O-037

ダークデータ抽出における弱教師学習でのラベリング関数推薦技術
Recommendation of Labeling Functions in Weak Supervision for Dark Data Extraction

加藤 大羽‡　　田中剛‡
Daiba Kato　Tsuyoshi Tanaka

1. はじめに

近年，デジタルトランスフォーメーション(DX)で求められるデータドリブンな社会・経営の実現に向けて，AI を用いたデジタル技術への期待値が高まってきている。しかし，企業内には未だデジタル化されていない情報資産が多く存在し，データの収集・分析・解析の大きな障害となっている。企業内に眠る情報資産のおよそ 80%は，そのままでは分析できず価値が不明瞭なダークデータ[1]であると言われ，ダークデータ分析技術の開発とその技術を使いこなすデータサイエンティスト人材の育成が急務となっている。

ダークデータによって生じる課題例の 1 つとして，生命保険会社では，スキャン画像形式の健康診断書から，年齢や健康状態などを確認し，保険の引受が可能かを判断する引受査定業務を行っている。健康診断書のフォーマットが固定化されておらず病院ごとに異なるため，作業者が目視で点検し工程管理システムに入力している。そのため，年間 10 万件以上の健康診断書の対応に常時 100 名程度の人的リソースが割かれており，更には作業者の経験によって業務の効率が大きく左右されてしまっている。

このように，ダークデータ入力・準備作業は，人間の解釈を必要とする作業であるため，人手作業の工数削減が大きな課題となっている。健康診断書や決算書，会社登記情報のような非定型ドキュメントに含まれる様々な情報表現からの情報抽出コスト削減を目的として，スタンフォード大学発の Snorkel[2,3]や Fonduer[4]といった技術が公開されている。Snorkel は，弱教師学習により少量でノイズを含む正解データから情報抽出モデルを作成する技術である。Fonduer は非定型ドキュメントに含まれる複数のリッチテキスト情報(文字情報だけでなく，記載位置情報や表中のテキストであるかなどの情報)を用いた情報抽出を実現するフレームワーク技術である。

我々は，Snorkel および Fondue をコア技術として，帳票認識，顧客分析等の業務支援を目的としたデータ抽出ソリューションを開発した。Snorkel および Fonduer 単体では，高度な知識を有するデータサイエンティストの利用を前提としており，使いこなしの難しさと準備作業工数が大きい問題を抱えている。特に，データサイエンティストが書類上の表現を情報表現へと解釈し，機械的に処理可能なラベリング関数(LF: Labeling Function)と呼ばれるプログラムへと変換する作業が必要[5]で，作業難易度と開発工数のネックとなっている。そこで我々は，あらかじめ汎用的なラベリング関数をテンプレートとして用意しておき，そのテンプレートの中から入力データ種に合わせて最適なラベリング関数を選択するラベリング関数の推薦技術を開発した。これによりラベリング関数の設計作業を省略可し，データ抽出ソリューション利用の敷居を下げつつ作業工数削減を行った。

‡ (株) 日立製作所　Hitachi, Ltd.

2. 想定ユースケース

企業が社内に保有する大量の非定型ドキュメント(ダークデータ)の構造データ化には，人手による手作業の負担が大きく，業務効率化のコストメリットを得られにくかった。省工数で正解データを準備し AI によるダークデータ分析が迅速に可能になれば，社内外の非構造データを構造化して二次利用しやすくなり，業務効率化や，複数種のドキュメント間から見えてくる新たな関係性分析などの活用が期待される。

3. ダークデータ分析による情報抽出

社内に眠る非構造データの省工数での構造データ化を実現するため，我々はダークデータのデータ抽出ソリューションを開発した。図 1 は，本ソリューションでの情報抽出フロー概略図である。非構造データを入力データとし，ダークデータ分析エンジンにより構造化する。入力データは，紙書類のスキャン画像や，PDF などの電子書類，ニュースリリース記事など広く対象とするために，OCR や構文解析を備えている。これにより様々な形式の入力ドキュメントを html (もしくは hocr) 形式へ変換し，入力データの媒体を問わず同一の手順でダークデータ分析エンジンを実行することができる。

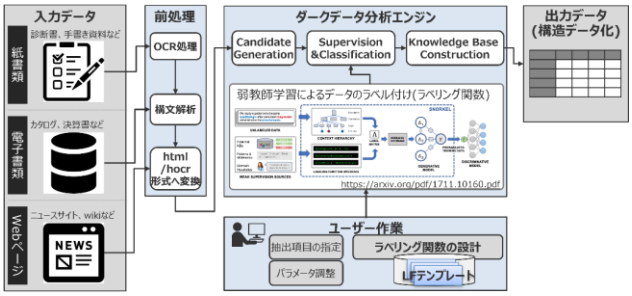

図 1 データ抽出ソリューションの概略図

3.1 ラベリング関数を用いたラベル付け

ダークデータ分析エンジンでは，Snorkel を用いて前処理部で html/hocr 形式ドキュメント化された入力データの情報表現構造と，あらかじめ作成しておいたラベリング関数(LF)の記述内容が一致する箇所を探索する。ここで情報表現構造とは「ドキュメントのヘッダ部分に存在する」「”数値+cm”で記述されている」「xx という文字と同じ表内に存在し，同一行の左側に記述されている」等，抽出対象文字の位置情報，文字情報，表情報を含むリッチテキスト表現構造のことを指す。LF とは，情報表現構造を機械で実行可能な形式で表現したプログラムである。

図 2 は LF 記述例として，有価証券報告書からある決算期の”売上高”の値を抽出したい場合に設計する LF の 1 例である。「『”売上高”という文字列が同じ表内に存在し，”売上高”が同一行左側に記述されている文字列』の場合，

一致(True)を返し，それ以外の場合は不明(Abstain)を返す」という処理をプログラムとして記述する。

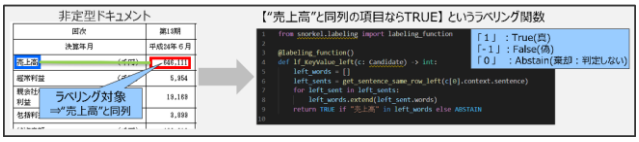

図 2 ラベリング関数の設計例

図 3 のように， LF をユーザが事前に複数準備しておき，パターンマッチにより非定型ドキュメントに含まれる情報表現構造との一致(True)，不一致(False)，不明(Abstain)を判定する。この処理を行うことで，非定型ドキュメント中の情報表現構造と，複数の LF とのパターンマッチ結果の対応表となる Label Matrix を生成することができる。この Label Matrix から Snorkel の弱教師学習によりラベル付け用モデルを生成する。対象の情報表現のラベル付け確からしさを示すスコアを導出し，このスコアが閾値以上であれば Knowledge Base(KB)に登録することで，利用する構造データを選択する。

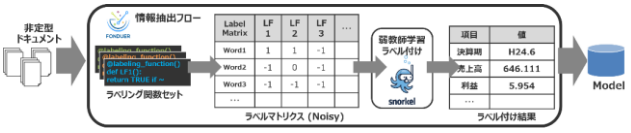

図 3 ラベリング関数を用いたラベル付け概要図

3.2 LF テンプレートによる LF 推薦

従来は，データサイエンティストがドキュメントに含まれている情報表現を 1 つ 1 つ解釈して LF を作成していたため，作業量工数が負担となっていた。そこで，あらかじめ汎用的な LF のテンプレートを作成しておき，その中から，ユーザが入力した抽出したい項目の情報と照合させ，必要な LF の組み合わせを推薦する技術を提案した。

図 4 は，ユーザーが操作する LF 推薦画面の一部である。ユーザーが入力データのドキュメントから抽出したい項目を選択すると，ドキュメント内の文字列の文字情報や位置関係から，利用可能な LF を LF 汎用テンプレートの中から選択し提案する。これにより，LF を作成することなく，コードレスで LF の設計・準備が可能になる。

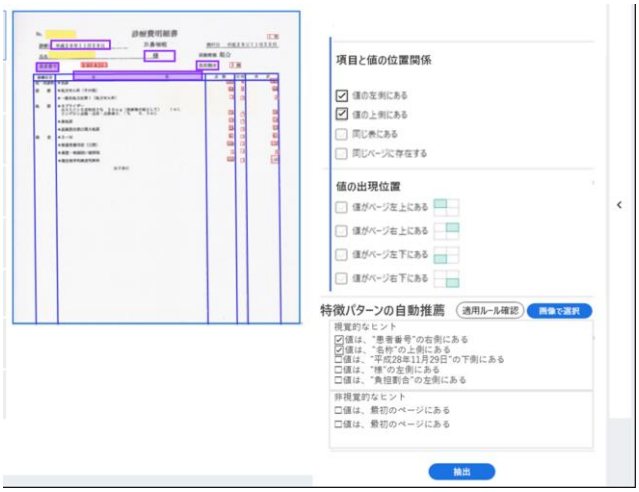

図 4 LF 推薦画面

4. 評価

実際の健康診断書 28 件および診療明細書 32 件を用いて，データ抽出ソリューションによる構造データ化の抽出精度と作業工数の見積もりを実施した。

表 1 は健康診断書と診療明細書それぞれの精度と合計作業工数の結果である。AI など使用せず人間による完全手作業・目視で構造データ化する場合，システムへの入力作業が膨大となり合計 35 時間程度の作業が必要になる。本研究のデータ抽出ソリューションを活用し，健康診断書と診療明細書それぞれの LF を自作設計した場合，抽出精度 76~81%程度で作業工数 15 時間となった。さらに，3.2 で述べた LF テンプレート推薦機能を使用して，LF 自作設計を省略した場合，抽出精度 73%~76%で作業工数 8.7 時間まで削減する見込みとなった。入力データに合わせてその都度 LF を自作設計した場合の方が，抽出しにくいフォーマットのドキュメントからも抽出しやすくなるため精度は高くなりやすいが，LF 推薦機能を活用した場合でも最小限の精度低下に留めつつ，作業工数の削減が確認できた。

表 1 抽出精度と想定作業工数

	健康診断書抽出精度	診療明細書抽出精度	合計作業工数
手作業判定	(100%)	(100%)	35 時間
データ抽出ソリューション (LF 自作)	81%	76%	15 時間
データ抽出ソリューション (LF 推薦使用)	76%	73%	8.7 時間

5. おわりに

本研究では，ダークデータと呼ばれる非定型ドキュメントを構造データ化するデータ抽出ソリューションの開発を行った。従来，データサイエンティストのスキルに依存していた非定型ドキュメントからの情報抽出の効率化を行うために，情報表現テンプレートとしてあらかじめ汎用的な LF を複数作成し，ユーザが入力した抽出したい項目の情報と照合させ，必要な LF の組み合わせを推薦する技術を提案した。その結果，構造データ化作業を省工数化実現の見込みを得た。

参考文献

[1] S. Astorino, “Bringing Light to Dark Data,” Inside Machine Learning, IBM Analytics, https://medium.com/inside-machine-learning/bringing-light-to-dark-data-6275549f07e8 (2018)
[2] Stanford University, “Stanford Data Science Initiative,” https://sdsi.stanford.edu/
[3] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu and C. Ré, “ Snorkel: Rapid Training Data Creation with Weak Supervision,” In Proceedings of the VLDB Endowment, Vol. 11, Issue 3, pp.269-282 (2017)
[4] S. Wu, L. Hsiao, X. Cheng, B. Hancock, T. Rekatsinas, T. Rekatsinas, P. Levis and C. Ré, “ Fonduer: Knowledge Base Construction from Richly Formatted Data,” In Proceedings of the 2018 International Conference on Management of Data, pp.1301-1316 (2018)
[5] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher R´e, “Training complex models with multi-task weak supervision,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 4763–4771, (2019)