言語処理学会第32回年次大会

北海道大学 Panasonic

法規・規格文書PDFからの階層ツリー型データセット構築

連絡先：hirayama.h77@gmail.com

平山 大世[1], 松藤 彰宏[2], 小川祐輝[2], 坂地泰紀[1], 野田五十樹[1]

[1]北海道大学, [2]パナソニック株式会社

C5-8

1. 背景と課題

前提

法規・規格PDFは多段にネストした階層構造と多様な番号体系を持つ。

問題

固定長チャンクのRAGでは上位の文脈や参照関係が失われやすい。

動機

階層ツリーを自動で構築したい！

第20 工場審査型式適合評価の方法（規程第19条関係）

↓ **親子**

1 工場審査型式適合評価の方法は、次のいずれかの方法により行うものとし…

親子

（1） 本章第2節を準用した最終検査による検査方法

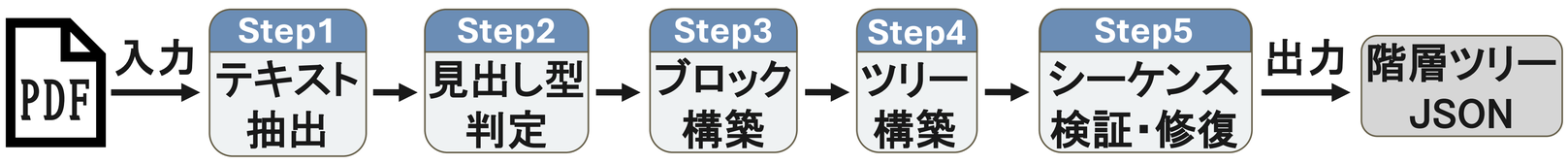
↕ **兄弟**

（2） 製造工程における検査結果等を含め…

困難

- **レイアウト由来の分断**
- **図表由来のノイズ**
- **テキストレイヤの品質問題**

2. 提案手法：パイプラインの全体像



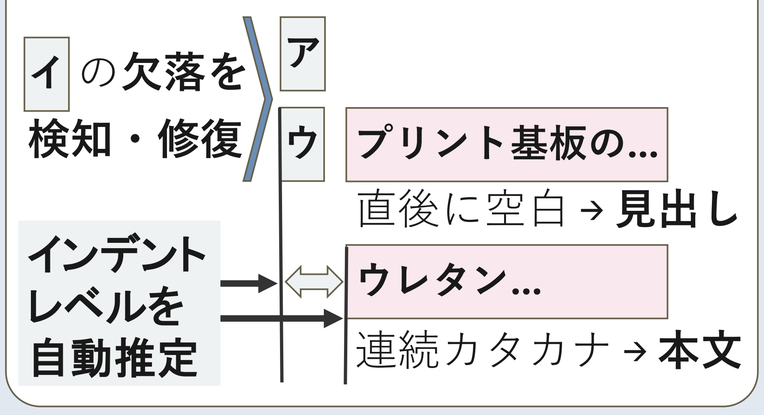
Step	処理内容
1	**pdfplumberで**座標付き単語を取得し、y座標の近接性で**行クラスタリング。** **図表領域・図形密集領域を除外。** ページ番号等ノイズを正規表現で除去。
2	**正規表現で見出し型判定。** 番号マーカー（(1)や(ア)など）を抽出。マーカー直後の**文字境界に着目して**カタカナ箇条（ア, イ, ...）**誤検出抑制。**
3	**行を論理ブロックに統合。** ページ跨ぎによる**分断**、見出し・本文が同一行に書かれる**混在などを解決。** **テキスト正規化**（全角半角統一など）。
4	**インデント分布から離散レベルに量子化**（中央値+中央絶対偏差で閾値推定）。**見出し型の階層レベルとスタックで親子関係を決定。**
5	**兄弟ノードのマーカーシーケンス検証。** 欠番検出時、直前ノード末尾から欠番マーカーで始まる段落を正規表現で検査し、新ノードとして挿入。

3. 主要な技術的工夫

(A) 境界条件による見出し判定
（マーカー直後の文字に注目）

(B) 文書内分布でインデント量子化
（文書固有のレイアウトに適応）

(C) シーケンス検証と欠番項目の復元
（構造誤りを自動修復）



4. 出力データ形式

属性	説明
テキスト	ノードの本文
見出し型	章・節・大見出し・大項目・括弧数字・イロハ・本文
親子関係	親ノード・子ノードリスト
兄弟関係	前後の兄弟ノードへの参照
祖先パス	ルートから当該ノードまで結合した**文脈情報**

5. 評価

- **対象：火災報知設備の感知器及び発信機の検定細則（総務省）**
- **範囲：11ページ（正解189ノード）**

表1: 検出指標

指標	提案手法
F1	1.00

表2: 属性指標

指標	提案手法
見出し型一致率	1.00
完全一致率	0.80
テキスト類似度	0.99

表3: 構造指標

指標	提案手法
親子正解率	1.00
兄弟順序正解率	1.00

6. アブレーション

設定	F1	テキスト類似度
全構成	1.00	0.99
No 誤検出抑制	0.98	0.99
No テキスト正規化	0.98	0.97
No シーケンス検証	0.97	0.98

- **テキスト正規化**: 無効化でF1が0.98に低下。テキスト類似度が0.97に低下。実質的なテキストの差異は小さいが**対応付けに影響。**
- **誤検出抑制**: 無効化でF1が0.98に低下。**カタカナが先頭にある行の誤認識が原因。**
- **シーケンス検証**: 無効化でF1が0.97に低下。改行崩れによる**欠落項目の吸収・飛び番号が原因**（テキストレイヤの品質の影響）。
- **提案パイプラインの骨格で十分強い**が、完全一致には追加の後処理が必要だった。

7. 考察

【構造指標で完全一致を達成】

1. 境界条件による見出し判定
→本文中の数値・一般語彙との**誤検出を抑制**

2. 文書内分布でインデント量子化
→固定閾値では捉えられない**文書固有のレイアウトに適応**

3. シーケンス検証による欠番修復
→PDF抽出時の改行崩れに起因**構造誤りを修正**

【テキストの不完全一致】
→**PDFテキストレイヤの品質問題**
字形類似文字（基板⇔某板）の誤認識、図表キャプション混入

8. まとめと今後の課題

見出し型の違いにどう対応する？
欠落項目の影響をどう抑える？

まとめ

- 境界条件・文書内分布・シーケンス検証により、固定閾値に依存しない構造推定を実現
- F1=1.00、親子/兄弟順序正解率=1.00を達成

今後の課題

- 複数文書での評価・汎化検証
- 抽出した構造情報を用いたRAGシステムでの有効性評価