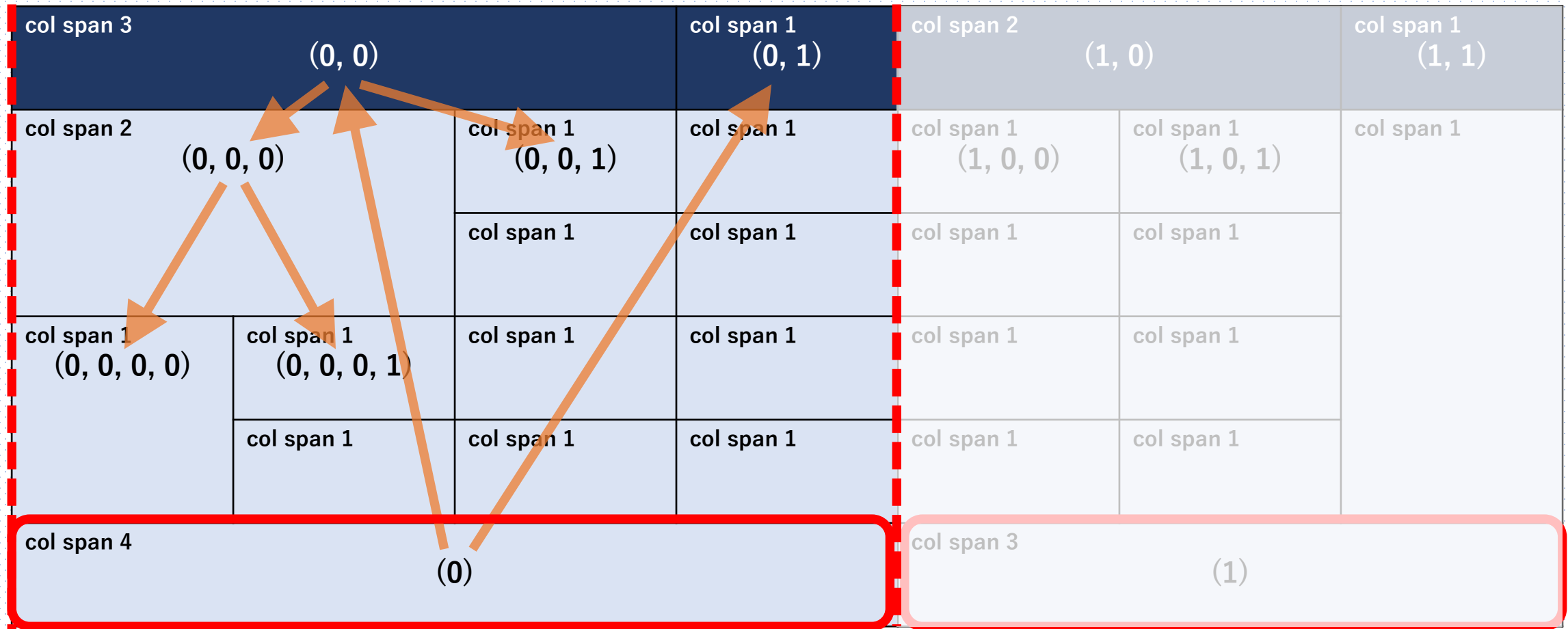


図4: exhaustive vertical tree

# 提案手法 表の表現

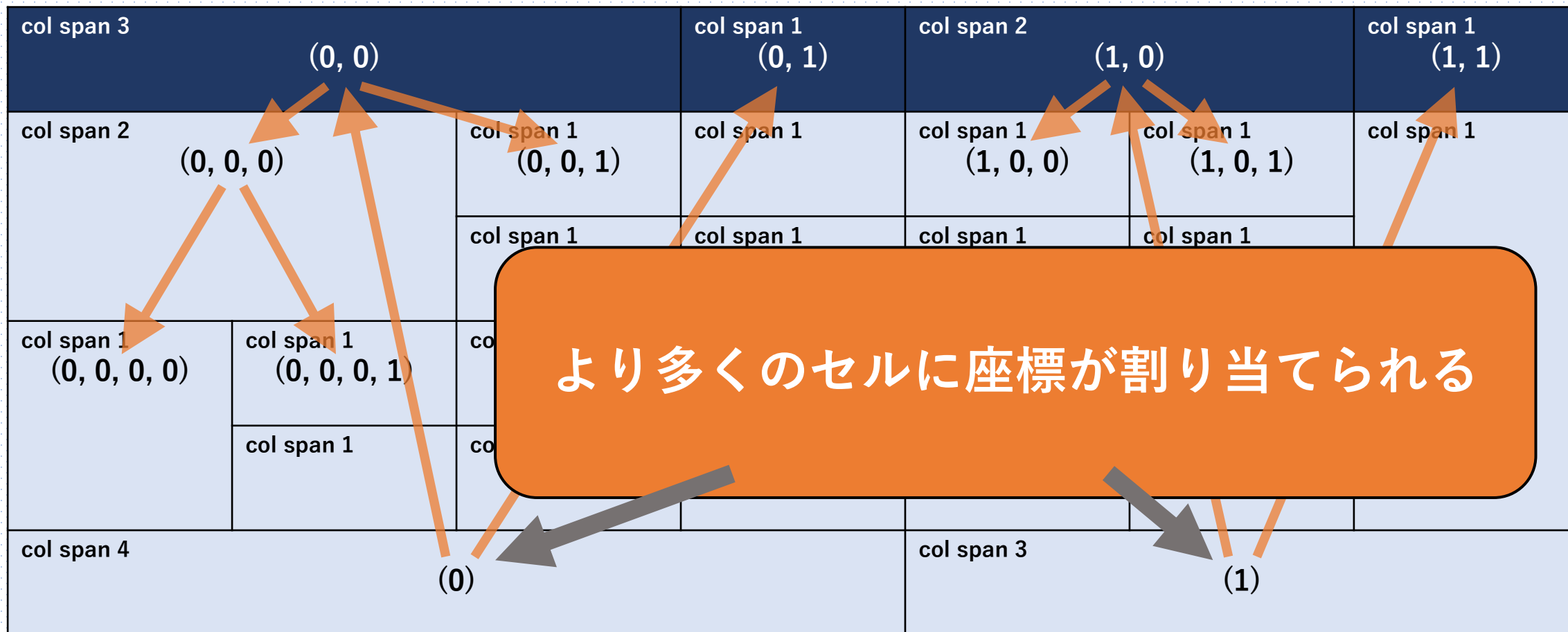


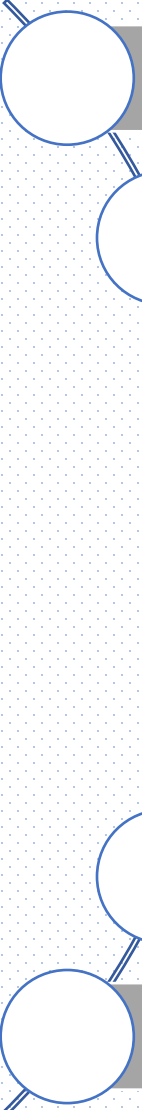
図4: exhaustive vertical tree

# 提案手法 用語の定義

表2: 木の種類の概説

木の種類	概説
default tree	TUTAで定まる二次元座標木
default vertical tree	垂直方向(列)に関する木
default horizontal tree	水平方向(行)に関する木
exhaustive tree	提案手法で定まる二次元座標木
exhaustive vertical tree	垂直方向(列)に関する木
exhaustive horizontal tree	水平方向(行)に関する木

# 目次



はじめに
関連研究
データセット
提案手法
実験
追加実験
まとめ

# 実験 目的

提案手法(exhaustive tree)が，表構造理解に有効であることを明らかにすることを目的とする

colspan 3 (0, 0)			colspan 1 (0, 1)	colspan 2 (1, 0)		colspan 1 (1, 1)
colspan 2 (0, 0, 0)		colspan 1 (0, 0, 1)	colspan 1	colspan 1 (1, 0, 0)	colspan 1 (1, 0, 1)	colspan 1
		colspan 1	colspan 1	colspan 1	colspan 1	
colspan 1 (0, 0, 0, 0)	colspan 1 (0, 0, 0, 1)	colspan 1	colspan 1	colspan 1	colspan 1	colspan 1
colspan 4 (0)				colspan 3 (1)		

より多くのセルに座標が割り当てられる

## 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

# 実験 評価方法

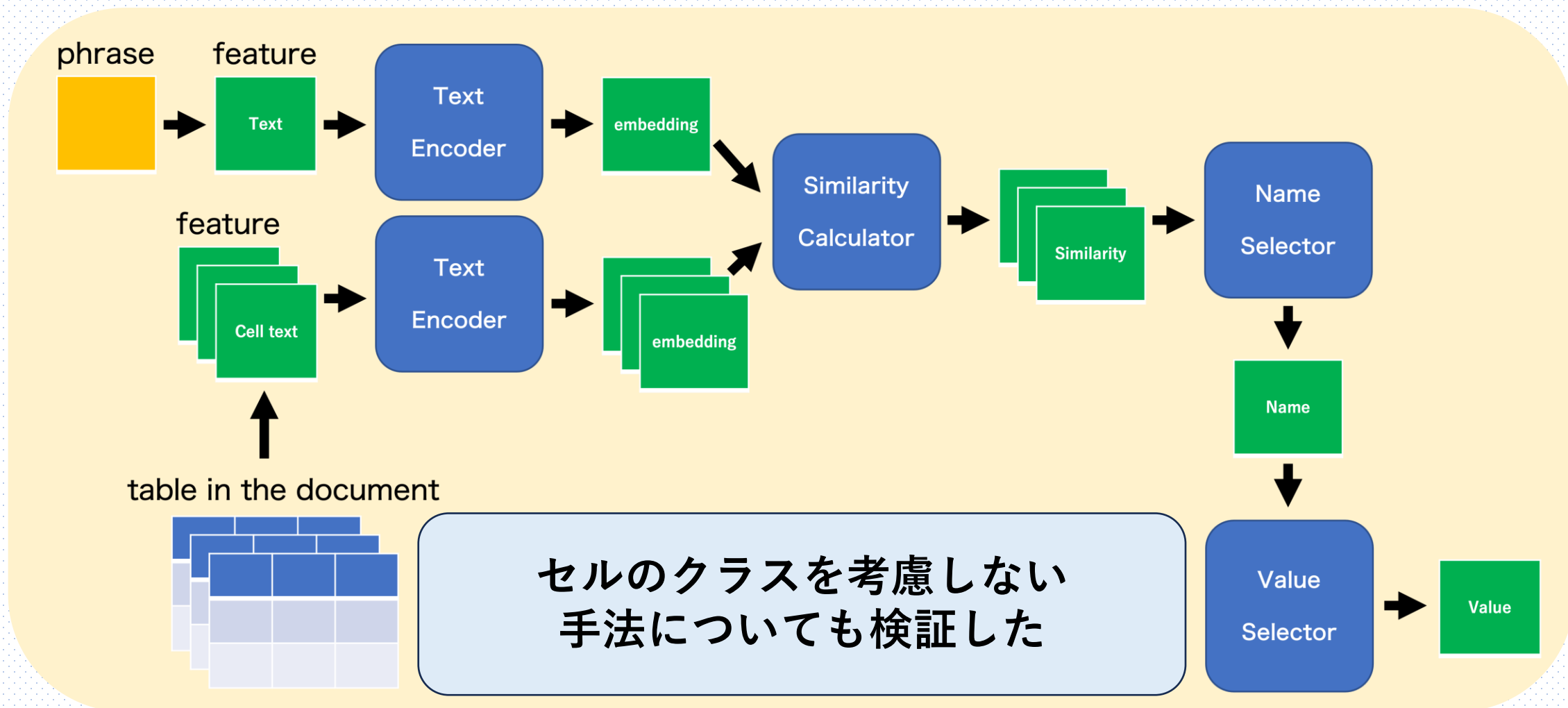


図3: 提案手法の処理概要



# 実験 評価方法

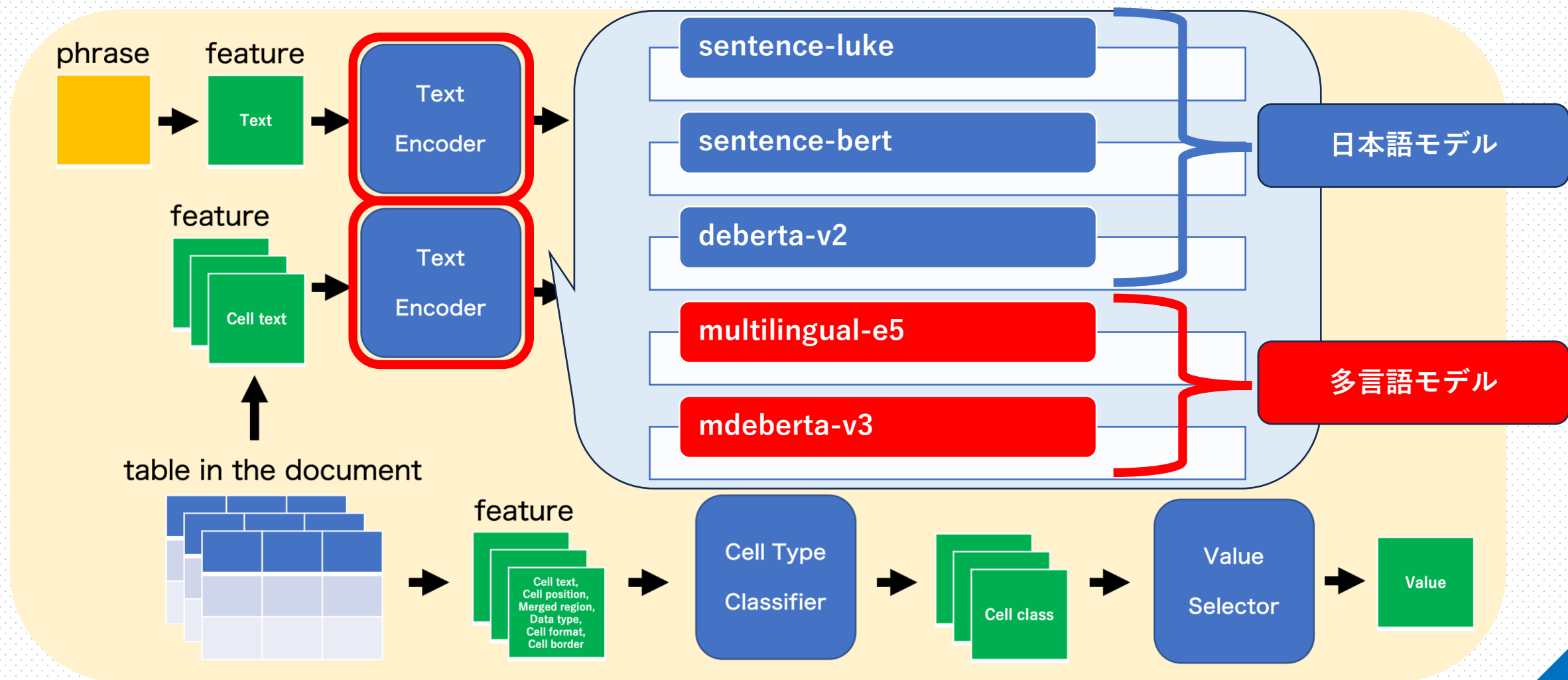


図3: 提案手法の処理概要

# 実験 TDEデータセットでの評価結果

表3: 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	<u>0.6744</u>	0.8620	0.9886	<u>0.9046</u>	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	<u>0.8679</u>	<u>0.9901</u>	0.9017	<u>0.8887</u>	<u>0.8375</u>	<u>0.8585</u>

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

# 実験 TDEデータセットでの評価結果

表3: 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	<u>0.6744</u>	0.8620	0.9886	<u>0.9046</u>	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	<u>0.8679</u>	<u>0.9901</u>	0.9017	<u>0.8887</u>	<u>0.8375</u>	<u>0.8585</u>

全ての手法において、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

# 実験 TTREデータセットでの評価結果

表4: 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	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	0.0857	<u>0.5069</u>	0.1186	0.2204
multilingual-e5 + TUTA no tree	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	<u>0.2733</u>	<u>0.5048</u>	<u>0.2742</u>	<u>0.2982</u>
multilingual-e5 + TUTA default tree	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	0.2514	0.5001	0.2596	0.2909
multilingual-e5 + TUTA exhaustive tree	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	<u>0.2613</u>	0.5044	<u>0.2639</u>	<u>0.2930</u>

“Total”において、Text Encoderにmultilingual-e5を用いて、Cell Type ClassifierにTUTA no treeを用いた手法が、最も高い性能を示した

# 実験 TTREでの考察

なぜ，TTREにおいて，TUTA no treeが最も高い性能を示したのか？

予想

TDEでのDataクラスを分類するための  
モデルの性能が，TTREに大きな影響を与える

# 実験 TTREでの考察

表3: 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	<u>0.6744</u>	0.8620	0.9886	<u>0.9046</u>	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	<u>0.8679</u>	<u>0.9901</u>	0.9017	<u>0.8887</u>	<u>0.8375</u>	<u>0.8585</u>

Dataクラスを分類するための  
モデルの性能を向上させることが重要

# 実験 TTREデータセットでの評価結果

表4: 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	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	0.0857	<b>0.5069</b>	0.1186	0.2204
multilingual-e5 + TUTA no tree	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	<b>0.2733</b>	<b>0.5048</b>	<u>0.2742</u>	<u>0.2982</u>
multilingual-e5 + TUTA default tree	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	0.2514	0.5001	0.2596	0.2909
multilingual-e5 + TUTA exhaustive tree	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	<b>0.2613</b>	0.5044	<b>0.2639</b>	<b>0.2930</b>

TUTAを使用した手法と、TUTAを使用しなかった手法で、Valueを比較すると、再現率には大きな差がなく、適合率には有意な差があった

# 実験 TTREでの考察

---

TUTAを使用した手法と，TUTAを使用しなかった手法で，Valueを比較すると，再現率には大きな差がなく，適合率には有意な差があった

Valueを決定する際に，セルのタイプを判定し，  
dataクラス以外を除外する手法は非常に有力



# 実験 TTREデータセットでの評価結果

表4: 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	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	0.0857	<u>0.5069</u>	0.1186	0.2204
multilingual-e5 + TUTA no tree	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	<u>0.2733</u>	<b>0.5048</b>	<u>0.2742</u>	<u>0.2982</u>
multilingual-e5 + TUTA default tree	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	0.2514	0.5001	0.2596	0.2909
multilingual-e5 + TUTA exhaustive tree	<u>0.3556</u>	<u>0.4574</u>	<u>0.3221</u>	<b>0.2613</b>	0.5044	<b>0.2639</b>	<b>0.2930</b>

全ての手法で、precisionよりもrecallの方が高い結果となった

# 実験 TTREでの考察

---

全ての手法で、precisionよりもrecallの方が高い結果となった

Nameが、所与のフレーズとセルの値の類似度のみで  
決められているため、実際に参照すべき表とは異なる表も  
まとめて取得してしまっていることが原因

# 実験 TTREでの考察

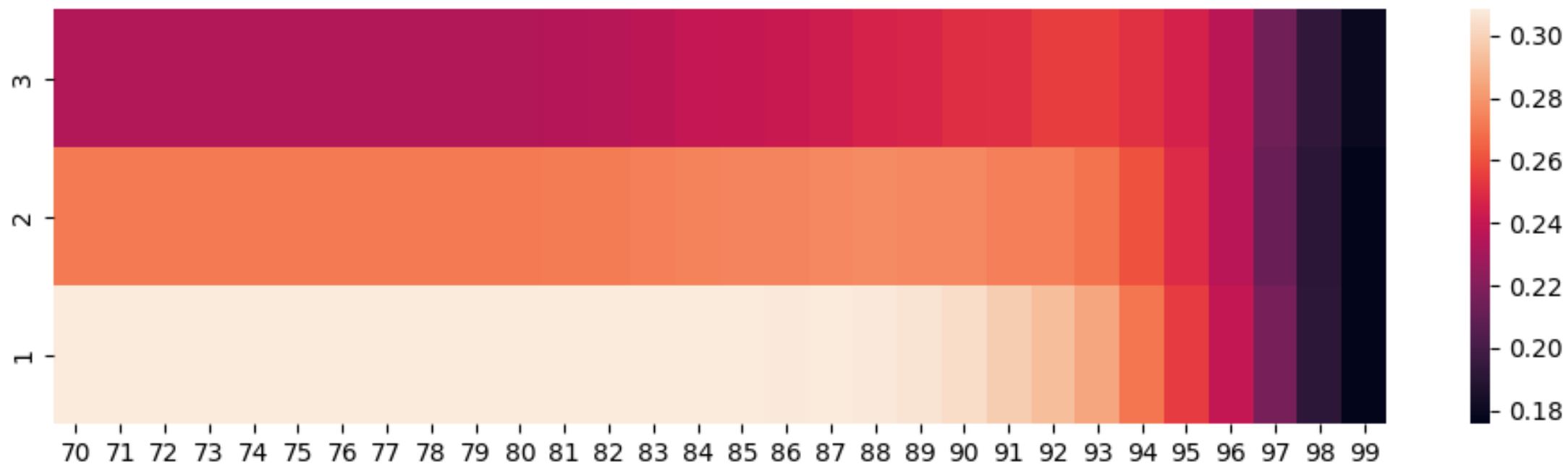


図5: Totalにおける類似度と上位K件による閾値ごとの評価結果のヒートマップ  
(横軸が類似度, 縦軸が上位K件)

Totalにおいて、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

# 実験 TTREでの考察

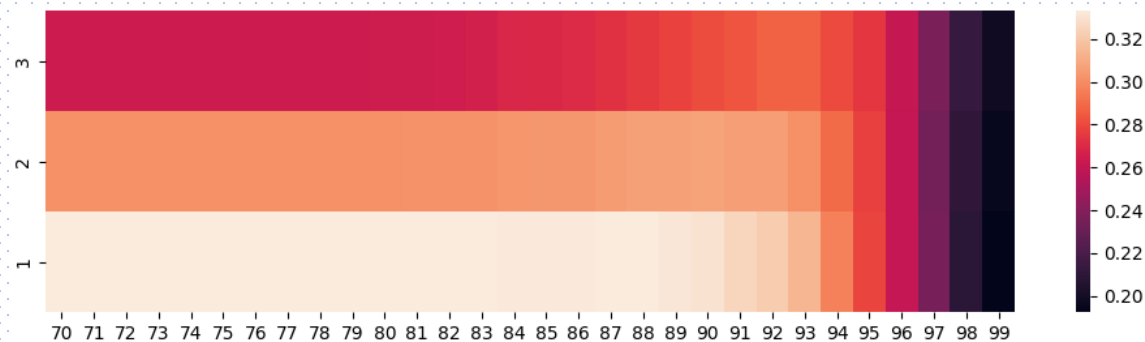


図6: Nameにおける類似度と上位K件による閾値ごとの評価結果のヒートマップ  
(横軸が類似度, 縦軸が上位K件)

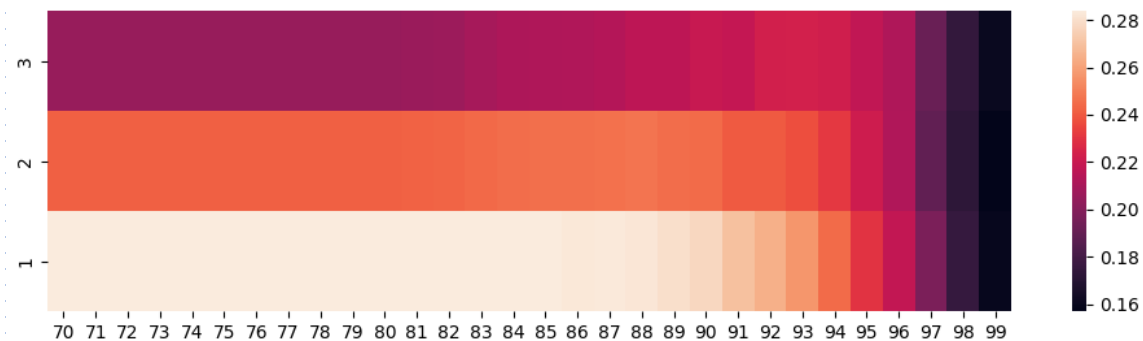
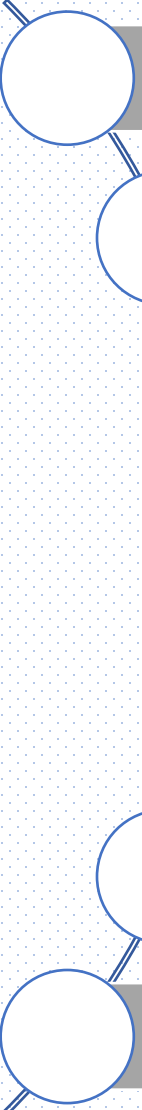


図7: Valueにおける類似度と上位K件による閾値ごとの評価結果のヒートマップ  
(横軸が類似度, 縦軸が上位K件)

Name, Valueにおいても, 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

# 目次



はじめに
関連研究
データセット
提案手法
実験
追加実験
まとめ

# 追加実験 目的

提案手法(exhaustive tree)が，スプレッドシート形式の表における表構造理解に有効であるかを明らかにする

colspan 3 (0, 0)			colspan 1 (0, 1)	colspan 2 (1, 0)		colspan 1 (1, 1)
colspan 2 (0, 0, 0)		colspan 1 (0, 0, 1)	colspan 1	colspan 1 (1, 0, 0)	colspan 1 (1, 0, 1)	colspan 1
		colspan 1	colspan 1	colspan 1	colspan 1	
colspan 1 (0, 0, 0, 0)	colspan 1 (0, 0, 0, 1)	colspan 1	colspan 1	colspan 1	colspan 1	colspan 1
colspan 4 (0)				colspan 3 (1)		

より多くのセルに座標が割り当てられる

## DeEx (deexcelerator)

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

# 追加実験 DeExデータセットでの評価結果

表5: DeExデータセットでの評価結果

Method	metadata	notes	data	attributes	header	derived	macro-F1
TUTA no tree	0.8574	<b>0.4825</b>	<u>0.9937</u>	0.8114	<u>0.8701</u>	<u>0.8092</u>	<b>0.8041</b>
TUTA default tree	<b>0.8585</b>	0.4419	0.9930	<u>0.8382</u>	0.8547	0.7391	0.7876
TUTA exhaustive tree	<u>0.8671</u>	<u>0.5589</u>	<b>0.9931</b>	0.8045	<b>0.8638</b>	0.7491	<u>0.8061</u>
TUTA raw tree	0.8516	0.4271	0.9911	0.7888	0.8070	<b>0.7767</b>	0.7737

提案手法(TUTA exhaustive tree)が最も高い性能を示した



# 追加実験 DeExデータセットでの評価結果

表6: DeExデータセットでの評価結果における関連研究との比較

Method	macro-F1
TaBERT [1]	50.0
TAPAS [2]	68.6
TUTA (従来手法) [3]	76.6
TUTA exhaustive tree (提案手法)	<u>80.6</u>

+ 4.0 %

提案手法(TUTA exhaustive tree)が最も高い性能を示した

[1] Pengcheng Yin and Graham Neubig and Wen-tau Yih and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data.

[2] Herzig, Jonathan and Nowak, Pawel Krzysztof and Müller, Thomas and Piccinno, Francesco and Eisenschlos, Julian. 2020. TaPas: Weakly supervised table parsing via pre-training.

[3] Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2021. TUTA: Tree-based Transformers for Generally Structured Table Pre-training.

# 目次



はじめに

関連研究

データセット

提案手法

実験

追加実験

まとめ

# まとめ

提案手法(exhaustive tree)が，表構造理解に有効であることを明らかにすることを目的して検証した

colspan 3 (0, 0)			colspan 1 (0, 1)	colspan 2 (1, 0)		colspan 1 (1, 1)
colspan 2 (0, 0, 0)		colspan 1 (0, 0, 1)	colspan 1	colspan 1 (1, 0, 0)	colspan 1 (1, 0, 1)	colspan 1
		colspan 1	colspan 1	colspan 1	colspan 1	
colspan 1 (0, 0, 0, 0)	colspan 1 (0, 0, 0, 1)	colspan 1	colspan 1	colspan 1	colspan 1	colspan 1
colspan 4 (0)			colspan 3 (1)			

より多くのセルに座標が割り当てられる

# まとめ

## TDE (Table Data Extraction)

- 提案手法(TUTA exhaustive tree)が最も高い性能を示した

## TTRE (Text-to-Table Relationship Extraction)

- セルを検索する際に、Dataクラスに該当するセル以外を除外する手法は有効
- TDEでのDataクラスを分類するためのモデルの性能が、TTREに大きな影響を与える

## DeEx (deexcelerator)

- 提案手法(TUTA exhaustive tree)が最も高い性能を示した